The Trilemma of Truth in Large Language Models

Germans Savcisens[1,2] **and Tina Eliassi-Rad**[1,2,3]
[1] Khoury College of Computer Sciences, Northeastern University, Boston, MA, USA
[2] Network Science Institute, Northeastern University, Boston, MA, USA
[3] Santa Fe Institute, Santa Fe, NM, USA
{g.savcisens, t.eliassirad}@northeastern.edu

Abstract

The public often attributes human-like qualities to large language models (LLMs), assuming that they "know" certain things. In reality, LLMs encode information retained during training as internal probabilistic knowledge. This study examines existing methods for probing the veracity of that knowledge and identifies three flawed underlying assumptions. To address these flaws, we introduce `sAwMIL` (Sparse-Aware Multiple-Instance Learning), a multiclass probing framework that combines multiple-instance learning with conformal prediction. `sAwMIL` leverages LLMs' internal representations to classify statements as *true*, *false*, or *neither*. We evaluate `sAwMIL` across 16 open-source LLMs, including default and chat-based variants, using three new curated datasets. Our results show that (1) common probing methods fail to provide a reliable and transferable veracity direction and, in some settings, perform worse than zero-shot prompting; (2) truth and falsehood are not encoded symmetrically; and (3) LLMs encode a third type of signal that is distinct from both true and false.[1]

1 Introduction

Can we trust the content generated by large language models (LLMs)? Recent literature suggests that LLMs possess internal probabilistic knowledge (Petroni et al., 2019; Meng et al., 2022; Chen et al., 2024; Chanin et al., 2024; Pan et al., 2025).[2] However, our understanding of *how* LLMs use this internal knowledge remains fragmented. Hicks et al. (2024) argue that LLMs lack intent regarding the veracity of their outputs and often hallucinate (Huang et al., 2025). Furthermore, it is often difficult for users to recognize hallucinations because LLMs produce fluent and persuasive text. For example, Church (2024) shows that students trust factually incorrect answers from LLMs due to their authoritative and confident tone, and Williams et al. (2025) demonstrate that users rate disinformation generated by LLMs as equally or even more credible than human-generated content. Finally, Suzgun et al. (2025) show that LLMs rely on inconsistent reasoning strategies for epistemic tasks and fail to separate what they believe from what is factually true. Therefore, *we need novel methods to assess the truthfulness of internal probabilistic knowledge to make user interactions with LLMs more reliable.*

Prompt-based methods (Figure 1.A) rely on simply asking an LLM about its knowledge. Abbasi-Yadkori et al. (2024) introduce information-theoretic prompt-based evaluation, while Xu et al. (2024) propose training frameworks to produce prompts with self-reflective rationales, and Farquhar et al. (2024) introduce uncertainty estimators to detect inconsistent text generations. However, prompt-based evaluations are sensitive to the input's phrasing (Lu et al., 2022) and content (McIlroy-Young et al., 2024).

A more direct approach is to examine *how* LLMs represent text internally (Figure 1.B). Consider an LLM $\mathcal{M}$ with a vocabulary $\mathcal{V}$. The LLM $\mathcal{M}$ maps the input text $\boldsymbol{x}$ to a probability distribution over subsequent tokens, denoted $P_{\mathcal{M}}$:

$$\mathcal{M}(\boldsymbol{x}) = P_{\mathcal{M}}(\tau \mid \boldsymbol{x}), \text{ where } \tau \in \mathcal{V}. \quad (1)$$

For any token $\tau \in \mathcal{V}$, the distribution $P_{\mathcal{M}}(\tau \mid \boldsymbol{x})$ denotes the probability that τ is the continuation of the sequence $\boldsymbol{x}$.[3] To compute the conditional distribution $P_{\mathcal{M}}$, an LLM transforms $\boldsymbol{x}$ into *intermediate* neural representations $h_i(\boldsymbol{x}) \in \mathbb{R}^{L \times d}$. Here $h_i(\boldsymbol{x})$ denotes the representation of $\mathcal{M}$ after the ith decoder, d stands for the dimensionality of the decoder, and L stands for the length of the sequence $\boldsymbol{x}$. We can probe these intermediate representations to identify veracity signals, i.e., to isolate

[1] **Code and data**: `carlomarxdk/trilemma-of-truth`.

[2] In this paper, we consider factual associations to be probabilistic knowledge.

[3] Tables A1 and A2, respectively, list the notations and abbreviations used in this paper.

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 120–194
October 24-29, 2026 ©2026 Association for Computational Linguistics

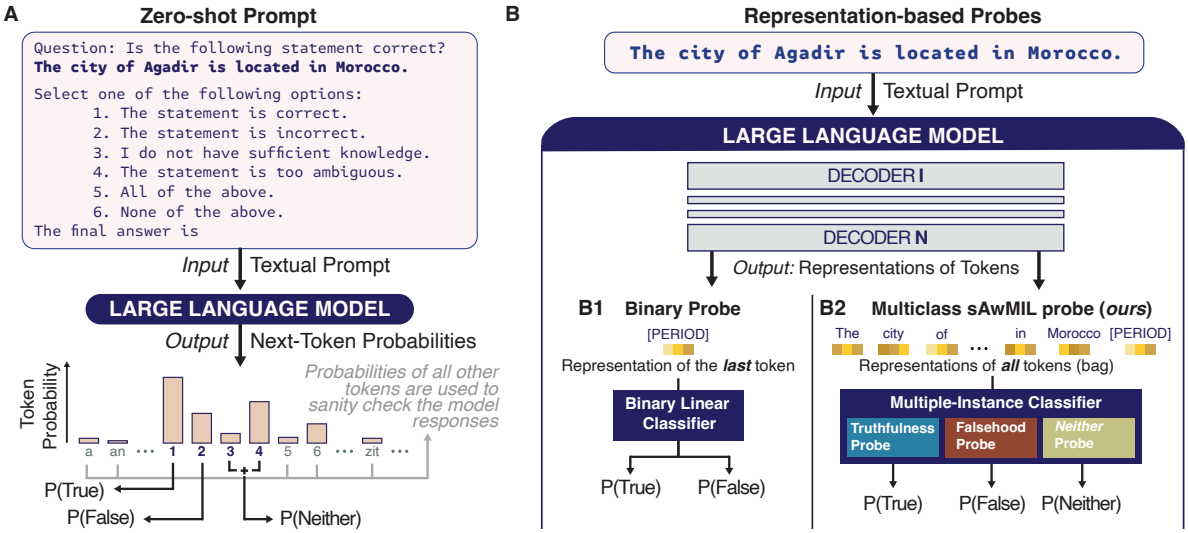

Figure 1: Overview of methods for probing veracity in LLMs. (**A**) In zero-shot prompting, a target statement is inserted into a structured prompt that instructs the LLM to select an answer. This method treats the model as a black box and examines its ⟨input, output⟩ pairs. (**B**) In representation-based probing, we perform analysis on the internal representations generated by intermediate decoders. (**B1**) Commonly used binary probes (e.g., mean-difference) output probabilities for *true* or *false* statements, but cannot account for statements that lack a definitive truth value. (**B2**) Our probe, multiclass sparse-aware MIL (`sAwMIL`), looks at the representation of every token in a statement and provides probabilities for three classes: *true*, *false*, and *neither*. Multiclass `sAwMIL` can account for cases when the LLM does not have any knowledge about the statement.

patterns that correspond to truthful statements. For example, Azaria and Mitchell (2023) train a neural network to classify statements as *true* or *false* based on these internal representations. Similarly, Marks and Tegmark (2024) use a mean-difference classifier to linearly separate *true* from *false* statements (Figure 1.B1). Bürger et al. (2024) introduce the truthfulness-polarity classifier, Burns et al. (2023) propose a semi-supervised method based on the contrastive pairs of statements, and Liu et al. (2026) describe a training-free ensemble approach based on correlations between the representations and truthfulness.

Collectively, these works rely on a shared idea: given statements $\boldsymbol{x}$ with veracity labels $y \in Z$ from a dataset $\mathcal{D}$, train a probe g_i that maps internal representations $h_i(\boldsymbol{x})$ to the distribution $G_{\mathcal{M}}$ over the LLM $\mathcal{M}$'s veracity labels:

$$g_i(h_i(\boldsymbol{x})) = G_{\mathcal{M}}(z \mid \boldsymbol{x}),\ z \in \{\text{true}, \text{false}\} \quad (2)$$

However, we observe that existing probing methods often rely on flawed assumptions, thereby limiting the reliability of current findings. Instead, we argue for a three-valued logic approach (Figure 1.B2) as the more appropriate method to model veracity in LLMs.

Our method `sAwMIL` (short for Sparse-Aware Multiple-Instance Learning) combines Multiple-Instance Learning (Bunescu and Mooney, 2007) and Conformal Prediction (Romano et al., 2020) to allow for a flexible probe that can handle *neither* statements and quantify uncertainty about statements in $\mathrm{K}_{\mathcal{M}}$.

Our contributions include the following: **(1)** We identify and discuss three flawed assumptions in the current veracity-probing literature. **(2)** We show that common linear classifiers do not capture reliable veracity directions. **(3)** We propose a novel multiclass linear-probing method, `sAwMIL`, based on Multiple-Instance Learning (MIL) and Conformal Prediction (CP). **(4)** We present three new datasets containing statements labeled *true*, *false*, and *neither*[4] to enable more rigorous evaluations of veracity probes.

2 Background and Flawed Assumptions

A large language model $\mathcal{M}$ has **internal probabilistic knowledge** $\mathrm{K}_{\mathcal{M}}$, which it acquires during training. Therefore, to determine the veracity of a statement ϕ, the model $\mathcal{M}$ must distinguish between the following three scenarios:

1. ϕ is **True** if there is sufficient support for ϕ given $\mathrm{K}_{\mathcal{M}}$, i.e., $\mathrm{P}(\phi \mid \mathrm{K}_{\mathcal{M}}) \geq \zeta$, where $\zeta \in$

[4]We italicize *neither*, *true*, and *false* to distinguish them from the regular use of these words.

$(0, 1]$ is a threshold.

2. ϕ is **False** if there is sufficient support for $\neg\phi$ given $\mathrm{K}_{\mathcal{M}}$, i.e., $\mathrm{P}(\neg\phi \mid \mathrm{K}_{\mathcal{M}}) \geq \xi$, where $\xi \in (0, 1]$ is a threshold.
3. ϕ is **Neither-true-nor-false** if there is insufficient support for both ϕ and $\neg\phi$ given $\mathrm{K}_{\mathcal{M}}$, i.e., $\mathrm{P}(\phi \mid \mathrm{K}_{\mathcal{M}}) < \zeta$ and $\mathrm{P}(\neg\phi \mid \mathrm{K}_{\mathcal{M}}) < \xi$, where $\xi \in (0, 1]$ is a threshold.

Consequently, if $\mathcal{M}$ has a mechanism to determine the veracity of a statement ϕ, then $\mathcal{M}$ should encode the signal associated with the veracity in its intermediate neural representations:

Truthfulness: $\mathcal{M}$ generates an activation pattern that encodes internal support for ϕ in $\mathrm{K}_{\mathcal{M}}$.

Falsehood: $\mathcal{M}$ produces an activation pattern that reflects a lack of sufficient support for ϕ, instead indicating that the internal knowledge $\mathrm{K}_{\mathcal{M}}$ provides stronger support for $\neg\phi$, signaling a contradiction or misalignment with known information.

Neither: $\mathcal{M}$ encodes the lack of support for ϕ and $\neg\phi$, indicating that the veracity of ϕ is currently undefined.

2.1 Flawed Assumptions When Probing Veracity in LLMs

A veracity probe g_i is trained on labeled datasets. These datasets consist of $\langle h_i(\boldsymbol{x}), y\rangle$ pairs, where $h_i(\boldsymbol{x})$ denotes an internal representation after the ith decoder and $y \in Z$ is the ground-truth veracity label; in most cases, $Z = \{\text{true, false}\}$. We focus on linear probes, where learned parameters define a veracity direction $\boldsymbol{\theta}_i$:

$$g_i(h_i(\boldsymbol{x})) = h_i(\boldsymbol{x})\, \boldsymbol{\theta}_i^\top + \beta_i. \tag{3}$$

Next, we detail three flawed assumptions prevalent in the literature.

Flawed Assumption I: The veracity signal lies along a single axis. Existing veracity probes (Marks and Tegmark, 2024; Burns et al., 2023; Li et al., 2023; Levinstein and Herrmann, 2024) implicitly assume that truth and falsehood are encoded bidirectionally along a single axis in the representation space:

$$P\left(\phi \mid \mathrm{K}_{\mathcal{M}}\right) = 1 - P\left(\neg\phi \mid \mathrm{K}_{\mathcal{M}}\right) \tag{4}$$

Equation (4) implies that truthfulness and falsehood are symmetric and consistent: any layer that encodes a signal for truth must encode an equal and opposite signal for falsehood. However, there is little empirical evidence to support this claim. In contrast, Bürger et al. (2024) suggest that veracity exists along more than one axis.

This assumed binary framing forces all statements into a binary partition: any statement not confirmed as *true* is classified as *false*, and vice versa. Yet an LLM often encounters statements for which its internal probabilistic knowledge $\mathrm{K}_{\mathcal{M}}$ provides insufficient support for *either* ϕ or $\neg\phi$. These include statements about entities absent from training data, contested claims, and underspecified propositions; binary probes cannot represent this third epistemic state and instead assign such statements to *true* or *false*. Figure A1 illustrates the consequence: a binary, single-axis probe (`MD+CP`) cannot carve out a region for the *neither* class, placing such statements inside the high-confidence intervals of *true* and *false* rather than in a space of their own.

Takeaway I: A veracity probe should therefore account for the *neither*[5] statements and should not presuppose that truth and falsehood share a single geometric direction.[6]

Flawed Assumption II: Ground-truth labels are faithful proxies for LLM knowledge. In practice, to train a probe g_i, we use $\mathcal{D}_{train}$ consisting of statement-label pairs $\langle \boldsymbol{x},\, y\rangle$. The labels y are assigned according to *our* understanding of the world: these labels follow a distribution $G_{\mathcal{D}}$ grounded in human knowledge. Ideally, the probe's goal is to recover the veracity signal corresponding to the LLM's internal knowledge $\mathrm{K}_{\mathcal{M}}$, which may follow a different label distribution $G_{\mathcal{M}}$.

This mismatch introduces two associated obstacles. First, we generally do not know what the LLM has learned: the precise composition of the LLM's training data often remains unknown (Longpre et al., 2024), and we lack straightforward methods to verify what has been incorporated into $\mathrm{K}_{\mathcal{M}}$. For instance, *we* know that "The city of Bissau is in Congo" has a ground-truth label $y =$ *false* since we can check maps, but we do not know how $\mathcal{M}$ labels it. The LLM may never have encountered Bissau during training, in which case the statement falls outside $\mathrm{K}_{\mathcal{M}}$ entirely.

Existing probing methods (Burns et al., 2023; Harding, 2023; Li et al., 2023; Bürger et al., 2024; Marks and Tegmark, 2024; Levinstein and Herrmann, 2024) do not account for mismatches between label distributions. Instead, these probes

[5] Cases where $\mathrm{P}(\phi \mid \mathrm{K}_{\mathcal{M}}) < \zeta$ and $\mathrm{P}(\neg\phi \mid \mathrm{K}_{\mathcal{M}}) < \xi$

[6] $\boldsymbol{\theta}_i^{\text{true}} \neq -\boldsymbol{\theta}_i^{\text{false}}$

introduce a systematic bias, in which g_i captures a signal reflecting *our* labeling choices rather than the model's true internal representations. That is, they assume $G_{\mathcal{D}} \overset{d}{=} G_{\mathcal{M}}$.

The second obstacle relates to the predicted labels. Even when the probe's predictions align with the correct labels, the probe's output scores are not necessarily interpretable as degrees of belief. Probes such as SVMs and mean-difference classifiers are based on decision boundaries. They output distances from a separating hyperplane rather than calibrated probabilities. However, Herrmann and Levinstein (2024) argue that veracity probes should provide not only discrete labels but also values that can be interpreted as uncertainty estimates. A common method is to wrap raw scores in a sigmoid function, which forces them into $[0, 1]$. However, it still does not yield calibrated probabilities.

Since we cannot interpret the score in terms of probabilities, it is unclear what these scores represent: a probe that returns a score of 0.5 for a statement ϕ could mean that $\mathcal{M}$ is genuinely uncertain under $\mathrm{K}_{\mathcal{M}}$, or it could mean the probe g_i lacks discriminative power in the representation space. Without calibration, these cases are indistinguishable.

Takeaway II: Existing probes do not provide accurate labels or reliable confidence estimates relative to the LLM's own knowledge $\mathrm{K}_{\mathcal{M}}$. Instead, a probe should (1) account for the possibility that $G_{\mathcal{D}} \overset{d}{\neq} G_{\mathcal{M}}$ and (2) produce calibrated outputs that can be interpreted as the LLM's degree of support for a given statement.

Flawed Assumption III: The veracity signal is localized at the last token. The majority of veracity probes are trained on the representation of the last token (Harding, 2023; Marks and Tegmark, 2024; Li et al., 2023). For example, it is assumed that the period alone carries the entire veracity signal in "Boston is in the US." Thus, the implicit assumption is that any factual signal appearing n tokens before the end of the statement will be faithfully preserved until the last token. A more reasonable strategy is to probe at the exact position where the statement is actualized, e.g., immediately after "in the", rather than relying on the LLM to *move* that signal all the way to the end of the statement.

Figure A2 provides an example: last-token and binary probes assign confident veracity labels to *non-actualized* tokens in the pre-actualized span $\boldsymbol{x}^p$ (e.g., "the", "city", "of"), where no truth value has yet been declared.

Takeaway III: The position of the relevant token is not known a priori. Veracity probes, therefore, must employ a flexible mechanism to identify the optimal token positions for extracting veracity signals, rather than relying on fixed positions.

Summary. A probe that directly addresses these flawed assumptions would more accurately reflect the internal knowledge of the LLM and provide (1) a multiclass veracity signal with distinct directions for *true*, *false*, and *neither* statements, rather than collapsing them onto a single axis; (2) calibrated uncertainty measures of $\mathcal{M}$'s probabilistic knowledge; and (3) a flexible mechanism for identifying informative tokens.

In Section F, we present a case study demonstrating that existing probing methods often misclassify tokens that do not have any associated veracity value, as well as confidently misclassify *neither* statements as *true* or *false*.

3 Method: `sAwMIL`

To address the flawed assumptions, we propose a multiclass probe, called `sAwMIL` (short for Sparse-Aware Multiple-Instance Learning). It classifies statements into three classes: *true*, *false*, and *neither*, and yields statistically guaranteed prediction sets. To do so, `sAwMIL` adopts two methodological frameworks: multiple-instance learning (MIL) and conformal prediction (CP).

3.1 Sparse-Aware Multiple-Instance Learning

Algorithms such as logistic regression, support vector machines, and mean-difference classifiers belong to the single-instance learning (SIL) family, where each instance in the dataset has an individual label. In contrast, multiple-instance learning is a type of weakly supervised learning that operates on a set of labeled *bags* (Carbonneau et al., 2018). A bag B contains a set of related instances (e.g., embeddings of individual words in a sentence). Each bag has an associated binary label, but the labels for individual instances within the bag often remain *unknown*. A bag with a positive label ($y = 1$) indicates that at least one instance in the bag B belongs to the positive class. Similarly, a negative label indicates that none of the instances in the bag B are positive.

Therefore, multiple-instance learning algorithms must consider the overall structure of the bag, sup-

press irrelevant instances, and simultaneously identify a few informative ones. Our proposed algorithm, sAwMIL, is an extension of the multiple-instance SVM proposed by Bunescu and Mooney (2007).

sAwMIL is specifically designed for sparse bags: only a few instances carry a relevant signal. It also allows us to incorporate prior knowledge about which instances are likely to carry a relevant signal. sAwMIL has three training stages (we provide a detailed description of the sAwMIL framework in Section H and Algorithm 1):

Stage 1 (MIL SVM): We start with the MIL adaptation of an SVM that operates on bags. The objective is to identify the most important instances within the positive bags, pushing all other instances and negative bags toward the opposite side of the separating hyperplane.

Stage 2 (Relabeling): Using the fitted SVM from Stage 1, we identify the most informative instances across all positive bags. We compute the distribution of scores assigned to each instance across all positive bags and assign a positive pseudo-label to instances scoring above the $1-\eta$-quantile. The remaining instances are relabeled as negative. Note that η is the *balancing hyperparameter* that encodes our prior assumption about the number of positive instances per bag.

We further exploit the structure of our statements. Consider the following statement: "The city of Riga is in Latvia." We can decompose it into a *pre-actualized* part ($\boldsymbol{x}^p$ = "The city of Riga is in"), which cannot carry any veracity signal, and an *actualized* part ($\boldsymbol{x}^a$ = "Latvia"), where we expect the veracity signal to appear. Formally, each statement decomposes into $\boldsymbol{x} \leftarrow [\boldsymbol{x}^p, \boldsymbol{x}^a]$, and a token is relabeled as positive if it scores above the $(1-\eta)$-quantile *and* belongs to the actualized part.

Stage 3 (SIL SVM): Finally, we train a single-instance SVM on the relabeled instances from Stage 2. Here, we disregard the original grouping into bags and use the instances and their pseudo-labels.

One-vs-all sAwMIL. We train three one-vs-all sAwMIL probes that isolate distinct veracity signals. This design choice allows us to account for Flawed Assumption I:

- is-true probe: separates tokens that carry a *true* signal from all others.
- is-false probe: separates tokens that carry a *false* signal from all others.
- is-neither probe: separates tokens that carry a *neither* signal from all others.

Multiclass sAwMIL. Ideally, we want a multiclass probe that assigns probabilities to a statement being *true*, *false*, or *neither*. Thus, we assemble the one-vs-all sAwMIL probes into a multiclass one via *softmax regression*. It takes the outputs of the one-vs-all probes and transforms them into multiclass probabilities: $p_k = \exp(z_k)/\sum_j \exp(z_j)$, where $z_k = g_i^k(\boldsymbol{x})\,\alpha_k + \beta_k$ and $k \in \{$is-true, is-false, is-neither$\}$. Recall that g_i is the trained probe in Equation (2).

3.2 Conformal Prediction

Raw support-vector-machine outputs (i.e., signed distances from the separating hyperplane) are not meaningful confidence measures. We introduce a conformal prediction framework into sAwMIL to obtain coverage guarantees.

Conformal prediction is a post-hoc method (Romano et al., 2020; Angelopoulos et al., 2021) that transforms raw scores into prediction sets with *guaranteed coverage*. That is, given a new instance $\boldsymbol{x}$ and a predicted score z, CP outputs a prediction set $\mathcal{C}(\boldsymbol{x}) \subseteq Z$ such that the ground-truth label is contained with probability at least $1-\alpha$: $P\big(y \in \mathcal{C}(\boldsymbol{x})\big) \geq 1-\alpha$. Prediction sets that contain multiple labels indicate uncertainty, singleton sets indicate confident predictions, and empty sets indicate that the instance $\boldsymbol{x}$ is too unusual to be classified as any of the classes.

For a detailed description of the nonconformity scores, refer to Johansson et al. (2017) and Section E.

4 Experiments

This section outlines our experimental setup, including datasets, our evaluation procedure, and the set of large language models.

4.1 Data

We introduce three new datasets comprising factually *true*, factually *false*, and *neither* statements. While several benchmarks for veracity and factuality evaluation exist, prior work has shown that some of these may be incorporated into the pre-training or fine-tuning stages of LLMs (Fang et al., 2025). In contrast, our goal is to minimize the risk of data contamination while also maintaining greater control over data provenance and quality.

Dataset	True	False	Neither	Examples
City Locations	A: 1392 N: 1376	A: 1358 N: 1374	A: 876 N: 876	(True) The city of Mâcon is located in France. (False) The city of Dharān is located in Ecuador. (Neither) The city of Staakess is located in Marbate.
Medical Indications	A: 1522 N: 1439	A: 1419 N: 1523	A: 478 N: 522	(True) PR-104 is indicated for the treatment of tumors. (False) Zolpidem is indicated for the treatment of angina. (Neither) Alostat is indicated for the treatment of candigemia.
Word Definitions	A: 1269 N: 1251	A: 1223 N: 1257	A: 1747 N: 1753	(True) Corsage is a synonym of a nosegay. (False) Towner is not a type of a resident. (Neither) Kharter is not a synonym of a greging.

Table 1: Composition of datasets used in this work showing the number of *true*, *false*, and *neither* statements per dataset. **A** stands for the number of affirmative statements, and **N** stands for the number of negated statements. The last column shows example statements with ground-truth labels.

Hence, we assemble new datasets that involve statements related to specific themes (see Table 1 for examples):

- **City Locations** dataset contains statements about cities and their corresponding countries extracted from the `GeoNames` geographical knowledge database (Wick, 2025).
- **Medical Indications** dataset consists of statements about medications and their corresponding indications from the `DrugBank 5.1.12` knowledge base (Wishart et al., 2018).
- **Word Definitions** dataset is based on the WordsAPI (2026) dictionary; these statements involve words and their synonyms or relations.

We deliberately restrict our study to *simple relational statements*. This is a construct-validity choice: to test whether a probe recovers a *veracity* signal, the stimuli must have an unambiguous, verifiable truth value that does not depend on context or compositional structure.

Every dataset consists of negated statements such as "The city of Riga **is not** located in Estonia."[7] and affirmative ones like "Menadione **is** indicated for the treatment of coughs."[8]

***Neither* statements.** If a statement ϕ is absent from the LLM's internal probabilistic knowledge $\mathrm{K}_{\mathcal{M}}$, then ϕ is *neither*, since the truth value cannot be determined at the present moment (due to the lack of information).

It is non-trivial to determine which statements are absent from $\mathrm{K}_{\mathcal{M}}$, because we generally do not have access to the training datasets used to train LLMs. Instead, we create *neither* statements with *synthetic entities*, which occur neither in the real world nor in works of fiction. Since these entities are specifically generated for our experiments, it is highly unlikely that an LLM has learned anything about them during training. Thus, we can use them as *substitutes* for content that LLMs could not have learned: from the point of view of an LLM, these should be considered neither *true* nor *false*.

Refer to Section B.1 for additional details on data provenance and generation of *neither* statements.

4.2 Evaluation

Language Models. We evaluate 16 open-source LLMs (ranging from approximately 3 to 14 billion parameters) across four families: `Gemma/Gemma-2`, `Llama-3` (`v3.1` and `v3.2`), `Mistral-v0.3`, and `Qwen-2.5`. These models run on consumer-grade hardware and are publicly available through Hugging Face (Wolf et al., 2020). We provide a detailed overview of these models in Section C.

Evaluation Metric. We evaluate probes based on their ability to correctly classify statements into three classes: *true*, *false*, and *neither*. We use Matthews correlation coefficient (MCC) to summarize the performance of each probe (on the test sets); refer to Section I for the definition of MCC.

Probing Methods. We compare performance across *six* probing methods, grouped into three categories. In *zero-shot prompting*, we insert each statement into a prompt formulated as a multiple-choice question. Section G provides additional results for the zero-shot evaluation.

We also evaluate three *binary representation-based probes* trained on the representations of the last token: the mean-difference (`MD`) probe introduced by Marks and Tegmark (2024), the `TTPD`

[7]This statement is a factually true and negated statement.

[8]This statement is a factually false and affirmative statement.

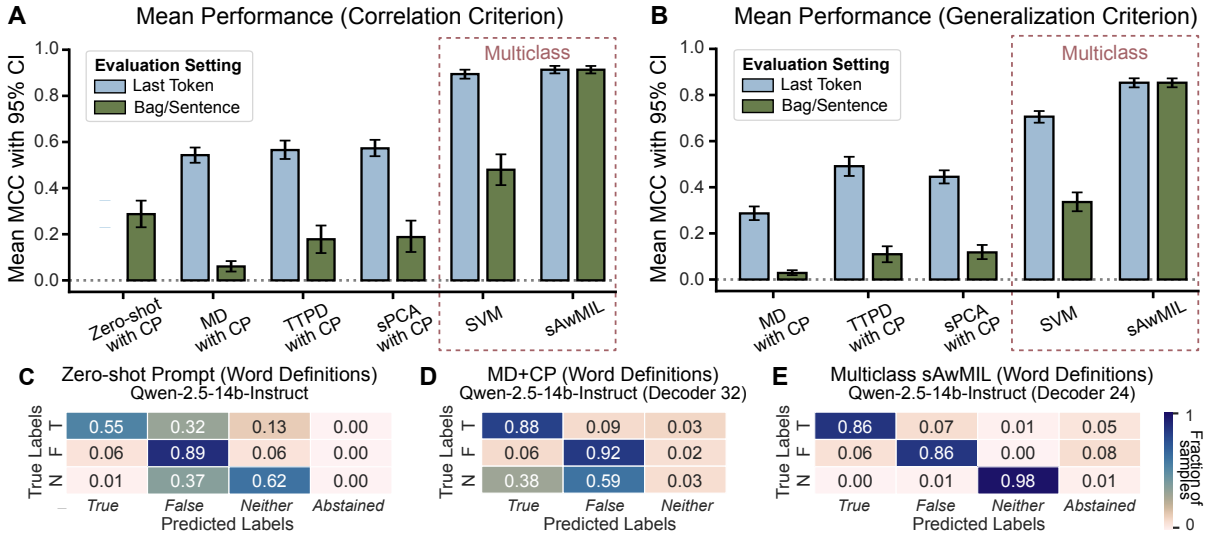

Figure 2: Mean performance of probing methods aggregated across 16 models and three datasets (with 95% confidence intervals), along with examples of confusion matrices. Each probe is evaluated under two settings: using only the representation of the last token and using predictions aggregated across the entire bag/sentence. For each probe, we aggregate the statistics using the best-performing layers. (**A**) *Correlation criterion*. Mean probe performance on the test split of the dataset on which each probe was trained. (**B**) *Generalization criterion*. Mean probe performance on datasets that were not used for training. (**C**) Confusion matrix for zero-shot prompting, where the LLM overpredicts the *false* class. (**D**) Confusion matrix for `MD+CP` evaluated on the last token representation, where the probe incorrectly predicts *neither* statements as *true* or *false*. (**E**) Confusion matrix for multiclass `sAwMIL` evaluated on the full bag. Binary probes trained on the last token representation (`MD+CP`, `TTPD+CP`, and `sPCA+CP`) exhibit both lower performance and poorer generalization. The multiclass `SVM` has comparable performance on the last token. However, performance degrades when evaluated on the full-bag setting; it also achieves lower generalization than `sAwMIL`.

probe (short for Training of Truth and Polarity Direction) introduced by Bürger et al. (2024), and a supervised PCA classifier (`sPCA`). We refer the reader to Algorithms 2 to 4 for details. Since `MD`, `TTPD`, and `sPCA` probes are trained to separate only *true* and *false* statements, we augment each with conformal prediction. Samples whose conformal prediction set does not contain exactly one of *true*/*false* are labeled *neither*. We refer to these methods as `MD+CP`, `TTPD+CP`, and `sPCA+CP`, respectively.

Finally, we include two *multiclass probes*: a multiclass `SVM` trained on the representation of the last token,[9] and the multiclass `sAwMIL` probe, which operates on the whole sentence—all token representations within a statement—and uses a conformal prediction framework.

Validity Criteria. We compare all probes under five criteria: (1) *Correlation*, (2) *Generalization*, (3) *Manipulation*, (4) *Locality*, and (5) *Selectivity*. See Sections I and J for details on the validity criteria. Here we focus on the first three criteria.

[9]Multiclass `SVM` is not equipped with conformal prediction.

The **correlation criterion** (Harding, 2023; Herrmann and Levinstein, 2024) assesses how well a probe g_i, trained on $\mathcal{D}_{train}$, performs on the corresponding $\mathcal{D}_{test}$, assuming that both are drawn from the same distribution. For example, a probe trained on statements about city locations should accurately classify other statements from the same domain. The **generalization criterion** (Bürger et al., 2024) evaluates whether a probe g_i, trained on $\mathcal{D}_{train}$, successfully generalizes to datasets $\mathcal{D}'_{test}$ containing statements from a different domain.

The **manipulation** criterion tests whether the identified direction $\boldsymbol{\theta}_i$ *causally* affects $\mathcal{M}$'s outputs. In addition, the **locality** criterion verifies whether these manipulations predominantly affect the veracity-relevant tokens rather than shifting the distribution of unrelated tokens (Harding, 2023).

Contrastive Token Intervention (CTI). We assess the manipulation and locality criteria by combining activation steering (Li et al., 2023; Todd et al., 2024) with a difference-in-differences (DiD) design. Recall that we have a tokenized statement $\boldsymbol{x} = [\boldsymbol{x}^p, \boldsymbol{x}^a]$ with a pre-actualized part $\boldsymbol{x}^p$ and

an actualized part $\boldsymbol{x}^a$. Further, let $\boldsymbol{\tau}_c$ denote the correct actualized sequence ($\boldsymbol{x}^a = \boldsymbol{\tau}_c$) and $\boldsymbol{\tau}_r$ a length-matched random control ($|\boldsymbol{\tau}_c| = |\boldsymbol{\tau}_r|$). At layer i, we translate the last-token representation of $\boldsymbol{x}^p$ along the veracity direction $\boldsymbol{\theta}_i$:

$$\tilde{h}_i^{\kappa}(\boldsymbol{x}^p) \leftarrow h_i(\boldsymbol{x}^p) + \kappa \cdot \alpha \cdot \boldsymbol{\theta}_i, \tag{5}$$

where $\kappa \in \{-1, +1\}$ is the sign of the intervention and α its dose (Section J). For each target sequence $\boldsymbol{\tau} \in \{\boldsymbol{\tau}_c, \boldsymbol{\tau}_r\}$ and κ, we record the resulting change in the output log-probability $\Delta_{\boldsymbol{\tau}}^{\kappa}$ relative to the unperturbed model. The positive and negative translations, each for the correct and control tokens, result in four measurements (and form the outcome of a 2×2 model):

$$\begin{aligned}\Delta_{\boldsymbol{\tau}}^{\kappa} =& \beta_0 + \beta_1 \cdot \mathbb{1}[\boldsymbol{\tau} = \boldsymbol{\tau}_c] + \beta_2 \cdot \mathbb{1}[\kappa = +] \\ &+ \beta_3 \cdot \mathbb{1}[\boldsymbol{\tau} = \boldsymbol{\tau}_c] \cdot \mathbb{1}[\kappa = +] + \varepsilon\end{aligned} \tag{6}$$

The interaction term β_3 is the veracity-specific effect: the additional change in log-probability on the correct tokens $\boldsymbol{\tau}_c$ after we account for the shift β_2 that the intervention induces on all tokens. A non-significant β_3 means the direction $\boldsymbol{\theta}_i$ has the same effect on correct $\boldsymbol{\tau}_c$ and control tokens $\boldsymbol{\tau}_r$; a significant β_3 meets both criteria: $\boldsymbol{\theta}_i$ causally affects $\mathcal{M}$'s outputs (manipulation) and does so specifically for veracity-relevant tokens (locality). Since the raw β_3 is scale-dependent across ⟨LLM, dataset⟩ configurations, we summarize results for each probe via the *normalized interaction effect* (NIE): $|\beta_3/\hat{\sigma}|$, where $\hat{\sigma}$ is the residual standard deviation of the fit. Thus, NIE is a scale-independent measure of the veracity-specific shift.

We estimate the pooled NIE across the 48 ⟨LLM, dataset⟩ configurations using a DerSimonian–Laird random-effects estimator (Kelley and Kelley, 2023), which accounts for between-configuration heterogeneity. A complementary significance-based *success rate* is reported in Section J. We apply CTI to the four probes (`MD`, `TTPD`, and the `is-true` directions of `SVM` and `sAwMIL`); Section J provides full details.

5 Results

We first look at the performance of probes under the *correlation* and *generalization* criteria. Figure 2 shows the mean performance aggregated across all 16 LLMs and three datasets. For representation-based probes, we report results for two settings: (1) using only the last token's representation, and (2) using all token representations (the full bag).[10]

Binary representation-based probes. The `MD+CP`, `TTPD+CP`, and `sPCA+CP` probes achieve moderate MCC when evaluated on the last-token representation (Figure 2.A). However, their performance drops substantially when evaluated on the full-bag setting, suggesting sensitivity to non-actualized tokens or noise. Under the generalization criterion (Figure 2.B), these probes exhibit limited transferability and degraded performance on the full-bag setting. A common failure, illustrated in Figure 2.D and detailed in Tables A26–A31, is that these probes frequently and confidently misclassify *neither* statements as *true* or *false*. Thus, while effective under controlled conditions, these binary probes appear to identify proxy directions that partially reflect veracity but are confounded by spurious correlations.

Zero-shot prompting. Unlike representation-based probes, prompting does not require information about where a factual statement begins or ends. This makes it more robust when working with unstructured text. However, zero-shot prompting shows a strong preference for a single class: models tend to overpredict one label. For example, in Figure 2.C, Qwen-2.5-14B-Instruct overpredicts the false label under prompting. We see similar skews in the confusion matrices of other models with zero-shot prompting (Table A25).

Multiclass representation-based probes. Both multiclass probes, `SVM` and `sAwMIL`, achieve substantially higher performance (Figure 2.A). However, the `SVM` probe, like its binary counterparts, exhibits a performance drop under the bag setting, indicating reliance on non-generalizable cues. In contrast, the `sAwMIL` probe maintains consistent performance across both settings and yields superior generalization performance (Figure 2.B).

Manipulation & locality. Finally, we test whether the veracity directions causally affect $\mathcal{M}$'s outputs. At the strongest dose, steering along the `is-true` direction from `sAwMIL` produces the largest normalized interaction effect (NIE 0.439; Figure 3), while `MD` achieves the weakest (0.155); NIE increases with dose (Table A9), consistent with a valid directional effect. Notably, `TTPD` nearly matches `sAwMIL` (NIE 0.411). That is, it too lo-

[10] For single-instance probes, when the full bag is provided, the predicted label corresponds to the class with the highest score among all tokens in the statement.

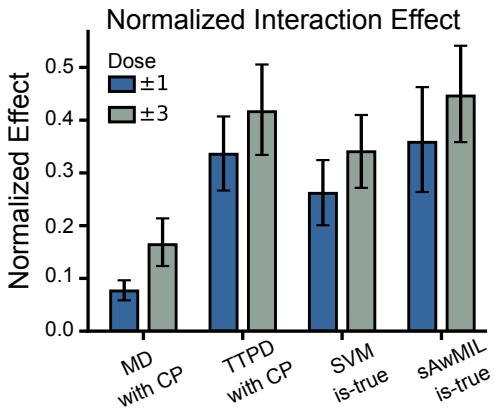

Figure 3: Manipulation & locality evaluation pooled over all 48 ⟨LLM, dataset⟩ configurations for each probe: pooled normalized interaction effect (NIE) adjusted via the DerSimonian–Laird random-effects estimator (Kelley and Kelley, 2023). The error bars are 95% confidence intervals. The normalized effect size quantifies the magnitude of the effect. MD shows the weakest causal effect, followed by the single-instance SVM; TTPD and sAwMIL's is-true achieve similar effect sizes. These results are also presented in Table A9.

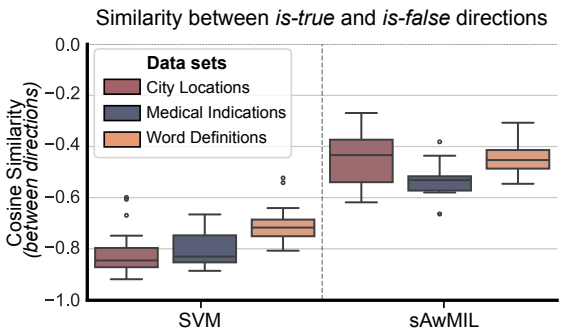

Figure 4: Cosine similarity between the is-true and is-false directions across datasets and probes, using values extracted from the best-performing layer and averaged across 16 LLMs. If the directions were perfectly opposite, their cosine similarity would be -1, indicating a single bidirectional axis of truth and falsehood. Instead, all probes exhibit less opposition, with the better-performing and more generalizable sAwMIL probe yielding higher (less negative) cosine similarity.

cates a causally relevant direction, but TTPD's direction does not satisfy the correlation and generalization criteria, suggesting it captures a causally relevant but non-generalizable (noisy) signal. Together, these findings indicate that sAwMIL is the only probe whose veracity directions are at once robust, transferable, and causally relevant.

Veracity Directions

To show that the truthfulness and falsehood directions are not perfect opposites, we examine the is-true and is-false directions identified by both the multiclass SVM probe and sAwMIL probe.

In Figure 4, we see how different these directions are by computing their cosine similarity. If the two were perfectly opposite, we would expect a cosine similarity of approximately -1. However, this is not the case. The better-performing and more generalizable sAwMIL probe yields directions that are *less opposed*, suggesting that LLMs encode *true* and *false* as related rather than polar concepts.

Similarly, when forming a matrix $\Theta = [\boldsymbol{\theta}^{\text{true}}, \boldsymbol{\theta}^{\text{false}}]$, we find that its rank is 2 for both SVM and sAwMIL, while the effective rank is 1.73 ± 0.012 for SVM and 1.93 ± 0.004 for sAwMIL.[11] If the two directions were linear combinations of one another (i.e., lying on a single axis), the rank would be 1 and the effective rank would be closer to 1.

[11]The number after $\pm$ is standard error.

These findings indicate that the is-true and is-false directions are not strict opposites but span a low-dimensional subspace that captures shared yet distinct representational components.

6 Conclusion

In this work, we critically examine popular methods for probing the veracity of large language models (LLMs) and show that these methods fail to learn reliable and transferable veracity patterns. We argue that probing methods rest on flawed assumptions about how veracity is encoded in LLMs' internal representations.

To address the flaws, we introduce sAwMIL, a multiclass linear probe that combines multiple-instance learning with conformal prediction. Unlike prior methods, sAwMIL models veracity along three axes: *true*, *false*, and *neither*. Across sixteen models and three datasets, sAwMIL outperforms existing probes and shows that truthfulness and falsehood are not represented as simple opposites within LLMs, but as directions spanning a subspace. These findings suggest that veracity-related signals in language models span multiple dimensions rather than a single axis of truthfulness.

Beyond diagnosis, sAwMIL is a practical tool for knowledge auditing and editing. Its calibrated three-way signal indicates whether a given statement is encoded in $\mathcal{M}$ as *true*, *false*, or *neither*, and our intervention experiments indicate that the recovered directions are causally related to $\mathcal{M}$'s outputs.

Limitations

Data and Statements. This study focuses on a specific subset of factual statements, namely those that involve the relation between two entities: city-country, medicine-disease, and noun-noun pairs. As such, it remains unclear how the sAwMIL probe behaves in cases involving multiple relational facts, or in statements describing relations among more than two entities. All statements in our datasets are written in English, and negated statements are limited to the simple "not" construction.

Furthermore, the *neither* statements used in this study are synthetically generated to simulate cases where LLMs lack knowledge. While these proxies enable controlled evaluation, they should be further validated to ensure they accurately capture how models represent unknown, ambiguous, or context-dependent facts. Future work could explore more natural examples from temporal holdout datasets or corpora that reflect human uncertainty.

Evaluation. While our experiments cover 16 LLMs spanning multiple families, our results are limited to models ranging from approximately 3 billion to 14 billion parameters. As such, results might not generalize to much smaller or larger LLMs. Moreover, our study considers only publicly available LLMs and excludes proprietary (closed-source) models. That is, we cannot guarantee that our findings generalize beyond the models that we use.

Our evaluation covers only single-turn scenarios, and we do not explore how probes behave in multi-turn interactions or under varying contexts and instructions. Finally, we consider probes that can only identify linear interactions. It remains an open question whether non-linear probes can achieve better generalization or identify more complex veracity-related patterns.

Ethical Considerations

In this study, we introduce a probing method, sAwMIL, designed to identify veracity-related signals encoded within large language models (LLMs). The primary applications of our approach are auditing and enhanced interpretability of LLMs. Our framework offers a clearer overview of the information captured by LLMs. Additionally, our method can be used to mitigate harmful or incorrect outputs. However, the method may also be exploited by malicious actors to manipulate LLM behavior.

For instance, activation steering along learned veracity directions could bias model outputs toward incorrect or misleading information. Knowledge of where truthfulness signals arise could be exploited to suppress or modify information (e.g., by removing specific knowledge, censoring information, or introducing targeted misinformation).

Importantly, we do not assess the robustness of sAwMIL against adversarial actors and cannot guarantee that the identified veracity directions are resistant to manipulation or obfuscation. Nevertheless, we release our code and data to promote open, reproducible research and encourage future investigations into the adversarial robustness of sAwMIL.

Use of AI

We used Claude and GitHub Copilot for code annotations and formatting assistance. Grammarly was used for assistance with grammar, spelling, and style. Claude and GPT were used for a final proofread of the manuscript.

References

Yasin Abbasi-Yadkori, Ilja Kuzborskij, András György, and Csaba Szepesvári. 2024. To believe or not to believe your LLM: Iterative prompting for estimating epistemic uncertainty. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, volume 37, pages 58077–58117.

Anastasios Nikolas Angelopoulos, Stephen Bates, Michael Jordan, and Jitendra Malik. 2021. Uncertainty sets for image classifiers using conformal prediction. In *Proceedings of International Conference on Learning Representations*.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. 2024. Refusal in language models is mediated by a single direction. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*.

Amos Azaria and Tom Mitchell. 2023. The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976, Singapore. Association for Computational Linguistics.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300.

Steven Bird. 2006. NLTK: The Natural Language Toolkit. In *Proceedings of the COLING/ACL 2006*

Interactive Presentation Sessions, pages 69–72, Sydney, Australia. Association for Computational Linguistics.

Razvan C Bunescu and Raymond J Mooney. 2007. Multiple instance learning for sparse positive bags. In *Proceedings of the 24th International Conference on Machine learning*, pages 105–112.

Lennart Bürger, Fred A Hamprecht, and Boaz Nadler. 2024. Truth is universal: Robust detection of lies in LLMs. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, volume 37, pages 138393–138431.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2023. Discovering latent knowledge in language models without supervision. In *The 11th International Conference on Learning Representations*.

Marc-André Carbonneau, Veronika Cheplygina, Eric Granger, and Ghyslain Gagnon. 2018. Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognition*, 77:329–353.

Alvaro Alonso Casero. 2023. BioBERT fine-tuned for disease NER. Hugging Face repository.

David Chanin, Anthony Hunter, and Oana-Maria Camburu. 2024. Identifying linear relational concepts in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1524–1535, Mexico City, Mexico. Association for Computational Linguistics.

Yuheng Chen, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. 2024. Journey to the center of the knowledge neurons: Discoveries of language-independent knowledge neurons and degenerate knowledge neurons. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17817–17825.

Clément Christophe, Praveen K Kanithi, Tathagata Raha, Shadab Khan, and Marco AF Pimentel. 2024. Med42-v2: A suite of clinical LLMs. Hugging Face checkpoint: m42-health/Llama3-Med42-8B.

Kenneth Church. 2024. Emerging trends: When can users trust GPT, and when should they intervene? *Natural Language Engineering*, 30(2):417–427.

Contact Doctor. 2024. ContactDoctor-Bio-Medical-MultiModal-Llama-3-8B-V1: A high-performance biomedical multimodal LLM. Hugging Face checkpoint: ContactDoctor/Bio-Medical-Llama-3-8B.

Paul Dyson and Jason R. Coombs. 2024. inflect: Correctly generate plurals, ordinals, indefinite articles; convert numbers to words. Python package, version 7.5.0.

Yanai Elazar and Yoav Goldberg. 2018. Adversarial removal of demographic attributes from text data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 11–21, Brussels, Belgium. Association for Computational Linguistics.

Yixiong Fang, Tianran Sun, Yuling Shi, Min Wang, and Xiaodong Gu. 2025. LastingBench: Defend benchmarks against knowledge leakage. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 18304–18317, Suzhou, China. Association for Computational Linguistics.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, and 89 others. 2024a. Gemma: Open models based on gemini research and technology. *arXiv preprint*. Hugging Face checkpoints: google/gemma-7b and google/gemma-7b-it.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, and 179 others. 2024b. Gemma 2: Improving open language models at a practical size. *arXiv preprint*. Hugging Face checkpoints: google/gemma-2-9b and google/gemma-2-9b-it.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The Llama 3 herd of models. *arXiv preprint*. Hugging Face checkpoints: meta-llama/Meta-Llama-3-8B, meta-llama/Llama-3.2-3b, meta-llama/Llama-3.1-8B-Instruct and meta-llama/Llama-3.2-3B-Instruct.

Ramiro Gómez. 2025. geonamescache: GeoNames data for Python. Python package, version 3.0.0.

Jacqueline Harding. 2023. Operationalising representation in natural language processing. *British Journal for the Philosophy of Science*.

Daniel A Herrmann and Benjamin A Levinstein. 2024. Standards for belief representations in LLMs. *Minds and Machines*, 35(1).

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International*

Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 2733–2743, Hong Kong, China. Association for Computational Linguistics.

Michael Townsen Hicks, James Humphries, and Joe Slater. 2024. ChatGPT is bullshit. *Ethics and Information Technology*, 26(2):38.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2).

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *arXiv preprint*. Hugging Face checkpoints: Mistral-7B-Instruct-v0.3 and mistralai/Mistral-7B-v0.3.

Ulf Johansson, Henrik Linusson, Tuve Löfström, and Henrik Boström. 2017. Model-agnostic nonconformity functions for conformal classification. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 2072–2079. IEEE.

Daniel Jurafsky and James H. Martin. 2026. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*, 3rd edition. Online manuscript released January 6, 2026.

George A Kelley and Kristi S Kelley. 2023. Evolution of statistical models for meta-analysis and implications for best practice. *Current Opinion in Epidemiology and Public Health*, 2(3):39–44.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Benjamin A. Levinstein and Daniel A. Herrmann. 2024. Still no lie detector for language models: Probing empirical and conceptual roadblocks. *Philosophical Studies*, 182(7):1539–1565.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: Eliciting truthful answers from a language model. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, volume 36.

Runheng Liu, Heyan Huang, Xingchen Xiao, Yanghao Zhou, and Zhijing Wu. 2026. Training-free truthfulness detection via sparse mlp value vectors. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, pages 3270–3281. ACM.

Shayne Longpre, Robert Mahari, Anthony Chen, Naana Obeng-Marnu, Damien Sileo, William Brannon, Niklas Muennighoff, Nathan Khazam, Jad Kabbara, Kartik Perisetla, Xinyi Wu, Enrico Shippole, Kurt Bollacker, Tongshuang Wu, Luis Villa, Sandy Pentland, and Sara Hooker. 2024. A large-scale audit of dataset licensing and attribution in AI. *Nature Machine Intelligence*, 6(8):975–987.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.

Samuel Marks and Max Tegmark. 2024. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *First Conference on Language Modeling*.

Reid McIlroy-Young, Katrina Brown, Conlan Olson, Linjun Zhang, and Cynthia Dwork. 2024. Order-independence without fine tuning. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, volume 37, pages 72818–72839.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, pages 17359–17372.

Rick Moyer. 2023. namemaker: Random name generator for Python using Markov chains. Python package, version 1.2.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and robust models for biomedical natural language processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy. Association for Computational Linguistics.

Haowen Pan, Xiaozhi Wang, Yixin Cao, Zenglin Shi, Xun Yang, Juanzi Li, and Meng Wang. 2025. Precise localization of memories: A fine-grained neuron-level knowledge editing technique for LLMs. In *The 13th International Conference on Learning Representations*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, and 2 others. 2019. Pytorch: an imperative style, high-performance deep learning library. In *The 33rd International Conference on Neural Information Processing Systems*. Curran Associates Inc.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and

Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Yaniv Romano, Matteo Sesia, and Emmanuel Candes. 2020. Classification with valid and adaptive coverage. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, volume 33, pages 3581–3591.

Robyn Speer. 2022. Python Package: rspeer/wordfreq. Python package, version 3.0.2.

Mirac Suzgun, Tayfun Gur, Federico Bianchi, Daniel E Ho, Thomas Icard, Dan Jurafsky, and James Zou. 2025. Language models cannot reliably distinguish belief from knowledge and fact. *Nature Machine Intelligence*, 7(11):1780–1790.

Eric Todd, Millicent Li, Arnab Sen Sharma, Aaron Mueller, Byron C Wallace, and David Bau. 2024. Function vectors in large language models. In *The 12th International Conference on Learning Representations*.

Marc Wick. 2025. GeoNames geographical database. https://www.geonames.org/. Access date: 2025-03-20.

Angus R Williams, Liam Burke-Moore, Ryan Sze-Yin Chan, Florence E Enock, Federico Nanni, Tvesha Sippy, Yi-Ling Chung, Evelina Gabasova, Kobi Hackenburg, and Jonathan Bright. 2025. Large language models can consistently generate high-quality content for election disinformation operations. *PLOS One*, 20(3):e0317421.

Edwin B Wilson. 1927. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212.

David S Wishart, Yannick D Feunang, An C Guo, Elvis J Lo, Ana Marcu, Jason R Grant, Tanvir Sajed, Daniel Johnson, Carin Li, Zinat Sayeeda, and 1 others. 2018. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic acids research*, 46(D1):D1074–D1082.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

WordsAPI. 2026. WordsAPI: An API for the English language. https://www.wordsapi.com/. Accessed: 2026-01-26.

Tianyang Xu, Shujin Wu, Shizhe Diao, Xiaoze Liu, Xingyao Wang, Yangyi Chen, and Jing Gao. 2024. SaySelf: Teaching LLMs to express confidence with self-reflective rationales. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5985–5998, Miami, Florida, USA. Association for Computational Linguistics.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 23 others. 2024. Qwen2.5 technical report. *arXiv preprint*. Hugging Face checkpoints: Qwen/Qwen2.5-7B, Qwen/Qwen2.5-7B-Instruct, Qwen/Qwen2.5-14B and Qwen/Qwen2.5-14B-Instruct.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, and 2 others. 2023. Representation engineering: A top-down approach to AI transparency. *arXiv preprint*.

Appendix

A Notations and Abbreviations **15**

B Datasets **17**
B.1 *Neither* Statements . . . 17
B.2 Data Selection and Processing . . . 17

C Selection of Large Language Models **20**

D Criteria for Validating a Veracity Probe **21**

E Conformal Prediction Sets **23**

F More on Representation-based Probing Methods **24**
F.1 Mean-difference Probe with Conformal Prediction . . . 24
F.2 Multiclass Single-Instance Support Vector Machine . . . 25

G Zero-Shot Prompting: Instructions, Veracity Labeling, and Abstention **29**

H Technical details for `sAwMIL` **34**
H.1 Single-Instance SVM . . . 34
H.2 The Multiple-Instance Setting . . . 34
H.3 Dual Formulation . . . 36

I Additional Evaluation Details: Correlation, Selectivity and Generalization **37**
I.1 Evaluation Setup . . . 37
I.2 Generalization Across Datasets . . . 37
I.3 Recap: Overall Validity . . . 37

J Additional Evaluation Details: Causal Validation via Steering **40**
J.1 Setup . . . 41
J.2 Results . . . 41

K Algorithms **44**

L More Tables on Classification Performance, Generalization Performance, and Confusion Matrices **51**

M Related Resources **75**

A Notations and Abbreviations

Symbol	Description	Shape / Notes
$\mathcal{M}$	Large language model (LLM)	
$\mathrm{K}_{\mathcal{M}}$	Internal probabilistic knowledge of the model $\mathcal{M}$	
$\mathcal{V}$	Vocabulary of the model $\mathcal{M}$, consists of tokens	$[\tau_1, \ldots, \tau_{\lvert\mathcal{V}\rvert}] \in \mathcal{V}$
$P_{\mathcal{M}}(\tau \mid \boldsymbol{x})$	Output of $\mathcal{M}$: a conditional probability distribution on tokens	
	Inputs and Datasets	
$\mathcal{D}$	Dataset of statements	$\mathcal{D} = \{\mathcal{D}_{\text{train}}, \mathcal{D}_{\text{test}}, \mathcal{D}_{\text{cal}}\}$
$\boldsymbol{x}$	Input token sequence, e.g., "The city of Riga is in Latvia."	$L = \lvert\boldsymbol{x}\rvert$
$\boldsymbol{x}^p$	Pre-actualized part of a statement, e.g., "The city of Riga is in"	
$\boldsymbol{x}^a$	Actualized part of a statement, e.g., "Latvia."	
B	Bag of token representations (contains related instances)	
y	Veracity label assigned to $\boldsymbol{x}$	$y \in Z$
ϕ	A statement evaluated for truthfulness/veracity	
$\mathcal{T}_S$	Transition matrix for n-gram generation	See Equation (A1)
	Internal Representations	
d	Size of the hidden representation (of a decoder)	
$h_i(\boldsymbol{x})$	Activations after the ith decoder	$\mathbb{R}^{L \times d}$
$h_i(\boldsymbol{x})_{[j]}$	Activation of the token at index j after ith decoder	$\mathbb{R}^{1 \times d}$ and $j \in \{1 \ldots L\}$
$h_i(\boldsymbol{x})_{[a:b]}$	Activations from ath to bth tokens after ith decoder	$\mathbb{R}^{(b-a+1) \times d}$
	Veracity, Probes and Distributions	
Z	Set of veracity labels, e.g., $\{$*true*, *false*, *neither*$\}$	
$G_{\mathcal{D}}(z \mid \boldsymbol{x})$	Distribution of veracity labels $z \in Z$ in a dataset $\mathcal{D}$	
$G_{\mathcal{M}}(z \mid \boldsymbol{x})$	Distribution of veracity labels $z \in Z$ in the model $\mathcal{M}$	
g_i	Veracity probe trained on representations of the ith decoder	$g_i : h_i(\boldsymbol{x}) \mapsto G_{\mathcal{M}}$
$\boldsymbol{\theta}_i$	Linear veracity direction extracted from the probe g_i	$\mathbb{R}^{1 \times d}$
η	Balancing hyperparameter for `sAwMIL`	$\eta \in (0, 1]$
$\boldsymbol{m}$	`sAwMIL`'s intra-bag labels (i.e., per-token labels in each $\boldsymbol{x}$)	$\boldsymbol{m} \in \{0, 1\}^L$, $L = \lvert\boldsymbol{x}\rvert$
$\mathcal{C}(\boldsymbol{x})$	Prediction set (related to conformal prediction) $\mathcal{C}(\boldsymbol{x}) \subseteq Z$	
$P(y \in \mathcal{C}(\boldsymbol{x}))$	Probability that label y is part of the prediction set	
	Interventions via Activation Steering (Section J)	
κ	Sign for the intervention direction	$\kappa \in \{-1, +1\}$
α	Intervention dose	$\alpha = d' \cdot \sigma,\ d' \in \{1, 3\}$
d'	Dose scaler	$d' \in \{1, 3\}$
σ	Standard deviation of the scores provided by g_i (based on $\mathcal{D}_{cal}$)	See Equation (A28)
$\tilde{h}_i^{\kappa}(\boldsymbol{x}^p)$	Steered hidden representation at layer i (along $\alpha \cdot \kappa \cdot \boldsymbol{\theta}_i$)	
$\boldsymbol{\tau}_c$	Actualized sequence (containing correct tokens)	$\boldsymbol{x}^a = \boldsymbol{\tau}_c$
$\boldsymbol{\tau}_r$	Actualized sequence (contains random control tokens)	$\lvert\boldsymbol{\tau}_r\rvert = \lvert\boldsymbol{\tau}_c\rvert$
$\Delta_{\boldsymbol{\tau}}^{\kappa}$	Log-probability shift after intervention (along $\alpha \cdot \kappa \cdot \boldsymbol{\theta}_i$)	See Equation (A27)
$\beta_0, \beta_1, \beta_2, \beta_3$	DiD regression coefficients	See Equation (A29)
$\hat{\sigma}$	Residual standard deviation of the DiD fit	See Equation (A30)
τ^2	Between-configuration heterogeneity (DerSimonian–Laird estimator)	

Table A1: Notations used throughout the paper. Symbols are grouped by category: model definitions, inputs and datasets, internal representations, veracity distributions, and intervention-related symbols.

Abbreviation	Full Form	Description
LLM	Large Language Model	
SIL	Single-Instance Learning	Probes trained on single instances
MIL	Multiple-Instance Learning	Probes trained on bags (collections of instances)
SVM	Support Vector Machine	Type of a classifier
CP	Conformal Prediction	Uncertainty calibration method
`MD+CP`	Mean-Difference with conformal prediction	It classifies *neither* via abstention
`TTPD+CP`	Training of Truth and Polarity Direction (probe) with CP	It classifies *neither* via abstention
`sPCA+CP`	Supervised PCA (probe) with conformal prediction	It classifies *neither* via abstention
`sAwMIL`	Sparse-Aware MIL probe	Multiclass probe handling unknowns
MCC	Matthews Correlation Coefficient	Multiclass performance measure in eq. (A25)
NER	Named Entity Recognition	See Section B
DiD	Difference-in-differences design	See Section J
CTI	Contrastive Token Intervention	See Section J
NIE	Normalized Interaction Effect (related to CTI)	See Section J
SR	Success rate (related to CTI)	See Section J

Table A2: Abbreviations and naming conventions used throughout this paper.

B Datasets

We introduce three new datasets: *City Locations*, *Medical Indications*, and *Word Definitions*. Each dataset consists of statements that are factually *true*, factually *false*, or *neither*. These datasets contain both affirmative and negated statements. The datasets are publicly available on Hugging Face at `carlomarxx/trilemma-of-truth` (DOI: `10.57967/hf/5900`).

An example of a false negated statement is "Guaifenesin is **not** indicated for the treatment of coughs", and an example of the true affirmative statement is "Shouter is a type of a communicator."

Data Splits. We split each dataset into train, calibration, and test sets using *approximately* 55/20/25 percentage splits (see Table A3). We ensure that the objects mentioned in statements are exclusive to the split. For example, if Singapore is mentioned in a statement in the training set, all statements containing Singapore are moved to the training split.

B.1 *Neither* Statements

Since we do not have access to LLM training datasets, we cannot validate whether LLMs retain information about specific facts or entities. To overcome this issue, we create *neither* statements that contain *synthetic entities*: entities that do not exist in the real world or fictional works. The *neither* statements are the ones whose value cannot be determined at present (e.g., due to lack of information).

Generation of *Neither* Statements

We use synthetic names to generate *neither* statements. For example, "The city of *Staakess* is located in *Soldovadago*" mentions a town and a country that do not exist. From the point of view of an LLM, these statements should be considered *neither-true-nor-false*, as LLMs could not have captured anything about these in $\mathrm{K}_{\mathcal{M}}$.

To generate the *neither* statements, we use the Markov Chain technique (Jurafsky and Martin, 2026). Given a set of existing words $[w_1, w_2 \ldots, w_n] \in S$, we break each word w_i into n-grams. For instance, we break "ability" into the following 2-grams: `[start]a ab bi il li it ty y[end]`. We then compute a transition matrix $\mathcal{T}_S$, which provides the probability of transitioning from the n-gram i to n-gram j:

$$\mathcal{T}_S\left(j \mid i\right) = \frac{\text{count}\left(i \rightarrow j\right)}{\sum_x \text{count}\left(i \rightarrow x\right)} \tag{A1}$$

We use $\mathcal{T}_S$ to sample new synthetic words that follow the n-gram distribution of words in S. We generate the following types of synthetic entities: city names, country names, drug/medicine names, disease/indication names, and English nouns.

In our experiments, we use 3-grams for most entities, except for country names, which we generate with 2-grams. We use the `namemaker` (Moyer, 2023) package that implements a Markov chain word generator.

B.2 Data Selection and Processing

Next, we provide details on the source and processing steps for each dataset.

City Locations

The *City Locations* dataset is based on the `GeoNames` database (Wick, 2025); `GeoNamesCache` (Gómez, 2025) is a Python package that interacts with the `GeoNames` API (accessed under the CC BY 4.0 license). We use the following criteria to select a ⟨city, country⟩ pair:

1. The population of the city is at least 30,000.
2. The city is associated with a country.
3. If a city name is associated with multiple countries, we include the ⟨city, country⟩ pair for each country. We exclude all cities that have "Antarctica" as a location or a country.

Since the resulting set of ⟨city, correct country⟩ pairs is relatively large, we reduce the number of pairs by downsampling. In total, we select 1,400 unique city names: 700 cities with the highest populations, and 700 cities randomly sampled from the rest of the names.

Statement Structure. For each ⟨city, country⟩ pair, we create statements of the form:

> The city of `[city]` is (not) located in `[country]`.

If a city name already contains the word "City" (e.g., "Guatemala City"), we do not start a sentence with "The city of." Finally, we sample ⟨city, incorrect country⟩ pairs, and generate statements according to the template above.

Synthetic Entities. We use the technique described in Section B.1 to generate synthetic city and country names.

To generate synthetic *city* names, we collect all the city names in our dataset (including those that

Dataset	Train	Calibration	Test	Total
City Locations	3999 (.55)	1398 (.19)	1855 (.26)	7252 (1.00)
Medical Indications	3849 (.56)	1327 (.19)	1727 (.25)	6903 (1.00)
Word Definitions	4717 (.55)	1628 (.19)	2155 (.25)	8500 (1.00)

Table A3: **Dataset splits.** The number of statements per split. In the parentheses, we specify the fraction of the total number of statements.

we did not include) and input them to `namemaker` (with n-gram length of 3). We generate 500 synthetic city names. We validate these synthetic names in two stages:

1. We check whether a synthetic name exists in the `GeoNames` database by looking for matches across *name* and *alternative name* fields. We keep 310 cities after this first stage.

2. We use Google Search to validate that a synthetic city name does not exist via the following prompt: "city [city name]". If the search result returns a city with 1-2 character difference, we remove the synthetic name from the list. We keep 219 cities after this second stage.

For the synthetic country names, we collect all the country names and input them to `namemaker` (with n-gram length of 2). We generate 250 synthetic names and validate them using the workflow described in the previous paragraph. We keep 238 country names after the first stage, and 138 after the second stage. The Google Search prompt is: "country [country name]".

Finally, with probability 25%, we append a prefix or suffix to the synthetic country name. The list of prefixes and suffixes includes "Island," "Republic of," "Kingdom," "West," "East," "North," "South," and "Land." Finally, we randomly assign each synthetic city name to a synthetic country name.

Medical Indications

The *Medical Indications* dataset is based on the `DrugBank` (version 5.1.12) (Wishart et al., 2018). We obtained access to the `DrugBank` on October 4th, 2024, via the academic license (for research purposes only). Our code repository does not contain the raw data from the `DrugBank`, but the reader can apply for the academic license.[12] We extract 2 fields from this knowledge base:

1. **Name**, which specifies the official name of the drug or the chemical (e.g., Lepirudin).

2. **Indication**, which is a text field that describes the indication of the drug. If this field consists of multiple sentences, we keep only the first sentence (e.g., "Lepirudin is indicated for anticoagulation in adult patients with acute coronary syndromes (ACS) such as unstable angina and acute myocardial infarction without ST elevation.")

To extract diseases and conditions from the *Indications* field, we use two named entity recognition (NER) models:

1. `SciSpaCy`'s `en_ner_bc5cdr_md` model for the biomedical term annotations (Neumann et al., 2019),

2. `BioBERT`-based NER (Casero, 2023; Lee et al., 2019) for disease annotations.

We input the "Indication" text to both models. The disease/condition terms are extracted only if both models mark it as a disease or condition. For example, for "Lepirudin is indicated for anticoagulation in adult patients with acute coronary syndromes (ACS)," the `SciSpaCy` model marks *coronary syndromes* as a disease, but `BioBERT` does not. Thus, we do not add it to Lepirudin's disease/condition list. Similarly, we remove the abbreviation if the disease list contains the full name *and* its abbreviation, such as [acute coronary syndromes, ACS].

Name validation. We further validate the drug names using the `SciSpaCy` model and retain only those marked as `CHEMICAL`. Otherwise, we remove the drug from our dataset. Finally, if the disease list (for a given drug) is empty after the preprocessing, we remove the drug from our dataset.

Additionally, we use the `wordfreq` package (Speer, 2022) to check whether the name of the drug or the name of the indication appears in widely used corpora (e.g., Wikipedia or books dataset). In other words, we remove the pair if either the drug name or the indication has a Zipf frequency of

[12] Here is the link to the `DrugBank`'s academic license: https://go.drugbank.com/releases/5-1-12

0 – i.e., the word does not appear in any of the `wordfreq` corpora.

Statement Structure. For each ⟨drug, correct disease⟩ pair, we create statements of the form:

> `[drug]` is (not) indicated for the treatment of `[disease/condition]`.

We also sample the ⟨drug, incorrect disease⟩ pairs. We ensure that the "incorrect disease" did not share any words with the diseases in the correct list.

Synthetic Entities. To generate synthetic drug names and disease names, we use the approach described in Section B.1 (with n-gram length of 3). We generate 500 synthetic drug names. We validate these synthetic names in two stages:

1. We pass each generated name through the `SciSpaCy` model and remove the ones marked as `CHEMICAL`. We keep 315 names after this first stage.
2. We use Google Search to validate that each drug name does not exist via the prompt "medicine [drug name]." If the search result returned a drug with 1-2 character difference, we remove it from the dataset. We keep 243 names after this second stage.

We generate 200 disease names and check whether they exist in our list of diseases. We keep 181 names after this first stage. Next, we use Google Search with the prompt "disease [disease/condition name]." We keep 131 disease names after this second stage. Finally, we randomly match synthetic drug names to synthetic disease names to generate *neither* statements.

Word Definitions

The *Word Definitions* dataset is based on a publicly available sample from WordsAPI (2026) database. It contains 10% of randomly sampled words from the database.

For each word in the sample, we keep the ones that satisfy the following criteria:

1. The word is a noun.
2. The word has at least one definition in the *definition* field.
3. The word has at least one of the following fields: *synonym*, *typeOf*, or *instanceOf*.

Statement Structure. Depending on the specified field (i.e., *synonym*, *typeOf*, *instanceOf*), we generate three types of statements:

1. "`[word]` is (not) `[instanceOf]`."
2. "`[word]` is (not) a type of `[typeOf]`."
3. "`[word]` is (not) a synonym of `[synonym]`."

Before inserting a word from *synonym*, *typeOf*, *instanceOf* fields into a corresponding spot, we determine whether the appropriate indefinite article is 'a' or 'an.' When possible, we change words into singular forms. To do so, we use the `inflect` package (Dyson and Coombs, 2024).

Synthetic Entities. To generate synthetic entities, we use the approach described in Supplementary Sec. B.1 (with n-gram length of 3). We generate four categories of synthetic entities:

1. Words that go at the beginning of each statement: We use all the words we have in the dataset.
2. Types: We use all the words from the *typeOf* field for the Markov Chain generation.
3. Synonyms: We use all the words from the `synonym` field.
4. Instances: We use words from the *instanceOf* field.

We generate 1,000 synthetic words for each of the four categories. We validate the non-existence of words. We use the `english_words` package[13] to check whether a word exists in "GNU Collaborative International Dictionary of English 0.53," or `web2` word list. Furthermore, we check whether there is a word in the *words* list of the `nltk` package (Bird, 2006). After this stage, we end up with 3,305 words. Finally, we randomly sample pairs of ⟨word, property⟩, where the property is a type, instance, or synonym.

[13] pypi.org/project/english-words/

C Selection of Large Language Models

In this section, we provide an overview of the large language models used in our experiments. Table A4 provides a list of all 16 models.

We use default models (pre-trained on general tasks) and chat models fine-tuned on instruction- and chat-like interactions. Every default model in our selection has a corresponding chat-based model. We add two additional chat-tuned `Llama` models, each fine-tuned on biomedical data. Furthermore, we do not use the full official model names but use short names with a version, such as "chat" or "default". For example, `Llama-3.2` (chat) refers to the `Llama-3.2-3b-Instruct` model. Finally, we list the Hugging Face repositories in the notes for each reference.

Official Model Name	Type	# Decoders	Size	Release Date	Source
Gemma-7b	Default	28	8.54 B	Feb 21, 2024	Gemma Team et al. (2024a)
Gemma-2-9b	Default	26	9.24 B	Jun 27, 2024	Gemma Team et al. (2024b)
Llama-3-8b	Default	32	8.03 B	Apr 18, 2024	Meta (Grattafiori et al., 2024)
Llama-3.2-3b	Default	28	3.21 B	Sep 25, 2024	Meta (Grattafiori et al., 2024)
Mistral-7B-v0.3	Default	32	7.25 B	May 22, 2024	Mistral AI (Jiang et al., 2023)
Qwen2.5-7B	Default	28	7.62 B	Sep 19, 2024	Alibaba Cloud (Yang et al., 2024)
Qwen2.5-14B	Default	38	14.80 B	Sep 19, 2024	Alibaba Cloud (Yang et al., 2024)
Gemma-7b-it	Chat	28	8.54 B	Feb 21, 2024	Gemma Team et al. (2024a)
Gemma-2-9b-it	Chat	26	9.24 B	Jun 27, 2024	Gemma Team et al. (2024b)
Llama-3.2-3b-Instruct	Chat	28	3.21 B	Sep 25, 2024	Meta (Grattafiori et al., 2024)
Llama-3.1-8b-Instruct	Chat	32	8.03 B	Jul 23, 2024	Meta (Grattafiori et al., 2024)
Llama3-Med42-8B	Chat	32	8.03 B	Jul 2, 2024	Christophe et al. (2024)
Bio-Medical-Llama-3-8B	Chat	32	8.03 B	Aug 8, 2024	Contact Doctor (2024)
Mistral-7B-Instruct-v0.3	Chat	32	7.25 B	May 22, 2024	Mistral AI (Jiang et al., 2023)
Qwen2.5-7B-Instruct	Chat	28	7.62 B	Sep 19, 2024	Alibaba Cloud (Yang et al., 2024)
Qwen2.5-14B-Instruct	Chat	38	14.80 B	Sep 19, 2024	Alibaba Cloud (Yang et al., 2024)

Table A4: **List of LLMs used in our experiments.** We provide the official names of the models in the Hugging Face repository. Furthermore, we specify the model's *type*: 'default' denotes the pre-trained models, and 'chat' denotes the chat- or instruction-tuned versions. Finally, we provide the number of decoders, the number of parameters, the release date, and the source of the model. These models are publicly available through Hugging Face (Wolf et al., 2020).

D Criteria for Validating a Veracity Probe

Herrmann and Levinstein (2024), Zou et al. (2023) and Harding (2023) have proposed criteria for measuring the validity of veracity probes. We aggregate these into five major categories and provide an overview in Table A5. We propose to evaluate a probe g_i along the following criteria:

(i) **Correlation**. The probe, trained to predict a veracity property Z, should achieve high predictive accuracy on unseen samples that possess this property, i.e., on samples from $\mathcal{D}_{test}$. When the criterion is satisfied, the ith decoder embeds the information about Z to some degree. We cannot rule out the fact that it captures proxies associated with Z.

(ii) **Generalization** extends *Correlation* by requiring that the probe generalizes beyond the dataset it was trained on. The probe should achieve high predictive accuracy on samples with veracity Z, but have different phrasing or come from different domains. For example, if a probe is trained on representations associated with veracity for ecological statements, it should exhibit similar predictive accuracy on geographical statements.

(iii) **Selectivity**. The probe g_i should avoid classifying statements that are not (or cannot be) *true* or *false*. Instead, the probe g_i should abstain from making predictions on the *neither* statements—i.e., statements that the LLM could not have learned from its training data or that inherently lack any truthfulness or falsehood. Poor selectivity indicates that the probe might capture spurious correlations.

(iv) **Manipulation**. We should be able to use the identified direction $\boldsymbol{\theta}_i$ to update $h_i(\boldsymbol{x})$ and have a predictable change in the distribution of the output tokens $P_{\mathcal{M}}(\tau \mid \boldsymbol{x})$.

(v) **Locality**. When asking a question such as "Is X true? Answer yes or no," the manipulation should primarily influence the probabilities related to the tokens "yes" or "no". It should minimally affect unrelated tokens. For example, if a manipulation does not increase the likelihood of the LLM generating "no," but increases the likelihood of generating tokens such as "elephant," then the identified veracity direction affects not only veracity-related signals but also unrelated ones.

Evaluating a probe according to these criteria allows us to determine how well g_i captures the signals associated with veracity Z and how manipulations (a.k.a. interventions) affect the LLM's output $P_{\mathcal{M}}$. Part of our future work includes adding a new criterion to assess whether a probe can determine whether an LLM can "reason" logically. For example, if $\mathcal{M}$ classifies a statement ϕ_1 as true and another statement ϕ_2 as true, then will $\mathcal{M}$ also classify $\phi_1 \wedge \phi_2$ as true?

Results. We evaluate all five criteria across our experiments. Results for **Correlation**, **Selectivity**[14] and **Generalization** are reported in Figure 2 and additional details are provided in Section I. The results for **Manipulation** and **Locality** are reported in Figure 3 and additional details are provided in Section J.

[14]The multiclass MCC allows us to evaluate both the correlation and selectivity criteria at the same time.

Criteria	Definition	If Satisfied	Similar Concepts
Correlation	A probe g_i trained on $\langle h_i(\boldsymbol{x}), y\rangle \in \mathcal{D}_{\text{train}}$ should perform well (i.e., have high predictive accuracy) on $\mathcal{D}_{\text{test}}$, assuming the same input and label distributions.	$\mathcal{M}$ encodes information correlated with veracity; for the result refer to Figure 2.A.	Information (Harding, 2023), Accuracy (Herrmann and Levinstein, 2024)
Generalization	A probe g_i trained on $\langle h_i(\boldsymbol{x}), y\rangle \in \mathcal{D}_{\text{train}}$ should have high predictive accuracy on data from different domains.	$\mathcal{M}$ has a universal activation pattern correlated with veracity; for the results refer to Figure 2.B and Figure A9.	Generalization as defined by Bürger et al. (2024), Uniformity (Herrmann and Levinstein, 2024)
Selectivity	A probe g_i trained on $\langle h_i(\boldsymbol{x}), y\rangle \in \mathcal{D}_{\text{train}}$ should not assign *true* or *false* labels to samples where truthfulness is absent or undefined.	$\mathcal{M}$ has a distinct mechanism that correlates exclusively with veracity.	Misrepresentation as defined by Harding (2023), Control Task (Hewitt and Liang, 2019)
Manipulation	Modifying $h_i(\boldsymbol{x})$ along $\boldsymbol{\theta}_i$ should systematically alter $P_{\mathcal{M}}(\tau \mid \boldsymbol{x})$ for tokens τ related to the veracity property Z.	$\mathcal{M}$ has a *linear* mechanism to track veracity and uses it to compute the output $P_{\mathcal{M}}(\tau \mid \boldsymbol{x})$.	Use (Harding, 2023), Addition (Arditi et al., 2024), Intervention (Meng et al., 2022)
Locality	Modifying $h_i(\boldsymbol{x})$ along $\boldsymbol{\theta}_i$ should not significantly alter $P_{\mathcal{M}}(r \mid \boldsymbol{x})$ for random tokens r that are unrelated to Z.	$\mathcal{M}$ maintains a separate mechanism that tracks veracity, without being confused with other concepts.	Misrepresentation (Harding, 2023), Leakage (Elazar and Goldberg, 2018)

Table A5: **Validity criteria for representation-based probes.** If satisfied, these criteria serve as validation that g_i indeed captures signals associated with veracity Z. Here, we provide a formal definition of each criterion, along with the implications of satisfying the criterion. Finally, we provide the list of similar criteria and concepts used in the literature.

E Conformal Prediction Sets

In our work, we focus on *split conformal learning* (Angelopoulos et al., 2021), which requires a hold-out (or calibration) dataset to compute conformal prediction sets. Given a probe g_i, we use a calibration dataset $\mathcal{D}_{cal}$ of activation-label pairs $\langle h_i(\boldsymbol{x}), y\rangle$ to find prediction sets $C(\boldsymbol{x}) \subset \mathcal{Z}$ with a marginal coverage guarantee; assuming the calibration and test data are exchangeable, $C(\boldsymbol{x})$ contains the correct label with probability at least $1-\alpha$. Given $\alpha = 0.1$, the correct label is covered at least 90% of the time *on average over the data*. We provide pseudocode for the nonconformity functions in Algorithms 5 and 6.

Nonconformity scores. To identify the conformal sets, we compute a nonconformity score for each sample in $\mathcal{D}_{cal}$. For the binary cases (such as mean-difference probe and one-vs-all `sAwMIL`), we use the binary nonconformity scoring; see Algorithm 5. It is based on the distance between the prediction and the classifier's separating hyperplane. In Equation (A2), s is the signed distance of the sample to the separation hyperplane, and y is a ground-truth (or candidate) label:[15]

$$\text{binaryNC}\,(s, y) = \exp\big(-y \cdot s\big), \tag{A2}$$

where $y \in \{-1, 1\}$ and $s \in \mathbb{R}$. If the sample ends up on the wrong side of the separation hyperplane (e.g., $s > 0$ and $y = -1$), then the nonconformity score in Equation (A2) is high and the candidate label is weakly supported by the model.

For the multiclass `sAwMIL`, we use the multiclass nonconformity score (Johansson et al., 2017); see Algorithm 6. For a given candidate label y, the score is defined in terms of the difference between the predicted probability of the true class and the highest probability among the other classes (with K denoting the total number of classes). Formally, for a candidate label y with predicted probability p_y, we calculate the multiclass nonconformity score with the following function:

$$\text{multiNC}\,(\mathbf{p}, y) = \frac{1 - (p_y - \max_{i \neq y} p_i)}{2}, \tag{A3}$$

where $\mathbf{p} \in \Delta^{K-1}$ is the predicted probability vector, $\Delta^{K-1} := \{\mathbf{p} \in \mathbb{R}^K \mid p_i \geq 0, \sum_i p_i = 1\}$, and p_y denotes the predicted probability assigned to the candidate label y.

[15] In this case, labels should be either -1 or 1. Thus, all samples with label 0 are assigned label -1.

In both cases, lower scores in Equations (A2) and (A3) indicate that the candidate label y is strongly supported by the model. For each candidate label y, we compute its conformal p-value ψ_y from the nonconformity scores and include y in the conformal prediction set C if $\psi_y > \alpha$, where $\alpha = 0.1$. The probe abstains when the resulting prediction set is empty or contains more than one label (see Algorithms 5 and 6). The addition of conformal predictions allows us to distinguish between cases where the statements originate from different distributions, as compared to those in the calibration dataset.

F More on Representation-based Probing Methods

F.1 Mean-difference Probe with Conformal Prediction

The mean-difference probe (MD+CP) consists of two components: a binary mean-difference classifier (MD) and the conformal prediction (CP).

First, we fit the binary classifier with a linear decision boundary (Marks and Tegmark, 2024). We use it to separate *true* and *false* statements based on the internal representations h_i. For each pair $\langle \boldsymbol{x}_j, y_j \rangle$, we extract the activation of the last token $h_i(\boldsymbol{x}_j)_{[L]}$ and assemble a set of factually true $\mathcal{X}^+ = \{h_i(\boldsymbol{x_j})_{[L]} : y_j = \text{true}\}$ and a set of false $\mathcal{X}^- = \{h_i(\boldsymbol{x_j})_{[L]} : y_j = \text{false}\}$ activations. Here, L is the index of the last token in $\boldsymbol{x}$. We then compute the means of each set, denoted $\boldsymbol{\mu}^+$ and $\boldsymbol{\mu}^-$, and get the mean-difference vector

$$\boldsymbol{\theta} = (\boldsymbol{\mu}^+ - \boldsymbol{\mu}^-), \text{ where } \boldsymbol{\theta} \in \mathbb{R}^{1\times d}. \tag{A4}$$

Statements are then scored along the whitened direction $\boldsymbol{\Sigma}^{-1}\boldsymbol{\theta}^\top$, where $\boldsymbol{\Sigma}$ is the pooled within-class covariance matrix; refer to the pseudocode in Algorithm 2.

Second, we augment MD with conformal prediction (Angelopoulos et al., 2021). Conformal predictions help detect statements that fall outside MD's high-confidence regions for *true* or *false* classes. We use $\alpha = 0.1$ in our experiments; that is, the conformal prediction sets (or regions) have a marginal coverage of at least 90% (assuming the calibration and test data are exchangeable), the correct label lies in the region at least 90% of the time *on average over the data*. Note that we use the *true* and *false* statements from the calibration dataset to find the conformal prediction regions.

Finally, we test the MD+CP probe using *true*, *false*, and *neither* statements from the test set. *How can the binary* MD+CP *classifier identify neither statements in addition to true and false statements?* If the MD+CP probe accurately captures the veracity signal, the *neither* statements (from the test dataset) should fall outside of the conformal regions. We observe that this is not the case: MD+CP assigns high-confidence scores to the *neither*-valued statements in the *true* or *false* regions.

Figure A1 shows the score distributions[16] of MD+CP on the best-performing decoder (i.e., 13th) of the default Llama-3-8B model on the *City Locations* dataset. There are three distributions: one for *true*, one for *false*, and one for *neither*. If MD+CP correctly captures the veracity signal, we expect the distribution of *neither* statements (green bars) to be outside of the conformal prediction region (i.e., in the gray area) and it is not included in the conformal prediction set. However, this is not the case. Most of the *neither* statements fall within the conformal prediction region (i.e., not in the gray area) and are labeled as *true* or *false*.

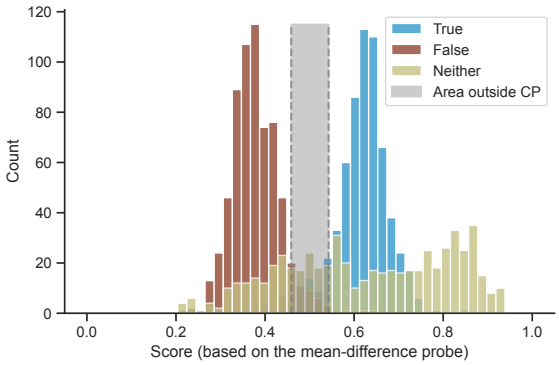

Figure A1: **Score distributions of mean-difference probe with conformal prediction (MD+CP) on the 13th decoder activations of the default Llama-3-8B model for the City Locations dataset.** The probe provides a good separation between *true* and *false* statements. If the MD+CP probe *truly* captures the veracity signal, we expect the scores for the *neither* statements to fall outside the conformal region (i.e., in the gray area), providing an **empty prediction set** (so the statement is labeled *neither*). However, MD+CP assigns high-confidence scores for the *neither* statements, labeling them *true* or *false*. This finding suggests that MD+CP relies on spurious proxies rather than genuine veracity signals.

Figure A2.A illustrates per-token MD+CP predictions across entire statements. If MD+CP correctly identifies the veracity signal, then (1) it should not assign any labels to the tokens in the pre-actualized parts of statements $\boldsymbol{x}^p$, and (2) the label should be consistent across the actualized part $\boldsymbol{x}^a$. Given a statement "The city of Tokyo is in Japan.", the pre-actualized part is "The city of Tokyo is in", and the actualized part is "Japan."

We observe that MD+CP assigns scores to the pre-actualized tokens that fall within the conformal prediction regions in cases #1–5 (see Figure A2). In case #5, MD+CP assigns a correct prediction at the *period* ($p = 0.39$ corresponds to a false label), but the prediction flips at the end of the text, where

[16] We use the embedding of the last token to compute the scores in Figure A1: $g_i\left(h_i\left(\boldsymbol{x}\right)\right) = \boldsymbol{\theta}^\top h_i(\boldsymbol{x})_{[L]} + \beta$. Here, L is the total length of the statement, which is the same as the index of the last token.

the *question mark* gets $p = 0.95$ corresponding to the *true* label. Similarly, in #7 case of Figure A2.A, the sentence does not have *any* truth value (i.e., it is not a factual claim). The MD+CP probe assigns high confidence scores to some of its tokens. These findings suggest that MD+CP probe captures proxies or spurious correlations. One cannot use it in real scenarios, where we do not know a priori where the factual claim ends. In contrast, Figure A2.B–D show the per-token predictions for the one-vs-all sAwMIL. These probes correctly identify positions where the veracity is actualized. For example, they do not assign predictions to the non-actualized parts of the statements and only label tokens in the actualized part. Moreover, these probes do not label tokens in cases where veracity is absent (e.g., see case #7 in Figure A2.D).

F.2 Multiclass Single-Instance Support Vector Machine

The multiclass sAwMIL probe is a multiple-instance learning version of a support vector machine, designed to operate on bags of token representations. To assess whether the MIL formulation offers any benefits, we construct a single-instance baseline by training a multiclass SVM on the last token representation only $h_i(\boldsymbol{x})_{[L]}$. As with multiclass sAwMIL, we first train three one-vs-all probes: `is-true`, `is-false`, and `is-neither`. Then, we assemble one-vs-all classifiers into a multiclass SVM using the same procedure described in Section 3.1 and Algorithm 1.

As before, to evaluate performance, we provide all token representations $h_i(\boldsymbol{x})$ (not only the last one), where the final prediction is computed based on

$$\hat{g}_i(\boldsymbol{x}) \;=\; \max_{1 \leq j \leq L} \; g_i\big(h_i(\boldsymbol{x})_{[j]}\big), \qquad \text{(A5)}$$

where $L = |\boldsymbol{x}|$ (number of tokens in $\boldsymbol{x}$).

Figure A3 depicts the performance of multiclass sAwMIL vs. multiclass SVM. The performance of multiclass SVM is closer to the performance of multiclass sAwMIL (as compared to MD+CP probe in Figure A3.A). However, multiclass sAwMIL still outperforms the multiclass single-instance SVM in 46 out of 48 cases ($=$ 16 LLMs $\times$ 3 datasets). For more results, we refer the reader to Table A11 and Table A20.

In Figure A3, we also observe that the multiclass sAwMIL performs better on the chat models (see bottom right portion of the plot) than the multiclass SVM. This supports our claim that veracity signals often emerge at positions other than the final token, and that multiple-instance learning can better isolate the veracity signal. Recall that the multiclass sAwMIL probe considers all the tokens in the statement and has additional training stages.

Figure A4 visualizes the per-token predictions. The one-vs-all SVM-based probes have better selectivity than the mean-difference probe with conformal prediction (MD+CP). However, in some cases, one-vs-all SVM assigns labels to the tokens in the pre-actualized part of the statement (see Figure A4.B, #5 statement). This suggests that one-vs-all SVM probes are capturing spurious correlations with potential proxies.

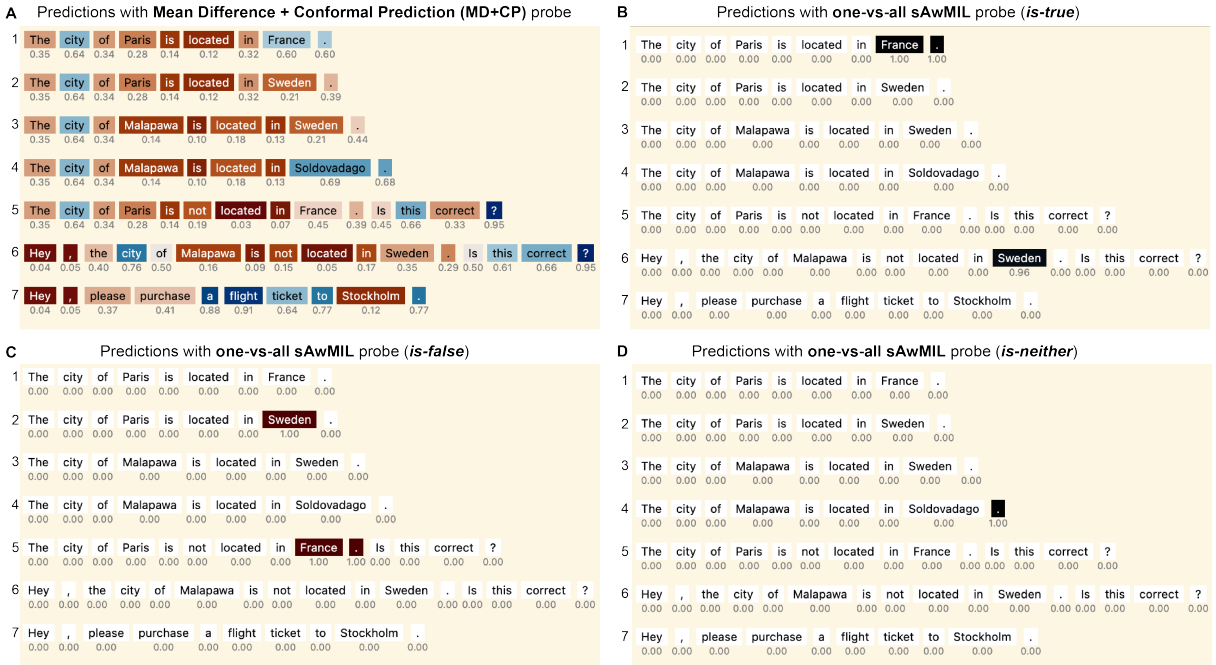

Figure A2: **Per-token predictions of mean-difference with conformal prediction (`MD+CP`) and one-vs-all `sAwMIL` on the 13th decoder activations of the default `Llama-3-8B` model.** Statements are from the *City Locations* dataset. We show per-token probabilities (printed beneath each word), assigned based on the token's representation. Words are shaded based on the predicted probability. When MD+CP outputs 0, the statement is labeled false; when it outputs 1, the statement is labeled true. If the per-token score falls outside the conformal prediction region, `MD+CP` assigns a score of 0.5 to that token (which corresponds to the highest uncertainty). The one-vs-all `sAwMIL` probe for `is-true` outputs 1 when the probe is 100% confident that the statement is true; and it outputs 0 when the per-token score is outside the conformal prediction region (i.e., there is an absence of truthfulness signal). Similarly, the one-vs-all `sAwMIL` probe for `is-false` outputs 1 when the probe is 100% confident that the statement is false; and it outputs 0 when the per-token score is outside the conformal prediction region (i.e., there is an absence of falsehood signal). The same logic applies for the one-vs-all `sAwMIL` probe for `is-neither`. **Panel A** shows the `MD+CP` predictions. It often assigns high confidence scores to pre-actualization tokens and makes mistakes on the *wrapped* prompts in cases #5 and 6 (e.g., statement #6: "Hey,___ Is this correct?"). Also, `MD+CP` probe assigns labels to the statement without any truth value (e.g., case #7). **Panels B–D** display one-vs-all `sAwMIL` probes (`is-true`, `is-false`, and `is-neither`). Unlike `MD+CP`, one-vs-all `sAwMIL` localizes the veracity signal to the actualized token and abstains elsewhere, demonstrating superior selectivity.

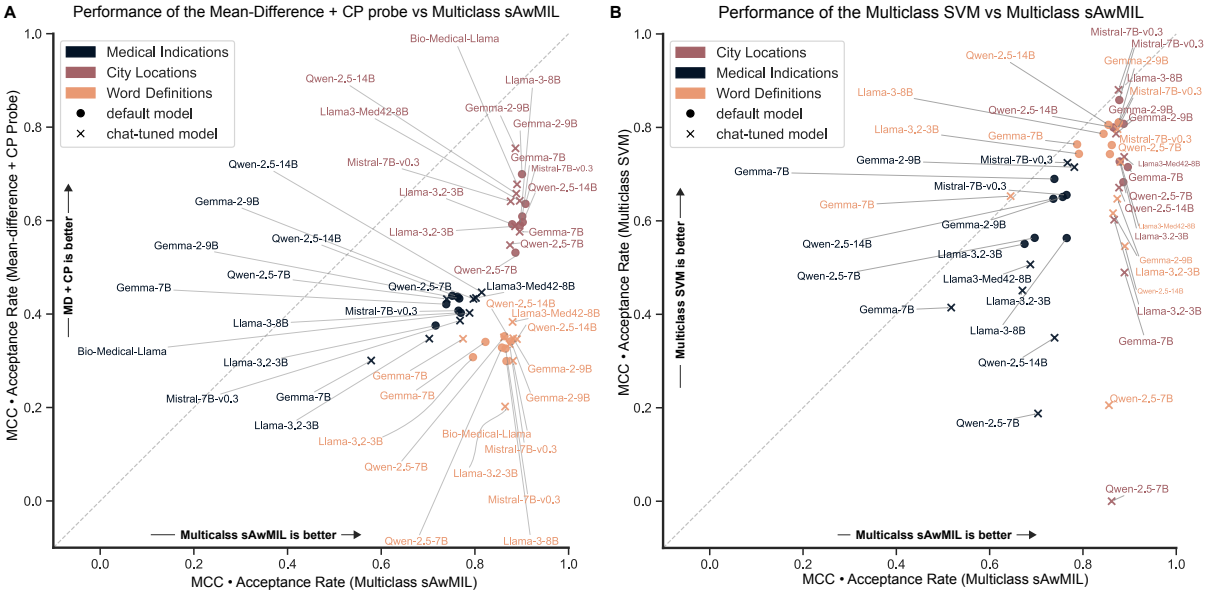

Figure A3: **Comparison of performance for `MD+CP`, a multiclass SVM and multiclass `sAwMIL` probes. Panels A & B:** Each marker shows a probe's performance for a ⟨model, dataset⟩ pair. Default models are shown with circles, while chat models are shown with crosses. Different colors indicate different datasets. **Panel A** shows the comparison between the multiclass `sAwMIL` probe on the x-axis and the mean-difference probe with conformal prediction (`MD+CP`) on the y-axis. **Panel B**: Performance of multiclass `sAwMIL` vs. multiclass single-instance SVM probes. The performance of the multiclass `sAwMIL` probe is specified on the x-axis, while the performance of the multiclass SVM is specified on the y-axis. We observe that multiclass `sAwMIL` outperforms multiclass SVM. The only exceptions are the `Gemma-7B` chat model and the `Mistral-7B-v0.3` chat model on *Word Definitions*, where the performances of the single-instance and multiple-instance probes are competitive. This experiment shows that multiple-instance learning (i.e., training on all the tokens in the statement) is beneficial when tracking the veracity of an LLM. Overall, the multiclass `sAwMIL` probe outperforms `MD+CP` and the multiclass SVM.

Figure A4: **Per-token predictions of one-vs-all single-instance SVM on the 13th decoder activations of the default `Llama-3-8B`**. Statements are from the *City Locations* dataset. We show per-token probabilities (printed beneath each word), assigned based on the token's representation. Words are shaded based on their predicted probability. **Panels A–C** display the one-vs-all `SVM` probes (`is-true`, `is-false`, and `is-neither`). The one-vs-all SVM probe isolates the signal better than the `MD+CP` probe. (See Figure A2.A for the `MD+CP` results.) However, in some cases, the one-vs-all SVM probe assigns high-certainty scores to tokens that lack any veracity signal. For example, in Panel A (case #6), the probe picks up on tokens including 'this,' 'is,' and '?,' which do not have inherent truth value. Overall, the `sAwMIL` in Figure A2.B–D has better selectivity.

G Zero-Shot Prompting: Instructions, Veracity Labeling, and Abstention

Instructions in Zero-shot Prompting. In zero-shot prompting, each LLM receives an instruction along with a statement $\boldsymbol{x}$ as input. These instructions outline the task and describe the output format. We use zero-shot prompts to evaluate how well LLMs can assess the veracity of a given statement and a multiple-choice question.

In our (zero-shot prompting) experiments, we use three different templates. All default models share the same template as displayed in Figure A5.A–B. We do not use this template for chat models, since they support "turn-based" conversations. Thus, for most chat models, we use templates that simulate user-assistant dialog. We use the template in Figure A5.D–E for the chat models that support context prompts. We use the template in Figure A5.C and in Figure A5.F for `Gemma` chat models since they do not support *context prompts*. Additionally, we use three different phrasings of the instructions:

1. **Original instructions** are Figure A5.A and Figure A5.D. In the main paper, we report results based on these instructions.
2. **True-False instructions**, where we change the phrasing of the original instructions by replacing the words "correct" and "incorrect" with "true" and "false". These are displayed in Figure A5.B and Figure A5.E.
3. **Shuffled instructions**, where we change the original instructions by shuffling choice identifiers. These are displayed in Figure A5.C and Figure A5.F.

Figures A6 to A8 depict the Weighted Matthews Correlation Coefficient (W-MCC) performance of zero-shot prompting[17] across three datasets and 16 LLMs for each template type and instruction type.

In some cases, minor changes to the instructions can significantly affect performance. Overall, the `Qwen` and `Gemma-2` models achieve the best performance among LLMs. However, `Qwen` models seem to be more sensitive to the phrasing of the prompt.

From token probabilities to veracity labels in Zero-shot prompting. Given the original instructions and the statement $\boldsymbol{x}$, an LLM outputs token-level probabilities over its vocabulary $\mathcal{V}$ as

$$P_{\mathcal{M}}(\tau \mid \text{instruction} \wedge \boldsymbol{x}) \quad \text{with} \sum_{\tau \in \mathcal{V}} P_{\mathcal{M}}(\tau \mid \text{instruction} \wedge \boldsymbol{x}) = 1$$

We are interested in the probabilities of the tokens that correspond to the multiple choices, i.e., numbers 1–6 in any of the panels in Figure A5. We denote tokens associated with these numbers as: `[1]`, `[2]`, `[3]`, etc. We map these token-level probabilities $P_{\mathcal{M}}$ into the veracity-label probabilities $G_{\mathcal{M}}$ as follows:

$$\begin{aligned} G_{\mathcal{M}}(\text{true} \mid \boldsymbol{x}) &= P_{\mathcal{M}}(\texttt{[1]} \mid \text{instruction} \wedge \boldsymbol{x}) \\ G_{\mathcal{M}}(\text{false} \mid \boldsymbol{x}) &= P_{\mathcal{M}}(\texttt{[2]} \mid \text{instruction} \wedge \boldsymbol{x}) \\ G_{\mathcal{M}}(\text{neither} \mid \boldsymbol{x}) &= P_{\mathcal{M}}(\texttt{[3]} \mid \text{instruction} \wedge \boldsymbol{x}) \\ &\quad + P_{\mathcal{M}}(\texttt{[4]} \mid \text{instruction} \wedge \boldsymbol{x}) \end{aligned}$$

Abstention in Zero-shot prompting. We include options `[5]` and `[6]` to check the "sanity" of the model $\mathcal{M}$. For example, option #5 in Figure A5.A suggests that a statement $\boldsymbol{x}$ is *true*, *false*, and *neither* – all at the same time. If the LLM concentrates most of its probability mass on tokens in $\mathcal{V}_{\perp} = \mathcal{V} \setminus \{\texttt{[1]}, \ldots, \texttt{[4]}\}$ (the sanity-check options $\{\texttt{[5]}, \texttt{[6]}\}$ or any token outside the answer set $\{\texttt{[1]}, \ldots, \texttt{[4]}\}$), we assume the LLM did not follow the instructions and *abstains* from prediction:

$$G_{\mathcal{M}}(\text{abstain} \mid \boldsymbol{x}) = \sum_{\tau \in \mathcal{V}_{\perp}} P_{\mathcal{M}}(\tau \mid \text{instruction} \wedge \boldsymbol{x}) \quad \text{(A6)}$$

Recall that the zero-shot prompting relies solely on the token-level probabilities, i.e., $\mathcal{M}$'s output. It does not look at the intermediate hidden representation of $\boldsymbol{x}$.

[17] W-MCC is an MCC score weighted by the acceptance rate (1-abstention rate).

A **Original Instructions (Default Model)**

```
Question: Is the following statement correct?
The city of Agadir is located in Morocco.

Select one of the following options:
      1. The statement is correct.
      2. The statement is incorrect.
      3. I do not have sufficient knowledge.
      4. The statement is too ambiguous.
      5. All of the above.
      6. None of the above.

Please respond with the corresponding number.
The final answer is
```

B **True-False Instructions (Default Model)**

```
Question: Is the following statement true?
The city of Agadir is located in Morocco.

Select one of the following options:
      1. The statement is true.
      2. The statement is false.
      3. I do not have sufficient knowledge.
      4. The statement is too ambiguous.
      5. All of the above.
      6. None of the above.

Please respond with the corresponding number.
The final answer is
```

C **Shuffled Instructions (Default Model)**

```
Question: Is the following statement correct?
The city of Agadir is located in Morocco.

Select one of the following options:
      2. The statement is correct.
      6. The statement is incorrect.
      4. I do not have sufficient knowledge.
      5. The statement is too ambiguous.
      1. All of the above.
      3. None of the above.

Please respond with the corresponding number.
The final answer is
```

D **Original Instructions with context (Chat Model)**

System (Context Prompt)

```
You are an expert in fact-checking. Your task
is to assist the user by answering questions
based on your comprehensive knowledge.
Please respond with the corresponding number.
```

User

```
Question: Is the following statement correct?
The city of Agadir is located in Morocco.

Select one of the following options:
      1. The statement is correct.
      2. The statement is incorrect.
      3. I do not have sufficient knowledge.
      4. The statement is too ambiguous.
      5. All of the above.
      6. None of the above.
```

Assistant

```
The final answer is
```

E **True-False Instructions *with* context (Chat Model)**

System (Context Prompt)

```
You are an expert in fact-checking. Your task
is to assist the user by answering questions
based on your comprehensive knowledge.
Please respond with the corresponding number.
```

User

```
Question: Is the following statement true?
The city of Agadir is located in Morocco.

Select one of the following options:
      1. The statement is true.
      2. The statement is false.
      3. I do not have sufficient knowledge.
      4. The statement is too ambiguous.
      5. All of the above.
      6. None of the above.
```

Assistant

```
The final answer is
```

F **Shuffled Instructions *without* context (Chat Model)**

User

```
You are an expert in fact-checking. Your task
is to assist the user by answering questions
based on your comprehensive knowledge.
Please respond with the corresponding number.

Question: Is the following statement correct?
The city of Agadir is located in Morocco.

Select one of the following options:
      2. The statement is correct.
      6. The statement is incorrect.
      4. I do not have sufficient knowledge.
      5. The statement is too ambiguous.
      1. All of the above.
      3. None of the above.
```

Assistant

```
The final answer is
```

Figure A5: **Zero-shot prompt templates.** We use these templates in our experiments. **Panel A–C:** The prompts for the default models. **Panel D–F:** Examples of prompts used for the chat models. Note that chat models like `Gemma` do not have a context (or system) prompt; hence, we provide instructions in the first message (see Panel F). In the main manuscript, we report the performance over the *original instructions*, i.e., instructions in Panels A and D. For the chat LLMs without the context prompt, we apply the two-message template in Panel F, but use the *original instructions*. Note: The statement used in these examples is factually correct.

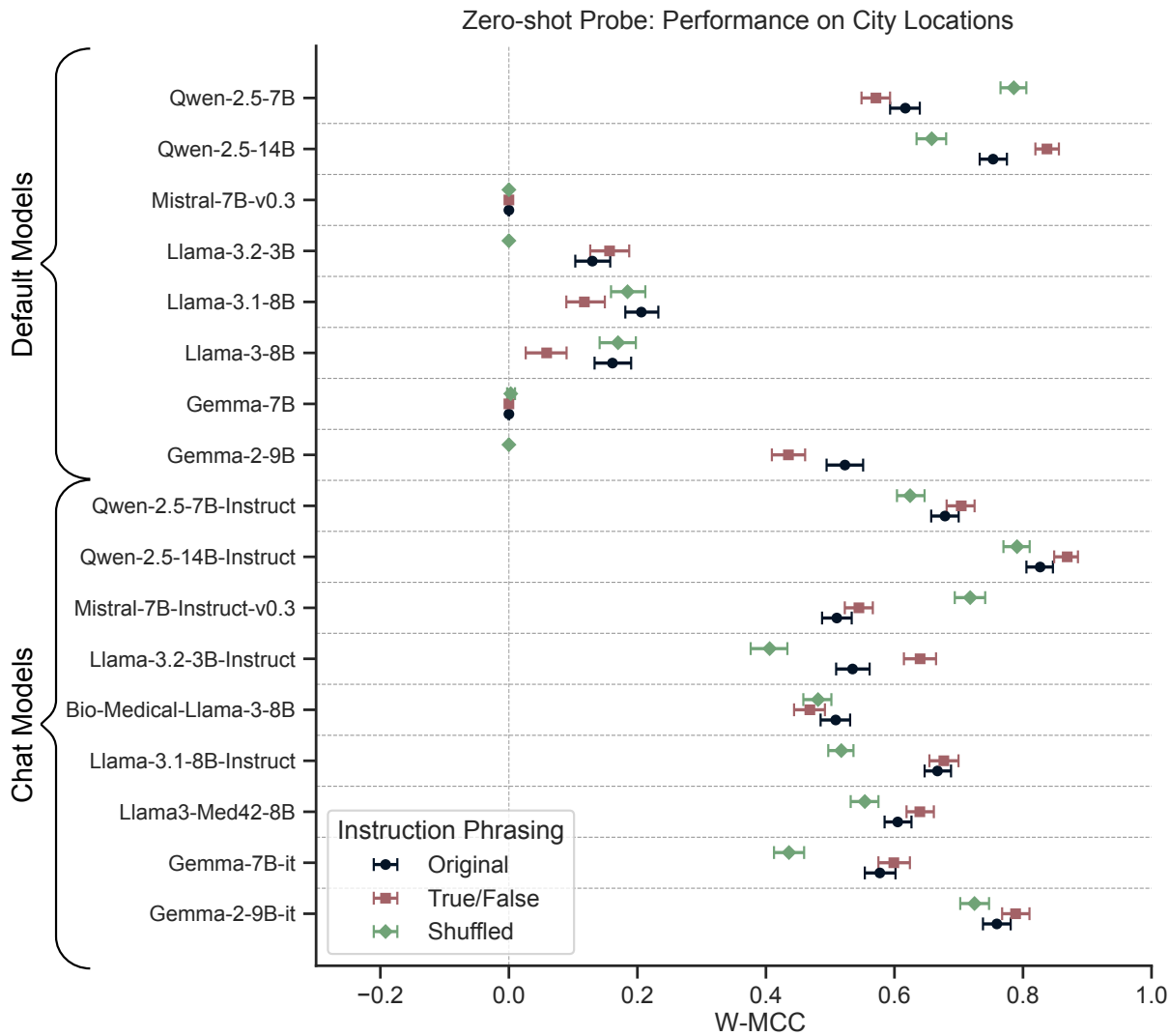

Figure A6: **Performance of zero-shot prompting on the *City Locations* dataset across different models and instruction phrasings.** We use the Weighted Matthews Correlation Coefficient (W-MCC) to quantify the performance. The marker shows the mean value and the error bars show the 95% confidence intervals (based on bootstrapping with $n = 1{,}000$ bootstrap samples). Minimal changes to the prompt instructions can skew the performance of zero-shot prompting. Chat models exhibit the highest performance across all instruction phrasings. However, the default `Qwen` models match the performance of other chat-based models. Shuffled instructions appear to lead to worse performance in chat models. We expected that the phrasings would have only a minor effect on their performance. The default `Gemma` and `Mistral` models seem to fail (their performance is around 0).

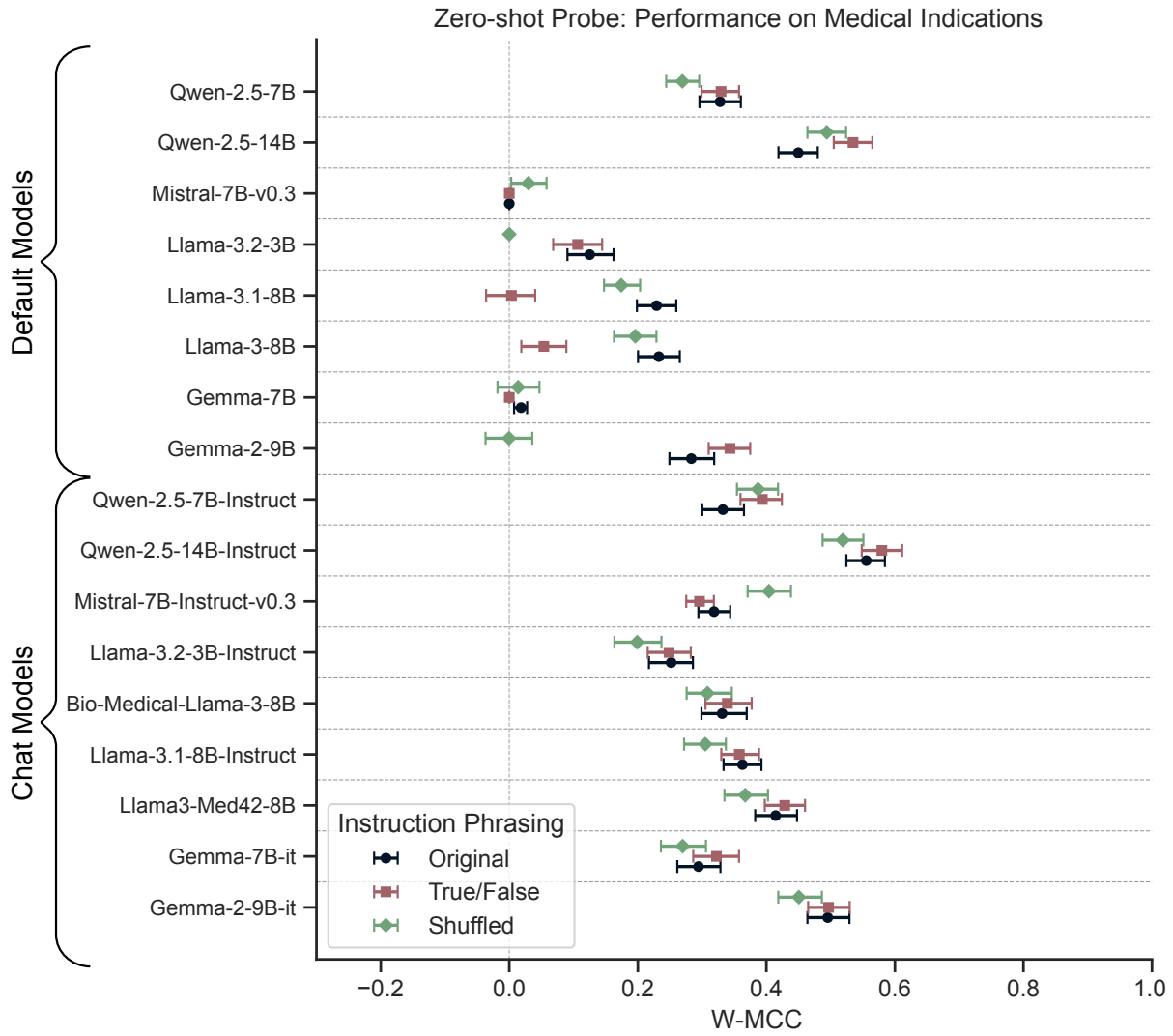

Figure A7: **Performance of zero-shot prompting on the *Medical Indications* dataset across different models and instruction phrasings.** We use the Weighted Matthews Correlation Coefficient (W-MCC) to quantify performance. The marker shows the mean value, and the error bars show the 95% confidence intervals (based on bootstrapping with $n = 1{,}000$ bootstrap samples). Minimal changes to the prompt instructions can skew the performance of zero-shot prompting. We observe a slight performance misalignment depending on the phrasing of the instruction. The best-performing LLMs are the largest chat models: `Gemma-2-9b` and `Qwen-2.5-14b`. We expected the biomedical `Llama` models to outperform on the medical indications dataset.

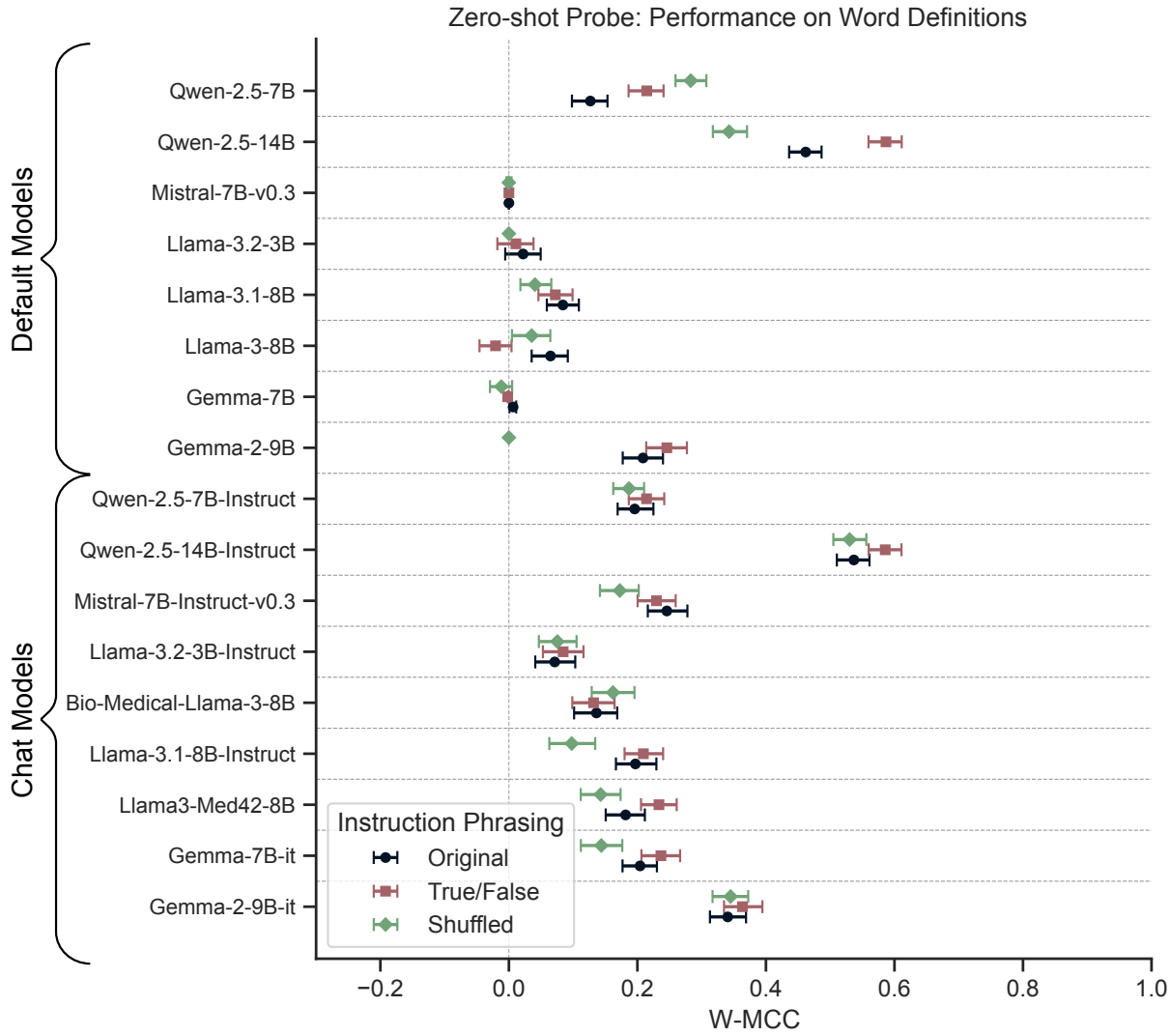

Figure A8: **Performance of the zero-shot prompting on the *Word Definitions* dataset across different LLMs and instruction phrasings.** We use the Weighted Matthews Correlation Coefficient (W-MCC) to quantify performance: the marker shows the mean value and the error bars show the 95% confidence intervals (based on bootstrapping with $n = 1{,}000$ bootstrap samples). Minimal changes to the prompt instructions can skew the performance of zero-shot prompting. The overall performance on the *Word Definitions* dataset is much lower than that on the other datasets. Generally, the misalignment in the performance between different instructions is much lower (except for the default `Qwen` models, where the difference is significant). The largest `Qwen` models, `Qwen-2.5-14b` and `Qwen-2.5-14b-instruct`, are the top-performing models on this task.

H Technical details for sAwMIL

Our proposed method, sAwMIL (sparse-aware multiple-instance learning), is a generalization of the standard Single-Instance Support Vector Machine (SVM) to the multiple-instance setting. In this section, we briefly recap the single-instance SVM (SIL-SVM) formulation and describe what changes when we switch to the multiple-instance setting. To make this section easier to read, Table A6 provides an overview of the relevant notations.

Symbol	Description
$\boldsymbol{x}$	Feature vector for a single instance, $\boldsymbol{x} \in \mathbb{R}^d$
B	Bag consisting of instances
$\mathcal{X}^+$, $\mathcal{X}^-$	Sets of positive and negative bags
$\tilde{X}^+$	Set of instances from all the positive bags
$\tilde{X}^-$	Set of instances from all the negative bags
$\mathcal{D}$	Dataset
$\tilde{\mathcal{D}}$	Relabeled dataset
$\boldsymbol{m}$	Intra-bag labels, $\boldsymbol{m} = \{0,1\}^{\lvert B\rvert}$

Table A6: Notations related to the sAwMIL description.

H.1 Single-Instance SVM

Support Vector Machines are maximum margin classifiers. The general idea behind the single-instance SVM (SIL-SVM)[18] is to find a separating hyperplane that maximizes the margin between the samples of the two classes. Thus, in the single-instance setting, our training dataset

$$\mathcal{D} = \{\langle \boldsymbol{x}_i, y_i \rangle\}_{i=1}^N \tag{A7}$$

consist of N instance-label pairs, with feature vectors $\boldsymbol{x}_i \in \mathbb{R}^d$ and binary labels $y_i \in \{-1, +1\}$.

The separating hyperplane is parameterized by a weight vector $\boldsymbol{\theta} \in \mathbb{R}^{1\times d}$ and bias $\beta \in \mathbb{R}$. The class boundaries are defined by $\boldsymbol{\theta}^\top \phi(\boldsymbol{x}) + \beta = \pm 1$, where $\phi(\cdot)$ is a feature transformation. In the case of the linear SVM, $\phi(\cdot)$ is an identity function.

The margin width equals $2/\|\boldsymbol{\theta}\|$, thus, maximizing the margin is equivalent to minimizing $\|\boldsymbol{\theta}\|^2$. At the same time, to allow for misclassifications, we introduce per-instance slack variables $\xi_i \geq 0$, which measure how far each point lies from the correct side of the margin. The resulting (primal) problem is:

$$\begin{aligned} \min_{\boldsymbol{\theta},\beta,\boldsymbol{\xi}} \quad & \frac{\|\boldsymbol{\theta}\|^2}{2} + \frac{C}{N}\sum_{i=1}^N \xi_i \\ \text{s.t.} \quad & y_i\big(\boldsymbol{\theta}^\top \phi(\boldsymbol{x}_i) + \beta\big) \geq 1 - \xi_i, \\ & \xi_i \geq 0, \quad \forall\, i \end{aligned} \tag{A8}$$

where $C > 0$ controls the trade-off between margin maximization and penalizing margin violations[19]. The decision function is $f(\boldsymbol{x}) = \boldsymbol{\theta}^\top \phi(\boldsymbol{x}) + \beta$, and the predicted label is $\hat{y} = \text{sign}(f(\boldsymbol{x}))$.

H.2 The Multiple-Instance Setting

In the multiple-instance setting, we no longer have an individual instance per label. Instead, our training data consists of *bags*, such that

$$\mathcal{D} = \{\langle B_i, y_i \rangle\}_{i=1}^N . \tag{A9}$$

Each bag B_i contains a set $\{\boldsymbol{x}_i^{(1)}, \ldots, \boldsymbol{x}_i^{(L_i)}\}$ of L_i instances that share a single bag-level label $y_i \in \{-1, +1\}$. The important assumptions for the multiple-instance SVMs are as follows:

1. A **negative bag** ($y_i = -1$) contains *only* negative instances.
2. A **positive bag** ($y_i = +1$) contains *at least one* positive instance.

The sAwMIL training consists of three stages. Our framework mirrors the sbMIL framework proposed by Bunescu and Mooney (2007):

1. We first train a multiple-instance SVM that operates both on bags and instances. Bunescu and Mooney refer to this SVM as a sparse MIL or sMIL.
2. In the second stage, we use the classifier from Stage 1 to relabel instances in every positive bag.
3. In the third stage, we "abandon" the bag structure and instead use all the instances together with the new labels from Stage 2. That is, we train a single-instance SVM, refer to Section H.1.

The only difference between Bunescu and Mooney's sbMIL and our sAwMIL is how we relabel instances in Stage 2. Below, we provide a detailed description of each stage.

[18] In most cases, when people mention SVMs, they refer to the single-instance variant.

[19] Larger values penalize margin violations more heavily, yielding a narrower margin but fewer misclassifications on the training set.

Stage 1: The Hybrid Bag–Instance SVM

Let $\mathcal{X}^-$ denote the set of all negative bags and $\mathcal{X}^+$ the set of all positive bags. Further, we denote

$$\tilde{X}^- = \bigcup_{B \in \mathcal{X}^-} \{\boldsymbol{x} : \boldsymbol{x} \in B\}, \tag{A10}$$

$$\tilde{X}^+ = \bigcup_{B \in \mathcal{X}^+} \{\boldsymbol{x} : \boldsymbol{x} \in B\} \tag{A11}$$

as the flattened sets of all instances from negative and positive bags, respectively. Bunescu and Mooney design the sparse MIL (sMIL) algorithm specifically for sparse bags. That is, sMIL is suitable for cases in which only a few instances in the positive bags are true positives, while the majority of instances are negative. For that reason, sMIL treats positive and negative bags differently by introducing two types of constraints.

Negative bags: instance-level constraints. Since every instance in a negative bag is a true negative, we impose a per-instance margin constraint over instances in $\tilde{X}^-$:

$$\boldsymbol{\theta}^\top \phi(\boldsymbol{x}_j) + \beta \;\leq\; -1 + \xi_{\boldsymbol{x}_j}, \forall\, \boldsymbol{x}_j \in \tilde{X}^-. \tag{A12}$$

Each negative instance contributes independently to the optimization, fully exploiting the *unambiguous* negative labels. Note that eq. (A12) is identical to the negative-class constraints in the single-instance SVM from Section H.1, i.e., we do not exploit the bag structure.

Positive bags: bag-level constraint. For positive bags, we do not know *which* instances are positive, so we cannot impose per-instance constraints. Instead, we impose a single constraint on the representation of each positive bag $B \in \mathcal{X}^+$. Under the assumption that a positive bag has at least one positive instance in B and the remaining $|B| - 1$ are likely to be negative, the expected average hidden label is:

$$\frac{(+1) + (|B| - 1) \cdot (-1)}{|B|} \;=\; \frac{2 - |B|}{|B|}. \tag{A13}$$

Using Equation (A13) as the required margin for the bag mean gives the bag-level constraint

$$\boldsymbol{\theta}^\top \frac{\phi(B_j)}{|B_j|} + \beta \;\geq\; \frac{2 - |B_j|}{|B_j|} - \xi_{B_j}, \quad \forall\, B_j \in \mathcal{X}^+, \tag{A14}$$

where $\xi_B \geq 0$ is a bag-level slack and

$$\phi(B) = \sum_{\boldsymbol{x} \in B} \phi(\boldsymbol{x}) \tag{A15}$$

In the linear case, ϕ is the identity. Note that the right-hand side of (A14) becomes more negative as $|B|$ grows, reflecting that the bag mean is increasingly diluted by negative instances in larger bags, i.e., the constraint on the bag mean naturally relaxes accordingly.

Finally, we combine eqs. (A12) and (A14) and now solve the following convex problem (can be solved with standard QP solvers):

$$\begin{aligned} \min_{\boldsymbol{\theta}, \beta, \boldsymbol{\xi}} \; & \frac{\|\boldsymbol{\theta}\|^2}{2} + \frac{C}{|\tilde{X}^-|} \sum_{\boldsymbol{x} \in \tilde{X}^-} \xi_{\boldsymbol{x}} \\ & + \frac{C}{|\mathcal{X}^+|} \sum_{B \in \mathcal{X}^+} \xi_B \\ \text{s.t.} \quad & \text{eqs. (A12) and (A14).} \end{aligned} \tag{A16}$$

Output. At the end of this stage, we get coefficients $\hat{\boldsymbol{\theta}} \in \mathbb{R}^{1 \times d}$ and $\hat{\beta} \in \mathbb{R}$.

Stage 2: Pseudo-label assignment

We use the estimated parameters $\hat{\boldsymbol{\theta}}$ and $\hat{\beta}$ to relabel all the instances in the positive bags. That is, for each instance $\boldsymbol{x}_j \in \tilde{X}^+$, we compute a score s_j according to eq. (A17):

$$s_j = \hat{\boldsymbol{\theta}}^\top \phi(\boldsymbol{x}_j) + \hat{\beta}, \tag{A17}$$

that further forms

$$S^+ = \{s_j \mid \boldsymbol{x}_j \in \tilde{X}^+\}. \tag{A18}$$

Further, we find the $(1 - \eta)$-quantile of the score distribution S^+. Here, η represents the expected fraction of the positive instances across positive bags. Formally,

$$q = Q_{1-\eta}(S^+)\,. \tag{A19}$$

Recall that in our experiments, the statements decompose into pre-actualized and actualized parts: $\boldsymbol{x} \leftarrow [\boldsymbol{x}^p, \boldsymbol{x}^a]$. That is, we know where the veracity signal is more likely to be. Thus, in our experiments, each bag B has a bag label y **and** intra-bag labels $\boldsymbol{m}$. Hence, the dataset is

$$\mathcal{D} = \{\langle B_i, y_i, \boldsymbol{m}_i \rangle\}_{i=1}^N\,. \tag{A20}$$

These intra-bag labels $\boldsymbol{m} = \{0, 1\}^{|B|}$ specify what instances within the bag are likely to contain the veracity signal and which are not.

Using these two pieces of information, we can now relabel instances in every positive bag:

$$\tilde{y}_j = \begin{cases} +1 & \text{if } s_j \geq q \text{ and } m_j = 1 \\ -1 & \text{otherwise,} \end{cases} \quad \forall\, \boldsymbol{x}_j \in \tilde{X}^+ \tag{A21}$$

Output. The result of this stage is the relabeled dataset $\tilde{\mathcal{D}}$ consisting of instance–pseudolabel pairs:

$$\tilde{\mathcal{D}} = \{\langle \boldsymbol{x}_j,\, \tilde{y}_j \rangle \mid \boldsymbol{x}_j \in \tilde{X}^+ \cup \tilde{X}^-\}, \tag{A22}$$

where $\tilde{y}_j = -1$ for all $\boldsymbol{x}_j \in \tilde{X}^-$.

Stage 3: Instance-Level SVM (SIL)

In the final stage, we discard the bag structure and instead use the relabeled dataset:

$$\tilde{\mathcal{D}} = \{\langle \boldsymbol{x}_i, \tilde{y}_i \rangle\}_{i=1}^{N}. \tag{A23}$$

Here, we fit the single-instance SVM to get the final $\boldsymbol{\theta}$ and β. The resulting decision function

$$f(\boldsymbol{x}) = \boldsymbol{\theta}^\top \phi(\boldsymbol{x}) + \beta \tag{A24}$$

produces per-instance scores. We provide a pseudocode in Algorithm 1. Further, we can pass these scores to the conformal prediction algorithm (Section E).

H.3 Dual Formulation

In this section, we define SVM algorithms in terms of the primal formulation. However, one can convert the primal problem in Equations (A8) and (A16) into an equivalent dual formulation via Lagrangian multipliers; refer to Bunescu and Mooney (2007) for details.

I Additional Evaluation Details: Correlation, Selectivity and Generalization

Section D provides the full list of the validity criteria. Here, we provide additional details behind the evaluation of the **Correlation**, **Selectivity** and **Generalization**, not covered in Section 5.

I.1 Evaluation Setup

Here, we describe a pipeline to evaluate our sAwMIL probe over the validity criteria specified in Section 3 and Section D.

Correlation and Selectivity. We use the test split of each dataset to evaluate the performance of the probe. We use Matthews Correlation Coefficient (MCC) to summarize the statistical accuracy of probes. The multiclass MCC value is calculated using Equation (A25):

$$\mathrm{MCC} = \frac{c\,s - \sum_k p_k t_k}{\sqrt{\left(s^2 - \sum_k p_k^2\right)\left(s^2 - \sum_k t_k^2\right)}} \quad \text{(A25)}$$

where c is the number of correct predictions, s is the total number of samples, K is the total number of classes, t_k is the number of k-class samples in the dataset, and p_k is the number of times k-class was predicted. $\mathrm{MCC} = 1$ indicates that a classifier predicted every instance correctly. $\mathrm{MCC} = 0$ implies that the predictions are random. $\mathrm{MCC} = -1$ indicates that the predictions are inversely correlated with the ground-truth labels.

Since *neither* statements are included in the test dataset, MCC provides a sense of how well the probe classifies factually *true* or *false* statements and indicates whether the probes can handle *neither*-type cases.

Generalization. To test how well a particular probe g_i trained on a dataset $\mathcal{D}_i$ generalizes, we evaluate its performance using the test split of another dataset $\mathcal{D}_j$. That is, g_i trained on the city locations dataset is evaluated using the test split of the word definitions dataset.

I.2 Generalization Across Datasets

To further support the claim that the multiclass sAwMIL captures veracity signals (and not merely a proxy), we demonstrate generalization performance across datasets. Figure A9 provides results for each dataset and LLM. The columns correspond to three test datasets, and the cells specify the multiclass sAwMIL's performance for a specific LLM.

Multiclass sAwMIL provides reasonable generalization performance (see Table A7). However, it may overfit to the highly specialized *City Locations* dataset. Therefore, using more diverse datasets that span a broader range of entities and cover a larger set of topics better isolates the veracity signal and yields better generalization performance. We refer the reader to Table A22 through Table A24 for detailed statistics.

In Table A7, the values reported are for sAwMIL computed without conformal prediction, i.e., over all test statements. Averaging the off-diagonal cells of this table therefore gives 0.63, 0.76, and 0.81 for probes trained on City Locations, Medical Indications, and Word Definitions. In Table A21, we report 0.82, 0.88, and 0.86 for the sAwMIL with conformal prediction, where MCC is measured only on non-abstained predictions.

We also observe that generalization performance is higher for chat models, where the average MCC score (on the non-training dataset) is 77.2% (standard error: 0.2%), versus 68.2% (standard error: 0.2%) for the default models. This is more noticeable in Figure A9.A, where default models have much lower MCC values than their chat model counterparts. For example, the chat Llama-3.2-3B has an MCC value 1.6 times as high as the default Llama-3.2-3B.

Across the three panels, Gemma-7B seems to be an outlier, as its generalization performance drops significantly for the chat version of the model.

Multiclass sAwMIL satisfies the generalization criterion defined in Section D by transferring veracity probes trained on one dataset to another while maintaining strong performance. This provides further evidence that multiclass sAwMIL captures a veracity signal that is not dataset-specific.

I.3 Recap: Overall Validity

In this section and Section 5 of the manuscript, we established that the multiclass sAwMIL probe satisfies the validity criteria. Specifically, we confirmed that it satisfies the correlation and selectivity criteria and outperforms zero-shot prompting and the mean-difference probe using conformal prediction sets. We further demonstrated that multiclass sAwMIL satisfies the generalization criterion, indicating that we can successfully apply probes trained on multiclass sAwMIL to statements from other domains.

In addition, we showed that the one-vs-all sAwMIL probes satisfy the manipulation and local-

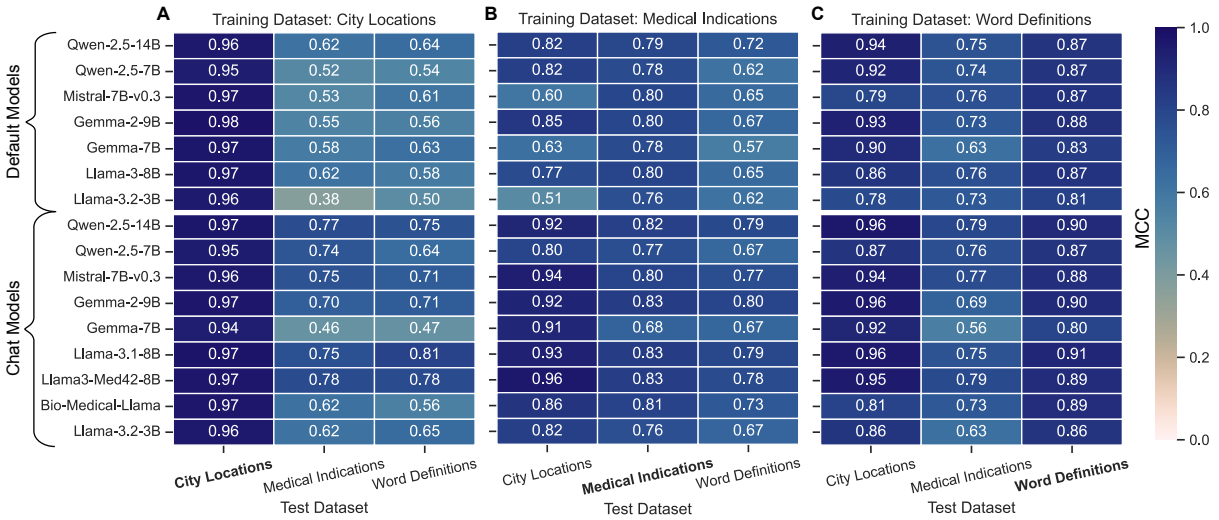

Figure A9: **Generalization performance of the multiclass `sAwMIL` probe across datasets.** Each panel corresponds to a different *training* dataset: *City Locations*, *Medical Indications*, and *Word Definitions*. Each column corresponds to a different test dataset. Each cell displays the MCC value, which quantifies how well the probe generalizes to the test dataset (higher values are better). For each model and dataset pair, we report the maximum MCC achieved across all decoders. In general, probes trained on chat models exhibit better generalization performance than the default models. **Panel A:** Generalization performance of multiclass `sAwMIL` when trained on *City Locations*. While the MCC values are significantly higher than a random baseline (with MCC = 0), the generalization ability is lower than that in Panels B or C. **Panel B:** Generalization performance of multiclass `sAwMIL` when trained on *Medical Indications*. In Panel A, we observe that training on *City Locations* and testing on *Medical Indications* provides good but not excellent MCC values (average MCC of 0.624 with standard error of 0.030). This is not the case in this panel, where *Medical Indications* is the training dataset and *City Locations* is the test dataset (average MCC of 0.818 with standard error of 0.033). **Panel C:** Generalization performance of multiclass `sAwMIL` when trained on *Word Definitions*. This probe has high generalization performance across datasets. When *City Locations* is the test dataset, the average MCC is 0.896 with a standard error of 0.015; and when *Medical Indications* is the test dataset, the average MCC is 0.723 with a standard error of 0.016. For aggregated statistics, see Table A7.

ity criteria (Section J). In most cases, we can intervene to change the log-probabilities of correct responses. Together, these findings provide strong evidence for the overall validity of `sAwMIL` probes. Not all LLMs have a veracity mechanism that has a linear relationship with the output. Exploring this non-linear relationship is part of our future work.

	Testing Dataset		
Training Dataset	City Locations	Medical Indications	Word Definitions
City Locations	0.963 (0.003)	0.624 (0.030)	0.633 (0.025)
Medical Indications	0.818 (0.033)	0.790 (0.009)	0.698 (0.018)
Word Definitions	0.896 (0.015)	0.723 (0.016)	0.868 (0.008)

Table A7: **Aggregated generalization performance of the multiclass `sAwMIL` for each dataset** (the values reported here are computed without conformal prediction, i.e., over all test statements, including the ones where `sAwMIL` would abstain). Each cell shows an MCC value, which quantifies the performance of the multiclass `sAwMIL` trained and tested on different combinations of the datasets. The value in the parentheses is the standard error. *Word Definitions* dataset provides better generalization performance because it contains statements covering a diverse set of topics, while the *City Locations* dataset provides lower generalization performance.

J Additional Evaluation Details: Causal Validation via Steering

Our causal validation follows the directional intervention paradigm (Li et al., 2023; Todd et al., 2024), where hidden activations are translated along a learned direction vector $\boldsymbol{\theta}_i$. Unlike standard activation patching (Meng et al., 2022)[20], which swaps activations between specific prompts to localize *where* knowledge is stored, our intervention tests *whether* a learned veracity direction causally influences model $\mathcal{M}$'s outputs. To identify this causal effect, we propose **Contrastive Token Intervention** (CTI) to estimate the causal effect of learned veracity directions on LLM outputs, i.e., the difference-in-differences design.

Activation Steering. Let $\boldsymbol{x} = [\boldsymbol{x}^p, \boldsymbol{x}^a]$ be a tokenized statement consisting of a pre-actualized part $\boldsymbol{x}^p$ and an actualized part $\boldsymbol{x}^a$. Let $\boldsymbol{\tau}_c$ denote the correct actualized token sequence (in this case, $\boldsymbol{x}^a = \boldsymbol{\tau}_c$) and $\boldsymbol{\tau}_r$ a random control token sequence (in this case, $\boldsymbol{x}^a = \boldsymbol{\tau}_r$ and $|\boldsymbol{\tau}_c| = |\boldsymbol{\tau}_r|$). Let $\boldsymbol{\theta}_i$ denote the learned veracity direction at layer i, let α denote the intervention dose, and κ indicate the sign of the translation.

We intervene by translating the hidden representation of the <u>last token</u> of $\boldsymbol{x}^p$ at layer i along $\boldsymbol{\theta}_i$ (representations of other tokens stay the same):

$$\tilde{h}_i^{\kappa}(\boldsymbol{x}^p) \leftarrow h_i(\boldsymbol{x}^p) + \kappa \cdot \alpha \cdot \boldsymbol{\theta}_i, \tag{A26}$$

where $\kappa \in \{-1, +1\}$, $\tilde{h}_i^{\kappa}(\boldsymbol{x}^p) \in \mathbb{R}^{|\boldsymbol{x}^p| \times d}$ and $h_i(\boldsymbol{x}^p) \in \mathbb{R}^{|\boldsymbol{x}^p| \times d}$. After we perform the intervention, we measure the change in log-probability of a target sequence $\boldsymbol{\tau}$ relative to the unperturbed model:

$$\begin{aligned}\Delta_{\boldsymbol{\tau}}^{\kappa} = &\log P_{\mathcal{M}}\Big(\boldsymbol{\tau} \mid \tilde{h}_i^{\kappa}(\boldsymbol{x}^p)\Big) \\ &- \log P_{\mathcal{M}}(\boldsymbol{\tau} \mid h_i(\boldsymbol{x}^p))\end{aligned} \tag{A27}$$

Evaluating Equation (A27) for each combination of target sequence $\boldsymbol{\tau} \in \{\boldsymbol{\tau}_c, \boldsymbol{\tau}_r\}$ and translation sign $\kappa \in \{-1, +1\}$ yields four values per statement:

Intervention Dosage. Our analysis shows that fixing α at a single value across probes yields intervention effects with substantially different magnitudes. This happens because each probe produces scores that have a different scale: for a fixed $\alpha = 1$, a probe with high variance (of output scores) produces a larger perturbation (translation) than one with low variance. To account for this, we calibrate α for each $\langle$probe, LLM, dataset, layer$\rangle$. That is, we take each g_i and compute the distribution of scores on the calibration data, $\mathcal{D}_{cal}$, and we then estimate the standard deviation of the score distribution σ. The resulting intervention dosage is computed based on Equation (A28):

$$\alpha = d' \cdot \sigma, \quad d' \in \{1, 3\}, \tag{A28}$$

where d controls the intervention strength in units of σ.

	$(\kappa = +1)$	$(\kappa = -1)$
Correct ($\boldsymbol{\tau}_c$)	$\Delta^{+}_{\boldsymbol{\tau}_c}$	$\Delta^{-}_{\boldsymbol{\tau}_c}$
Random ($\boldsymbol{\tau}_r$)	$\Delta^{+}_{\boldsymbol{\tau}_r}$	$\Delta^{-}_{\boldsymbol{\tau}_r}$

Table A8: Outcome variables used in the difference-in-differences design. Each cell shows the expected log-probability shift Δ under positive ($\kappa = +1$) and negative ($\kappa = -1$) interventions, for the correct tokens $\boldsymbol{\tau}_c$ and random baseline tokens $\boldsymbol{\tau}_r$. We plug these terms into Equation (A29) as the outcome variable.

Difference-in-Differences. To isolate veracity-specific effects from overall distributional shifts, we employ a **Difference-in-Differences** analysis. To reiterate, for each statement, we measure changes in log-probability under positive and negative translations for both the correct-answer tokens and random control tokens. The interaction term β_3 from the resulting 2×2 factorial model captures the differential causal effect (Table A8): how much more the intervention shifts the log-probability of correct tokens than of random tokens.

$$\begin{aligned}\Delta_{\tau}^{\kappa} = \beta_0 &+ \beta_1 \cdot \mathbb{1}[\boldsymbol{\tau} = \boldsymbol{\tau}_c] \\ &+ \beta_2 \cdot \mathbb{1}[\kappa = +] \\ &+ \beta_3 \cdot \mathbb{1}[\boldsymbol{\tau} = \boldsymbol{\tau}_c] \cdot \mathbb{1}[\kappa = +] \\ &+ \varepsilon\end{aligned} \tag{A29}$$

In Equation (A29), standard errors are clustered by statement to account for within-statement correlation across the four cells of the 2×2 model. This design controls for the possibility that a direction merely shifts the entire output distribution without specifically targeting veracity-relevant tokens. Here we provide the interpretation of each coefficient:

[20] In activation patching, one creates a pair of contrastive prompts (one that elicits the target behavior and one that does not). Activations from specific components (e.g., a feed-forward layer or attention head) of the clean run are then swapped into the corrupted run to determine whether that component is causally responsible for the behavioral difference.

β_1 : The *baseline advantage* of correct tokens over random tokens before any shift. A positive β_1 indicates that correct tokens have a higher probability than random tokens.

β_2 : The *overall distribution shift* introduced by the intervention. That is, what is the average effect of the intervention?

β_3 : The *veracity-specific shift*, aka what is the additional log-probability shift on the correct tokens induced by the intervention.

One can think (abstractly) of β_3 as the veracity-specific effect of the intervention, net of the generic distributional shift β_2 induced by translating along $\alpha \cdot \kappa \cdot \boldsymbol{\theta}_i$ as

$$\beta_3 = \underbrace{\left(\Delta^+_{\tau_c} - \Delta^-_{\tau_c}\right)}_{\text{effect on correct token}} - \underbrace{\left(\Delta^+_{\tau_r} - \Delta^-_{\tau_r}\right)}_{\text{effect on random token}} .$$

Therefore, eq. (A29) allows us to test the following hypothesis:

H_0**:** $\beta_3 = 0$: the veracity direction shifts all tokens equally (both $\boldsymbol{\tau}_r$ and $\boldsymbol{\tau}_c$).

H_1**:** $\beta_3 \neq 0$: the veracity direction $\boldsymbol{\theta}_i$ *specifically* targets the correct-answer token $\boldsymbol{\tau}_c$.

Cross-comparison. While **intervention dosage** normalization controls for scale differences across ⟨probe, LLM, dataset, layer⟩ configurations, β_3 remains correlated with σ, indicating the interaction effect is not fully scale-independent. To allow unbiased cross-configuration comparisons, we report two complementary metrics:

1. (Primary) **Normalized Interaction Effect** in eq. (A30): To summarize the magnitude of the causal effect in a scale-independent manner, we additionally report the normalized interaction effect, pooled across configurations via the DerSimonian–Laird estimator (Kelley and Kelley, 2023) described below.
2. (Secondary) **Success Rate**: the fraction of ⟨LLM, dataset⟩ configurations yielding a statistically significant β_3 after correction for the false discovery rate (per probe).[21] This measures how reliably a probe identifies a veracity-specific direction across configurations.

$$\text{Normalized Interaction Effect} = \left|\frac{\beta_3}{\hat{\sigma}}\right| \quad \text{(A30)}$$

[21]We apply the Benjamini and Hochberg (1995) procedure to control the false discovery rate.

In Equation (A30), $\hat{\sigma}$ is the residual standard deviation of the difference-in-differences fit.

Pooling Across Configurations. Even after dosage calibration, β_3 varies across ⟨probe, LLM, dataset⟩ configurations due to between-configuration heterogeneity. That is, different LLMs encode veracity at different layers and with different magnitudes. To obtain a single, robust summary per probe, we pool the normalized interaction effects across configurations using the DerSimonian–Laird random-effects estimator (Kelley and Kelley, 2023).

J.1 Setup

Collecting results. For each ⟨ probe, LLM, dataset ⟩ configuration, we collect the intervention results from the five best-performing layers (according to the MCC values) and fit the difference-in-differences model. We then select the layer with the median β_3. This layer is then considered to be the representative of the ⟨ probe, LLM, dataset ⟩ configuration. In total, we have 48 configurations. **Why?** (1) We want to make our analysis robust to the outliers and quantify the effect across "good" layers; (2) In some cases, the best-performing layer is the last one, and when we apply interventions to the last layer, the `PyTorch` (Paszke et al., 2019) package produces `NaN` values. For instance, that happens with the `TTPD` probe.

Comparison setting. We apply CTI to four probes: the mean-difference (`MD`), Training of Truth and Polarity Direction (`TTPD`), SVM, and `sAwMIL` — the latter two using only the `is-true` direction. We exclude the supervised PCA (`sPCA`) probe, as it does not yield a direction in the same representation space. Hence, our goal is to test whether the truthfulness directions identified by each probe causally influence $\mathcal{M}$ outputs, i.e., whether they can be used to steer LLM behavior toward or away from correct answers.

J.2 Results

To assess whether the learned veracity directions have a causal effect on LLM outputs, we look at the fraction of successful interventions for each probe across all ⟨`LLM`, `dataset`⟩ configurations (Figure A10.A), as well as the pooled normalized interaction effect in Figure A10.B; note that the pooled NIE is adjusted via the DerSimonian–Laird random-effects estimator (Kelley and Kelley, 2023). We also provide these results in Table A9.

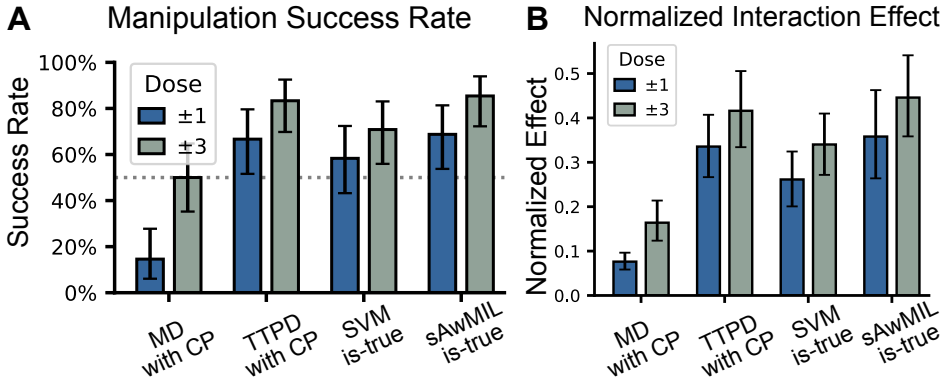

Figure A10: Manipulation + locality evaluation pooled over all 48 ⟨LLM, dataset⟩ configurations per each probe. Dose controls the magnitude of the interventions. **Panel A:** Intervention success rate per probe across all configurations (total: 48), with binomial proportion confidence intervals, i.e., fraction of configurations with significant β_3. Dose controls the magnitude of the translation along the learned veracity direction $\boldsymbol{\theta}_i$. A higher success rate indicates that activation steering along the veracity direction produced statistically significant changes in LLM output across more configurations. **Panel B**: Pooled normalized interaction effect, aka normalized β_3 adjusted via DerSimonian–Laird random-effects estimator (Kelley and Kelley, 2023). The success rate denotes the number of interventions with successful manipulation & locality, whereas the normalized effect size quantifies the magnitude of the effect. `MD` has the weakest causal effect, followed by the single-instance `SVM`. `TTPD` and the `sAwMIL`'s `is-true` probe achieve the highest success rates and similar effect sizes. These results are also displayed in Table A9.

Success indicates that the translation by the veracity direction ($\alpha \cdot \kappa \cdot \boldsymbol{\theta}_i$) yields a statistically significant change in the log-probability of the correct tokens as opposed to the irrelevant tokens (i.e., statistically significant β_3). The `MD` probe achieves the lowest success rate, indicating that the direction it identifies has little to no influence on majority of ⟨`LLM`, `dataset`⟩ configurations, while the single-instance `SVM` yields higher success. The `sAwMIL` and `TTPD` achieve the highest success rates, suggesting that both probes reliably identify directions the LLM actively uses when processing veracity-relevant information.

The pooled NIE in Figure A10 shows a similar trend, where `sAwMIL` and `TTPD` yield the highest intervention effects. That is, not only do these probes reliably find the veracity direction, but these directions are also used by the LLMs.

Meanwhile, Table A9 provides the between-configuration heterogeneity estimates τ^2. The higher this value is, the greater the variations in effect sizes across ⟨`LLM`, `dataset`⟩ configurations. Notably, `MD` has the lowest τ^2 value: the probe consistently finds directions with a weak causal effect. However, `TTPD` and `sAwMIL` both have higher τ^2 values: in some configurations, they identify directions with a high causal effect, whereas in others, they identify a direction with a smaller effect.

Together with findings related to correlation and generalization in Section 5, we see that `sAwMIL` yields better performance and also finds a direction with causal connections to LLM outputs. At the same time, `TTPD` seems to identify reliable veracity directions but is confounded by noise (since its performance and generalization do not hold as well).

Probe	Dose	Normalized Interaction Effect			Success Rate		
		NIE	95% CI	τ^2	SR	95% CI	k/n
MD	1	0.068	(0.050, 0.086)	0.0017	0.146	(0.061, 0.278)	7/48
	3	0.155	(0.116, 0.193)	0.0152	0.500	(0.352, 0.648)	24/48
TTPD	1	0.329	(0.261, 0.397)	0.0517	0.667	(0.516, 0.796)	32/48
	3	0.411	(0.332, 0.491)	0.0735	0.833	(0.698, 0.925)	40/48
SVM (is-true)	1	0.253	(0.198, 0.309)	0.0334	0.583	(0.432, 0.724)	28/48
	3	0.334	(0.267, 0.402)	0.0511	0.708	(0.559, 0.830)	34/48
sAwMIL (is-true)	1	0.350	(0.271, 0.428)	0.0722	0.688	(0.537, 0.813)	33/48
	3	0.439	(0.357, 0.520)	0.0768	0.854	(0.722, 0.939)	41/48

Table A9: Evaluation of the manipulation and locality criteria: normalized interaction effect and success rate per probe and intervention dose (across all 48 $\langle$LLM, dataset$\rangle$ configurations). Pooled **normalized interaction effect** (NIE) is estimated via the DerSimonian–Laird random-effects estimator (Kelley and Kelley, 2023). Here, τ^2 denotes between-configuration heterogeneity. **Success rate** (SR) is the fraction of configurations with statistically significant β_3 after BH correction (Benjamini and Hochberg, 1995), with k successes out of $n = 48$ configurations and binomial proportion 95% CIs (Wilson, 1927).

K Algorithms

In this section, we provide pseudocodes for several procedures described in the main text. In Algorithm 1, we provide pseudocode for the Sparse-Aware Multiple-Instance Learning (`sAwMIL`) probe. In Algorithms 2 to 4, we show pseudocodes for the mean-difference (`MD`), the training of truth and polarity directions (`TTPD`), and the supervised PCA (`sPCA`) probes, respectively. In Algorithms 5 and 6, we describe the procedure for the binary and multi-class conformal learning (described in Section 3.2).

Algorithm 1 Training a one-vs-all `sAwMIL` classifier

Input: A training dataset $\{(\boldsymbol{x}_i, y_i, \boldsymbol{m}_i)\}_{i=1}^{n}$ with binary bag labels $y_i \in \{0,1\}$, bags $\boldsymbol{x}_i \in \mathbb{R}^{L_i \times d}$, and intra-bag confidences $\boldsymbol{m}_i \in \{0,1\}^{L_i}$, where L_i is the number of items in a bag $\boldsymbol{x}_i$; also, a balancing parameter $\eta \in (0,1]$.

Output: Parameters $\boldsymbol{\theta} \in \mathbb{R}^{1\times d}$ and $\beta \in \mathbb{R}$. These parameters are subsequently given to a function f along with $\boldsymbol{z}$ to compute $f(\boldsymbol{z}) = \sigma\big(\boldsymbol{z}\,\boldsymbol{\theta}^\top + \beta\big)$, where σ is a sigmoid function.

1: Partition data into positive and negative sets:

$$\mathcal{X}^+ = \{\langle \boldsymbol{x}_i, \boldsymbol{m}_i \rangle : y_i = 1\}, \quad \mathcal{X}^- = \{\langle \boldsymbol{x}_i, \boldsymbol{m}_i \rangle : y_i = 0\}$$

2: Compute the **initial** coefficient vector and the intercept ▷ See Bunescu and Mooney (2007)

$$(\hat{\boldsymbol{\theta}}, \hat{\beta}) \leftarrow \texttt{solve_sMIL}(\mathcal{X}^+, \mathcal{X}^-), \text{ where } \hat{\boldsymbol{\theta}} \in \mathbb{R}^{1\times d} \text{ and } \hat{\beta} \in \mathbb{R}.$$

3: Let $\tilde{\mathcal{X}}^+$ denote the set of all instances from the positive bags and $\tilde{\mathcal{X}}^-$ denote all instances from the negative bags.

4: Compute scores for every instance in a positive set

$$S^+ \leftarrow \tilde{\mathcal{X}}^+ \hat{\boldsymbol{\theta}}^\top + \hat{\beta}, \text{ where } \tilde{\mathcal{X}}^+ \in \mathbb{R}^{|\tilde{\mathcal{X}}^+| \times d}.$$

5: Compute the threshold

$$q \leftarrow \texttt{quantile}(S^+, 1-\eta).$$

6: **for all** positive instances $\langle \bar{\boldsymbol{x}}_j, \bar{m}_j, \bar{y}_j \rangle \in \tilde{\mathcal{X}}^+$, where $\bar{\boldsymbol{x}}_j \in \mathbb{R}^{1\times d}$, $\bar{m}_j \in \{0,1\}$ and $\bar{y}_j = \emptyset$ **do**
  if $(\bar{\boldsymbol{x}}_j\,\hat{\boldsymbol{\theta}}^\top + \hat{\beta}) \geq q$ and $\bar{m}_j = 1$ **then** set $\bar{y}_j = 1$;
  else set $\bar{y}_j = 0$.

7: **end for**

8: Compute the **final** coefficient vector and the intercept ▷ via a SIL SVM

$$(\boldsymbol{\theta}, \beta) \leftarrow \texttt{solve_SIL}(\tilde{\mathcal{X}}^+, \tilde{\mathcal{X}}^-).$$

return Coefficient vector $\boldsymbol{\theta} \in \mathbb{R}^{1\times d}$ and the intercept $\beta \in \mathbb{R}$.

Algorithm 2 Training a mean-difference (MD) probe (Marks and Tegmark, 2024), sometimes referred to as mean-mass/mean-cluster difference classifier or linear discriminant analysis.

Input: A training dataset $\{\langle \boldsymbol{z}_i, y_i \rangle\}_{i=1}^{n}$ with binary labels $y_i \in \{0, 1\}$ and $\boldsymbol{z}_i \in \mathbb{R}^{1\times d}$. In our experiments, $\boldsymbol{z}$ is the embedding of the last token (unless otherwise noted).

Output: Parameters $\boldsymbol{\theta} \in \mathbb{R}^{1\times d}$, $\beta \in \mathbb{R}$, and $\boldsymbol{\Sigma}^{-1} \in \mathbb{R}^{d\times d}$. These parameters are subsequently given to a function f along with $\boldsymbol{z}$ to compute $f(\boldsymbol{z}) = \sigma\big(\boldsymbol{z}\,(\boldsymbol{\Sigma}^{-1}\,\boldsymbol{\theta}^{\top}) + \beta\big)$, where σ is a sigmoid function.

1: Partition data into positive and negative sets:

$$\mathcal{X}^{+} = \{\boldsymbol{z}_i : y_i = 1\}, \quad \mathcal{X}^{-} = \{\boldsymbol{z}_i : y_i = 0\}$$

2: Compute class means $\boldsymbol{\mu}^{+}$ and $\boldsymbol{\mu}^{-}$, and covariance matrices $\boldsymbol{\Sigma}^{+}$ and $\boldsymbol{\Sigma}^{-}$ for $\mathcal{X}^{+}$ and $\mathcal{X}^{-}$.

3: Compute pooled covariance matrix (where $n^{+} = |\mathcal{X}^{+}|$ and $n^{-} = |\mathcal{X}^{-}|$):

$$\boldsymbol{\Sigma} = \frac{(n^{+} - 1)\,\boldsymbol{\Sigma}^{+} + (n^{-} - 1)\,\boldsymbol{\Sigma}^{-}}{n^{+} + n^{-} - 2}$$

4: Compute the coefficient vector: $\boldsymbol{\theta} = \boldsymbol{\mu}^{+} - \boldsymbol{\mu}^{-}$, where $\boldsymbol{\theta} \in \mathbb{R}^{1\times d}$.

5: Compute scores for positive and negative sets:

$$s^{+} \leftarrow \mathcal{X}^{+}\,(\boldsymbol{\Sigma}^{-1}\boldsymbol{\theta}^{\top}) \;\text{ and }\; s^{-} \leftarrow \mathcal{X}^{-}\,(\boldsymbol{\Sigma}^{-1}\boldsymbol{\theta}^{\top}), \text{ where } s^{+} \in \mathbb{R}^{n^{+}} \text{ and } s^{-} \in \mathbb{R}^{n^{-}}.$$

6: Compute the intercept:

$$\beta = -\frac{1}{2}\left(\text{mean}\left(s^{+}\right) + \text{mean}\left(s^{-}\right)\right)$$

return Coefficient vector $\boldsymbol{\theta}$, intercept β, and the inverse covariance matrix $\boldsymbol{\Sigma}^{-1}$.

Algorithm 3 Training of Truth and Polarity Directions, aka `TTPD` probe (Bürger et al., 2024)

Input: A training dataset $\{\langle \boldsymbol{z}_i, y_i, p_i \rangle\}_{i=1}^n$ with binary truthfulness labels $y_i \in \{-1, 1\}$, binary polarity labels $p_i \in \{-1, 1\}$ that specify affirmative/negated statements and $\boldsymbol{z}_i \in \mathbb{R}^{1\times d}$. In our experiments, $\boldsymbol{z}$ is the embedding of the last token (unless otherwise noted).

Output: Parameters $\Theta \in \mathbb{R}^{2\times d}$, $\boldsymbol{\theta} \in \mathbb{R}^{1\times 2}$ and $\beta \in \mathbb{R}$. These parameters are subsequently given to a function f along with $\boldsymbol{z}$ to compute $f(\boldsymbol{z}) = \sigma\big((\boldsymbol{z}\,\Theta^\top)\,\boldsymbol{\theta}^\top + \beta\big)$, where σ is a sigmoid function.

1: Given data matrix $X = [\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n] \in \mathbb{R}^{n\times d}$ compute the centered data matrix

$$\bar{X} = X - \mathrm{mean}(X)$$

2: Find the truth direction $\boldsymbol{\theta}_t \in \mathbb{R}^{1\times d}$ via the Ordinary Least Squares:

$$\boldsymbol{\theta}_t = (\bar{X}^\top\,\bar{X})^{-1}\bar{X}^\top \boldsymbol{y}, \text{ where } \boldsymbol{y} = [y_1, \ldots, y_n]$$

3: Find the polarity direction $\boldsymbol{\theta}_p \in \mathbb{R}^{1\times d}$ via the Logistic Regression:

$$\boldsymbol{\theta}_p \leftarrow \mathrm{LogisticRegression}(X, \boldsymbol{p}), \text{ where } \boldsymbol{p} = [p_1, \ldots, p_n]$$

4: Project X onto $\boldsymbol{\theta}_t$ and $\boldsymbol{\theta}_p$:

$$\hat{X} \leftarrow \mathrm{stack}(\big[X\,\boldsymbol{\theta}_t^\top, X\,\boldsymbol{\theta}_p^\top\big]), \text{ where } \hat{X} \in \mathbb{R}^{n\times 2}$$

5: Use the *projected* data matrix $\hat{X}$ to get coefficient vector $\boldsymbol{\theta}$ and the intercept β:

$$\boldsymbol{\theta}\,,\ \beta \leftarrow \mathrm{LogisticRegression}(\hat{X}, \boldsymbol{y})$$

return Projection matrix $\Theta = \mathrm{stack}([\boldsymbol{\theta}_t, \boldsymbol{\theta}_p])$, coefficient vector $\boldsymbol{\theta}$ and intercept β.

Algorithm 4 Training a Supervised Principal Component Analysis (sPCA) probe

Input: A training dataset $\{\langle \boldsymbol{z}_i, y_i \rangle\}_{i=1}^{n}$ with binary labels $y_i \in \{0, 1\}$ and $\boldsymbol{z}_i \in \mathbb{R}^{1\times d}$, $k \in \mathbb{N}$, $k \leq d$ to specify the number of components and ridge penalty $\lambda > 0$. In our experiments, $\boldsymbol{z}$ is the embedding of the last token (unless otherwise noted).

Output: Parameters $\tilde{\Lambda} \in \mathbb{R}^{k\times d}$, $\boldsymbol{\theta} \in \mathbb{R}^{1\times k}$ and $\beta \in \mathbb{R}$. These parameters are subsequently given to a function f along with $\boldsymbol{z}$ to compute $f(\boldsymbol{z}) = \sigma\big((\boldsymbol{z}\,\tilde{\Lambda}^\top)\,\boldsymbol{\theta}^\top + \beta\big)$, where σ is a sigmoid function.

1: Partition data into positive and negative sets:

$$\mathcal{X}^+ = \{\boldsymbol{z}_i : y_i = 1\}, \quad \mathcal{X}^- = \{\boldsymbol{z}_i : y_i = 0\}$$

2: Compute means and centered matrices ($\mathcal{X} = \mathcal{X}^+ \cup \mathcal{X}^-$):

$$\boldsymbol{\mu}^+ = \mathrm{mean}(\mathcal{X}^+), \quad \boldsymbol{\mu}^- = \mathrm{mean}(\mathcal{X}^-), \quad \boldsymbol{\mu} = \mathrm{mean}(\mathcal{X}), \quad \mathcal{X}_c^+ = \mathcal{X}^+ - \boldsymbol{\mu}^+,\ \mathcal{X}_c^- = \mathcal{X}^- - \boldsymbol{\mu}^-.$$

3: Compute class covariance matrices $\boldsymbol{\Sigma}^+$ and $\boldsymbol{\Sigma}^-$ for $\mathcal{X}_c^+$ and $\mathcal{X}_c^-$.

4: Compute within-class covariance matrix (where $n^+ = |\mathcal{X}_c^+|$ and $n^- = |\mathcal{X}_c^-|$):

$$\boldsymbol{\Sigma}_w = \frac{(n^+ - 1)\,\boldsymbol{\Sigma}^+ + (n^- - 1)\,\boldsymbol{\Sigma}^-}{n^+ + n^- - 2}$$

5: Compute between-class covariance matrix:

$$\boldsymbol{d}_p = (\boldsymbol{\mu}^+ - \boldsymbol{\mu}), \quad \boldsymbol{d}_n = (\boldsymbol{\mu}^- - \boldsymbol{\mu}), \qquad \boldsymbol{\Sigma}_b = \frac{n^+\,\boldsymbol{d}_p^\top \boldsymbol{d}_p + n^-\,\boldsymbol{d}_n^\top \boldsymbol{d}_n}{n}.$$

6: Build symmetric scatter matrix with ridge penalty:

$$\boldsymbol{M} = \boldsymbol{\Sigma}_b + \boldsymbol{\Sigma}_w + \lambda \boldsymbol{I}_d, \qquad \boldsymbol{M} \leftarrow \tfrac{1}{2}(\boldsymbol{M} + \boldsymbol{M}^\top).$$

7: Compute top-k eigenpairs of M (largest algebraic), assemble the eigenmatrix $\Lambda \in \mathbb{R}^{k\times d}$ and corresponding eigenvalue vector $\boldsymbol{\nu} \in \mathbb{R}^k$:

$$(\nu_1, \boldsymbol{u}_1), \ldots, (\nu_k, \boldsymbol{u}_k) = \texttt{TopKEigs}(\boldsymbol{M}, k) \;\rightarrow\; \Lambda = [\,\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k\,]^\top, \quad \boldsymbol{\nu} = [\,\nu_1, \ldots, \nu_k\,]^\top.$$

Since $\boldsymbol{M}$ is PD (ridge $\lambda > 0$), every $\nu_i \geq \lambda > 0$, so the whitening in the next step is well-defined.

8: Project data matrix $X = [\boldsymbol{z}_1; \ldots; \boldsymbol{z}_n] \in \mathbb{R}^{n\times d}$ via the whitened eigenmatrix Λ:

$$\hat{X} = X\,\tilde{\Lambda}^\top, \text{ where } \tilde{\Lambda} = \mathrm{diag}(\boldsymbol{\nu})^{-\frac{1}{2}}\,\Lambda \in \mathbb{R}^{k\times d}, \quad \hat{X} \in \mathbb{R}^{n\times k}$$

9: Use the *projected* data matrix $\hat{X}$ to get coefficient vector $\boldsymbol{\theta}$ and the intercept β:

$$\boldsymbol{\theta}\,,\ \beta \leftarrow \mathrm{LogisticRegression}(\hat{X}, \boldsymbol{y})$$

return Projection matrix $\tilde{\Lambda}$, coefficient vector $\boldsymbol{\theta}$ and intercept β.

Algorithm 5 Inductive Conformal Predictions with binary nonconformity score.

Input: A calibration dataset $\{(\boldsymbol{z}_i, y_i)\}_{i=1}^{n} \subseteq \mathcal{D}_{cal}$, where $n = |\mathcal{D}_{cal}|$, $y_i \in \{-1, 1\}$ and $\boldsymbol{z}_i \in \mathbb{R}^{L \times d}$ ($L = 1$ in a single instance setup); a significance level $\alpha \in (0, 1)$ (target coverage $1 - \alpha$), a pretrained binary classifier g, and a new sample $\boldsymbol{z}_{new}$.

Output: Prediction set $\mathcal{Y}_{new}$

Initialize an empty score list $S \leftarrow \emptyset$
for all samples $\langle \boldsymbol{z}_i, y_i \rangle \in \mathcal{D}_{cal}$ **do**
    $\tilde{z}_i = g(\boldsymbol{z}_i)$ ▷ $\tilde{z}_i \in \mathbb{R}$ is a score
    $s_i = \exp(-y_i \cdot \tilde{z}_i)$ ▷ Nonconformity score, see Eq. A2.
    $S \leftarrow S \cup \{ s_i \}$
end for
Compute a score for the new samples: $\tilde{z}_{new} = g(\boldsymbol{z}_{new})$
Initialize an empty prediction set $\mathcal{Y}_{new} \leftarrow \emptyset$
for all $y \in \{-1, 1\}$ **do**
    $s_{new} = \exp(-y \cdot \tilde{z}_{new})$
    $\psi_y = \frac{1+|\{ s_i \in S : s_i \geq s_{new} \}|}{1+|S|}$ ▷ $\mathbb{I}(\cdot)$ is an indicator function
    if $\psi_y > \alpha$ **then**
        $\mathcal{Y}_{new} \leftarrow \mathcal{Y}_{new} \cup \{y\}$
    end if
end for
return Prediction set $\mathcal{Y}_{new}$

Algorithm 6 Inductive Conformal Predictions with multiclass nonconformity score.

Input: A calibration dataset $\{(\boldsymbol{z}_i, y_i)\}_{i=1}^{n} \subseteq \mathcal{D}_{cal}$, where $n = |\mathcal{D}_{cal}|$, $y_i \in \{1, \ldots, K\}$ (K is the number of classes) and $\boldsymbol{z}_i \in \mathbb{R}^{L \times d}$ ($L = 1$ in a single instance setup); a significance level $\alpha \in (0, 1)$ (target coverage $1 - \alpha$), a pretrained multiclass classifier g, and a new sample $\boldsymbol{z}_{new}$.

Output: Prediction set $\mathcal{Y}_{new}$

Initialize an empty score list $S \leftarrow \emptyset$
for all samples $\langle \boldsymbol{z}_i, y_i \rangle \in \mathcal{D}_{cal}$ **do**
    $\mathbf{p} = g(\boldsymbol{z}_i)$ ▷ $\mathbf{p} \in \Delta^{K-1}$ is a vector of probabilities
    $p_z = \max_{j \neq y_i} p_j$ ▷ Maximum non-target probability, where $j \in \{1, \ldots, K\}$
    $d_p = p_{y_i} - p_z$ ▷ Probability margin
    $s_i = \frac{1 - d_p}{2}$ ▷ Nonconformity score, see Eq. A3.
    $S \leftarrow S \cup \{ s_i \}$
end for
Compute probabilities for the new samples: $\boldsymbol{p}_{new} = g(\boldsymbol{z}_{new})$
Initialize empty prediction set $\mathcal{Y}_{new} \leftarrow \emptyset$
for all $y \in \{1, \ldots, K\}$ **do**
    $p_z = \max_{j \neq y} p_{new,j}$ ▷ where $j \in \{1, \ldots, K\}$ and $p_{new,j} \in \boldsymbol{p}_{new}$
    $d_p = p_{new,y} - p_z$
    $s_{new} = \frac{1 - d_p}{2}$
    $\psi_y = \frac{1 + |\{ s_i \in S : s_i \geq s_{new} \}|}{1 + |S|}$ ▷ $\mathbb{I}(\cdot)$ is an indicator function
    if $\psi_y > \alpha$ **then**
        $\mathcal{Y}_{new} \leftarrow \mathcal{Y}_{new} \cup \{y\}$
    end if
end for
return Prediction set $\mathcal{Y}_{new}$

L More Tables on Classification Performance, Generalization Performance, and Confusion Matrices

In this section, we report detailed tables of the following results.

- Classification performance for all ⟨model, dataset⟩ pairs for the setting when only the last token's representation is available, and the setting when the prediction is made over the full bag (see Tables A10 to A20);
- Generalization performance aggregated per setting, probe, and training dataset (see Table A21) and a detailed overview for the multiclass `sAwMIL` across all datasets (see Tables A22 to A24);
- Confusion matrices for each ⟨model, dataset⟩ pair (see Tables A25 to A32).

Model Name	Type	Setting	Probe	Dataset	$CI_{.025}$	MCC	$CI_{.975}$	Layer	Depth
Llama-3-8B	default	instance	sAwMIL	City Locations	0.99	**0.99**	1.00	16	0.52
Llama-3.2-3B	default	instance	sAwMIL	City Locations	0.98	**0.99**	1.00	10	0.37
Mistral-7B-v0.3	default	instance	sAwMIL	City Locations	0.99	**0.99**	1.00	17	0.55
Qwen-2.5-7B	default	instance	sAwMIL	City Locations	0.98	**0.99**	0.99	19	0.70
Qwen-2.5-14B	default	instance	sAwMIL	City Locations	0.98	**0.99**	0.99	28	0.60
Gemma-7B	default	instance	sAwMIL	City Locations	0.99	**1.00**	1.00	20	0.74
Gemma-2-9B	default	instance	sAwMIL	City Locations	0.99	**1.00**	1.00	21	0.51
Llama-3.1-8B	chat	instance	sAwMIL	City Locations	0.99	**1.00**	1.00	13	0.42
Llama-3.2-3B	chat	instance	sAwMIL	City Locations	0.98	**0.99**	0.99	12	0.44
Mistral-7B-v0.3	chat	instance	sAwMIL	City Locations	0.98	**0.99**	0.99	10	0.32
Qwen-2.5-7B	chat	instance	sAwMIL	City Locations	0.97	**0.98**	0.99	20	0.74
Qwen-2.5-14B	chat	instance	sAwMIL	City Locations	0.98	**0.99**	0.99	29	0.62
Gemma-7B	chat	instance	sAwMIL	City Locations	0.97	**0.98**	0.99	18	0.67
Gemma-2-9B	chat	instance	sAwMIL	City Locations	0.98	**0.99**	1.00	23	0.56
Bio-Medical-Llama	chat	instance	sAwMIL	City Locations	0.99	**0.99**	1.00	28	0.90
Llama3-Med42-8B	chat	instance	sAwMIL	City Locations	0.99	**0.99**	1.00	14	0.45
Llama-3-8B	default	instance	sAwMIL	Medical Indications	0.84	**0.86**	0.88	13	0.42
Llama-3.2-3B	default	instance	sAwMIL	Medical Indications	0.83	**0.85**	0.88	10	0.37
Mistral-7B-v0.3	default	instance	sAwMIL	Medical Indications	0.86	**0.88**	0.90	13	0.42
Qwen-2.5-7B	default	instance	sAwMIL	Medical Indications	0.83	**0.86**	0.88	16	0.59
Qwen-2.5-14B	default	instance	sAwMIL	Medical Indications	0.84	**0.87**	0.89	22	0.47
Gemma-7B	default	instance	sAwMIL	Medical Indications	0.84	**0.86**	0.89	17	0.63
Gemma-2-9B	default	instance	sAwMIL	Medical Indications	0.86	**0.88**	0.90	18	0.44
Llama-3.1-8B	chat	instance	sAwMIL	Medical Indications	0.85	**0.87**	0.89	18	0.58
Llama-3.2-3B	chat	instance	sAwMIL	Medical Indications	0.82	**0.85**	0.87	15	0.56
Mistral-7B-v0.3	chat	instance	sAwMIL	Medical Indications	0.86	**0.88**	0.90	16	0.52
Qwen-2.5-7B	chat	instance	sAwMIL	Medical Indications	0.83	**0.85**	0.87	17	0.63
Qwen-2.5-14B	chat	instance	sAwMIL	Medical Indications	0.86	**0.88**	0.91	23	0.49
Gemma-7B	chat	instance	sAwMIL	Medical Indications	0.80	**0.83**	0.85	15	0.56
Gemma-2-9B	chat	instance	sAwMIL	Medical Indications	0.86	**0.88**	0.90	21	0.51
Bio-Medical-Llama	chat	instance	sAwMIL	Medical Indications	0.85	**0.87**	0.89	11	0.35
Llama3-Med42-8B	chat	instance	sAwMIL	Medical Indications	0.85	**0.88**	0.90	8	0.26
Llama-3-8B	default	instance	sAwMIL	Word Definitions	0.86	**0.87**	0.89	13	0.42
Llama-3.2-3B	default	instance	sAwMIL	Word Definitions	0.83	**0.85**	0.87	10	0.37
Mistral-7B-v0.3	default	instance	sAwMIL	Word Definitions	0.85	**0.87**	0.89	13	0.42
Qwen-2.5-7B	default	instance	sAwMIL	Word Definitions	0.86	**0.88**	0.89	16	0.59
Qwen-2.5-14B	default	instance	sAwMIL	Word Definitions	0.86	**0.87**	0.89	21	0.45
Gemma-7B	default	instance	sAwMIL	Word Definitions	0.83	**0.85**	0.87	14	0.52
Gemma-2-9B	default	instance	sAwMIL	Word Definitions	0.88	**0.90**	0.91	17	0.41
Llama-3.1-8B	chat	instance	sAwMIL	Word Definitions	0.92	**0.93**	0.94	14	0.45
Llama-3.2-3B	chat	instance	sAwMIL	Word Definitions	0.85	**0.86**	0.88	12	0.44
Mistral-7B-v0.3	chat	instance	sAwMIL	Word Definitions	0.88	**0.90**	0.91	11	0.35
Qwen-2.5-7B	chat	instance	sAwMIL	Word Definitions	0.87	**0.89**	0.90	18	0.67
Qwen-2.5-14B	chat	instance	sAwMIL	Word Definitions	0.91	**0.93**	0.94	24	0.51
Gemma-7B	chat	instance	sAwMIL	Word Definitions	0.81	**0.83**	0.85	22	0.81
Gemma-2-9B	chat	instance	sAwMIL	Word Definitions	0.91	**0.92**	0.94	19	0.46
Bio-Medical-Llama	chat	instance	sAwMIL	Word Definitions	0.88	**0.90**	0.91	13	0.42
Llama3-Med42-8B	chat	instance	sAwMIL	Word Definitions	0.90	**0.92**	0.93	14	0.45

Table A10: **Classification performance of the multiclass sAwMIL probe across datasets and models** (evaluated on the *last token's representation*). This probe is trained on bag representations of the statements and evaluated using representations of the last tokens. We report Matthews Correlation Coefficient (MCC) with the 95% confidence intervals. Confidence intervals are based on bootstrapping with $n = 1,000$ samples. The **bold** values mark MCC with the confidence intervals that exclude zero. The 'Setting' column specifies the evaluation setting. The 'Layer' column specifies the layer at which the probe achieved the highest MCC, and 'Depth' denotes the relative depth of that layer within the model.

Model Name	Type	Setting	Probe	Dataset	$CI_{.025}$	MCC	$CI_{.975}$	Layer	Depth
Llama-3-8B	default	bag	sAwMIL	City Locations	0.99	**0.99**	1.00	16	0.52
Llama-3.2-3B	default	bag	sAwMIL	City Locations	0.98	**0.99**	1.00	10	0.37
Mistral-7B-v0.3	default	bag	sAwMIL	City Locations	0.99	**0.99**	1.00	17	0.55
Qwen-2.5-7B	default	bag	sAwMIL	City Locations	0.98	**0.99**	0.99	19	0.70
Qwen-2.5-14B	default	bag	sAwMIL	City Locations	0.98	**0.99**	0.99	28	0.60
Gemma-7B	default	bag	sAwMIL	City Locations	0.99	**1.00**	1.00	20	0.74
Gemma-2-9B	default	bag	sAwMIL	City Locations	0.99	**1.00**	1.00	21	0.51
Llama-3.1-8B	chat	bag	sAwMIL	City Locations	0.99	**1.00**	1.00	13	0.42
Llama-3.2-3B	chat	bag	sAwMIL	City Locations	0.98	**0.99**	0.99	12	0.44
Mistral-7B-v0.3	chat	bag	sAwMIL	City Locations	0.98	**0.99**	0.99	10	0.32
Qwen-2.5-7B	chat	bag	sAwMIL	City Locations	0.97	**0.98**	0.99	20	0.74
Qwen-2.5-14B	chat	bag	sAwMIL	City Locations	0.98	**0.99**	0.99	29	0.62
Gemma-7B	chat	bag	sAwMIL	City Locations	0.97	**0.98**	0.99	18	0.67
Gemma-2-9B	chat	bag	sAwMIL	City Locations	0.98	**0.99**	1.00	23	0.56
Bio-Medical-Llama	chat	bag	sAwMIL	City Locations	0.99	**0.99**	1.00	28	0.90
Llama3-Med42-8B	chat	bag	sAwMIL	City Locations	0.99	**0.99**	1.00	14	0.45
Llama-3-8B	default	bag	sAwMIL	Medical Indications	0.83	**0.86**	0.88	13	0.42
Llama-3.2-3B	default	bag	sAwMIL	Medical Indications	0.83	**0.85**	0.88	10	0.37
Mistral-7B-v0.3	default	bag	sAwMIL	Medical Indications	0.86	**0.88**	0.90	13	0.42
Qwen-2.5-7B	default	bag	sAwMIL	Medical Indications	0.83	**0.86**	0.88	16	0.59
Qwen-2.5-14B	default	bag	sAwMIL	Medical Indications	0.84	**0.87**	0.89	22	0.47
Gemma-7B	default	bag	sAwMIL	Medical Indications	0.84	**0.86**	0.89	17	0.63
Gemma-2-9B	default	bag	sAwMIL	Medical Indications	0.86	**0.88**	0.90	18	0.44
Llama-3.1-8B	chat	bag	sAwMIL	Medical Indications	0.85	**0.87**	0.89	18	0.58
Llama-3.2-3B	chat	bag	sAwMIL	Medical Indications	0.83	**0.85**	0.87	15	0.56
Mistral-7B-v0.3	chat	bag	sAwMIL	Medical Indications	0.86	**0.88**	0.90	16	0.52
Qwen-2.5-7B	chat	bag	sAwMIL	Medical Indications	0.82	**0.85**	0.87	17	0.63
Qwen-2.5-14B	chat	bag	sAwMIL	Medical Indications	0.86	**0.88**	0.90	23	0.49
Gemma-7B	chat	bag	sAwMIL	Medical Indications	0.80	**0.83**	0.85	15	0.56
Gemma-2-9B	chat	bag	sAwMIL	Medical Indications	0.86	**0.88**	0.90	21	0.51
Bio-Medical-Llama	chat	bag	sAwMIL	Medical Indications	0.85	**0.87**	0.89	11	0.35
Llama3-Med42-8B	chat	bag	sAwMIL	Medical Indications	0.85	**0.88**	0.90	8	0.26
Llama-3-8B	default	bag	sAwMIL	Word Definitions	0.85	**0.87**	0.89	13	0.42
Llama-3.2-3B	default	bag	sAwMIL	Word Definitions	0.83	**0.85**	0.87	10	0.37
Mistral-7B-v0.3	default	bag	sAwMIL	Word Definitions	0.85	**0.87**	0.89	13	0.42
Qwen-2.5-7B	default	bag	sAwMIL	Word Definitions	0.86	**0.88**	0.89	16	0.59
Qwen-2.5-14B	default	bag	sAwMIL	Word Definitions	0.85	**0.87**	0.89	21	0.45
Gemma-7B	default	bag	sAwMIL	Word Definitions	0.83	**0.85**	0.86	14	0.52
Gemma-2-9B	default	bag	sAwMIL	Word Definitions	0.88	**0.90**	0.91	17	0.41
Llama-3.1-8B	chat	bag	sAwMIL	Word Definitions	0.92	**0.93**	0.95	14	0.45
Llama-3.2-3B	chat	bag	sAwMIL	Word Definitions	0.85	**0.86**	0.88	12	0.44
Mistral-7B-v0.3	chat	bag	sAwMIL	Word Definitions	0.88	**0.90**	0.91	11	0.35
Qwen-2.5-7B	chat	bag	sAwMIL	Word Definitions	0.87	**0.89**	0.90	18	0.67
Qwen-2.5-14B	chat	bag	sAwMIL	Word Definitions	0.91	**0.93**	0.94	24	0.51
Gemma-7B	chat	bag	sAwMIL	Word Definitions	0.81	**0.83**	0.85	22	0.81
Gemma-2-9B	chat	bag	sAwMIL	Word Definitions	0.91	**0.92**	0.94	19	0.46
Bio-Medical-Llama	chat	bag	sAwMIL	Word Definitions	0.88	**0.90**	0.92	13	0.42
Llama3-Med42-8B	chat	bag	sAwMIL	Word Definitions	0.90	**0.92**	0.93	14	0.45

Table A11: **Classification performance of the multiclass sAwMIL probe across datasets and models** (evaluated on the *full bag*). This probe is trained and evaluated on the bag representation of the statements. We report Matthews Correlation Coefficient (MCC) with the 95% confidence intervals. Confidence intervals are based on bootstrapping with $n = 1{,}000$ samples. The **bold** values mark MCC with the confidence intervals that exclude zero. The 'Setting' column specifies the evaluation setting. The 'Layer' column specifies the layer at which the probe achieved the highest MCC, and 'Depth' denotes the relative depth of that layer within the model.

Model Name	Type	Setting	Probe	Dataset	$CI_{.025}$	MCC	$CI_{.975}$	Layer	Depth
Llama-3-8B	default	-	Zero-Shot	City Locations	0.05	**0.09**	0.12	-	-
Llama-3.2-3B	default	-	Zero-Shot	City Locations	0.09	**0.12**	0.15	-	-
Mistral-7B-v0.3	default	-	Zero-Shot	City Locations	0.00	0.00	0.00	-	-
Qwen-2.5-7B	default	-	Zero-Shot	City Locations	0.53	**0.56**	0.58	-	-
Qwen-2.5-14B	default	-	Zero-Shot	City Locations	0.61	**0.63**	0.65	-	-
Gemma-7B	default	-	Zero-Shot	City Locations	0.12	**0.15**	0.18	-	-
Gemma-2-9B	default	-	Zero-Shot	City Locations	0.50	**0.53**	0.56	-	-
Llama-3.1-8B	chat	-	Zero-Shot	City Locations	0.54	**0.56**	0.58	-	-
Llama-3.2-3B	chat	-	Zero-Shot	City Locations	0.45	**0.47**	0.50	-	-
Mistral-7B-v0.3	chat	-	Zero-Shot	City Locations	0.41	**0.44**	0.47	-	-
Qwen-2.5-7B	chat	-	Zero-Shot	City Locations	0.53	**0.55**	0.57	-	-
Qwen-2.5-14B	chat	-	Zero-Shot	City Locations	0.80	**0.82**	0.84	-	-
Gemma-7B	chat	-	Zero-Shot	City Locations	0.44	**0.46**	0.49	-	-
Gemma-2-9B	chat	-	Zero-Shot	City Locations	0.57	**0.59**	0.61	-	-
Bio-Medical-Llama	chat	-	Zero-Shot	City Locations	0.48	**0.51**	0.53	-	-
Llama3-Med42-8B	chat	-	Zero-Shot	City Locations	0.53	**0.55**	0.57	-	-
Llama-3-8B	default	-	Zero-Shot	Medical Indications	0.18	**0.22**	0.25	-	-
Llama-3.2-3B	default	-	Zero-Shot	Medical Indications	0.04	**0.08**	0.12	-	-
Mistral-7B-v0.3	default	-	Zero-Shot	Medical Indications	0.00	0.02	0.04	-	-
Qwen-2.5-7B	default	-	Zero-Shot	Medical Indications	0.31	**0.34**	0.37	-	-
Qwen-2.5-14B	default	-	Zero-Shot	Medical Indications	0.38	**0.41**	0.44	-	-
Gemma-7B	default	-	Zero-Shot	Medical Indications	0.13	**0.16**	0.19	-	-
Gemma-2-9B	default	-	Zero-Shot	Medical Indications	0.28	**0.32**	0.35	-	-
Llama-3.1-8B	chat	-	Zero-Shot	Medical Indications	0.34	**0.37**	0.40	-	-
Llama-3.2-3B	chat	-	Zero-Shot	Medical Indications	0.22	**0.25**	0.28	-	-
Mistral-7B-v0.3	chat	-	Zero-Shot	Medical Indications	0.15	**0.19**	0.22	-	-
Qwen-2.5-7B	chat	-	Zero-Shot	Medical Indications	0.30	**0.33**	0.36	-	-
Qwen-2.5-14B	chat	-	Zero-Shot	Medical Indications	0.49	**0.52**	0.55	-	-
Gemma-7B	chat	-	Zero-Shot	Medical Indications	0.27	**0.30**	0.34	-	-
Gemma-2-9B	chat	-	Zero-Shot	Medical Indications	0.41	**0.44**	0.47	-	-
Bio-Medical-Llama	chat	-	Zero-Shot	Medical Indications	0.30	**0.33**	0.37	-	-
Llama3-Med42-8B	chat	-	Zero-Shot	Medical Indications	0.38	**0.41**	0.45	-	-
Llama-3-8B	default	-	Zero-Shot	Word Definitions	0.03	**0.06**	0.09	-	-
Llama-3.2-3B	default	-	Zero-Shot	Word Definitions	-0.01	0.02	0.04	-	-
Mistral-7B-v0.3	default	-	Zero-Shot	Word Definitions	0.00	0.01	0.01	-	-
Qwen-2.5-7B	default	-	Zero-Shot	Word Definitions	0.12	**0.14**	0.17	-	-
Qwen-2.5-14B	default	-	Zero-Shot	Word Definitions	0.27	**0.30**	0.33	-	-
Gemma-7B	default	-	Zero-Shot	Word Definitions	-0.03	0.00	0.03	-	-
Gemma-2-9B	default	-	Zero-Shot	Word Definitions	0.05	**0.07**	0.10	-	-
Llama-3.1-8B	chat	-	Zero-Shot	Word Definitions	0.17	**0.20**	0.23	-	-
Llama-3.2-3B	chat	-	Zero-Shot	Word Definitions	0.05	**0.08**	0.11	-	-
Mistral-7B-v0.3	chat	-	Zero-Shot	Word Definitions	-0.09	**-0.06**	-0.03	-	-
Qwen-2.5-7B	chat	-	Zero-Shot	Word Definitions	0.13	**0.16**	0.18	-	-
Qwen-2.5-14B	chat	-	Zero-Shot	Word Definitions	0.32	**0.35**	0.37	-	-
Gemma-7B	chat	-	Zero-Shot	Word Definitions	0.11	**0.14**	0.17	-	-
Gemma-2-9B	chat	-	Zero-Shot	Word Definitions	0.24	**0.27**	0.30	-	-
Bio-Medical-Llama	chat	-	Zero-Shot	Word Definitions	0.10	**0.14**	0.17	-	-
Llama3-Med42-8B	chat	-	Zero-Shot	Word Definitions	0.15	**0.17**	0.20	-	-

Table A12: **Classification performance of zero-shot prompting across datasets and models**. We report Matthews Correlation Coefficient (MCC) with the 95% confidence intervals. Confidence intervals are based on bootstrapping with $n = 1{,}000$ samples. The **bold** values mark MCC with the confidence intervals that exclude zero. The 'Setting' column specifies the evaluation setting. The 'Layer' column specifies the layer at which the probe achieved the highest MCC, and 'Depth' denotes the relative depth of that layer within the model (do not apply to zero-shot prompting).

Model Name	Type	Setting	Probe	Dataset	$CI_{.025}$	MCC	$CI_{.975}$	Layer	Depth
Llama-3-8B	default	instance	MD+CP	City Locations	0.64	**0.66**	0.69	19	0.61
Llama-3.2-3B	default	instance	MD+CP	City Locations	0.60	**0.62**	0.64	27	1.00
Mistral-7B-v0.3	default	instance	MD+CP	City Locations	0.65	**0.68**	0.70	24	0.77
Qwen-2.5-7B	default	instance	MD+CP	City Locations	0.57	**0.60**	0.62	22	0.81
Qwen-2.5-14B	default	instance	MD+CP	City Locations	0.62	**0.65**	0.67	32	0.68
Gemma-7B	default	instance	MD+CP	City Locations	0.66	**0.68**	0.70	21	0.78
Gemma-2-9B	default	instance	MD+CP	City Locations	0.70	**0.73**	0.75	24	0.59
Llama-3.1-8B	chat	instance	MD+CP	City Locations	0.69	**0.71**	0.74	27	0.87
Llama-3.2-3B	chat	instance	MD+CP	City Locations	0.62	**0.64**	0.66	15	0.56
Mistral-7B-v0.3	chat	instance	MD+CP	City Locations	0.70	**0.72**	0.74	25	0.81
Qwen-2.5-7B	chat	instance	MD+CP	City Locations	0.60	**0.63**	0.65	19	0.70
Qwen-2.5-14B	chat	instance	MD+CP	City Locations	0.71	**0.73**	0.76	34	0.72
Gemma-7B	chat	instance	MD+CP	City Locations	0.61	**0.64**	0.66	19	0.70
Gemma-2-9B	chat	instance	MD+CP	City Locations	0.77	**0.80**	0.82	24	0.59
Bio-Medical-Llama	chat	instance	MD+CP	City Locations	0.69	**0.72**	0.74	18	0.58
Llama3-Med42-8B	chat	instance	MD+CP	City Locations	0.67	**0.69**	0.72	29	0.94
Llama-3-8B	default	instance	MD+CP	Medical Indications	0.50	**0.54**	0.57	25	0.81
Llama-3.2-3B	default	instance	MD+CP	Medical Indications	0.48	**0.51**	0.55	13	0.48
Mistral-7B-v0.3	default	instance	MD+CP	Medical Indications	0.51	**0.54**	0.57	15	0.48
Qwen-2.5-7B	default	instance	MD+CP	Medical Indications	0.54	**0.57**	0.60	23	0.85
Qwen-2.5-14B	default	instance	MD+CP	Medical Indications	0.56	**0.60**	0.63	42	0.89
Gemma-7B	default	instance	MD+CP	Medical Indications	0.53	**0.56**	0.60	21	0.78
Gemma-2-9B	default	instance	MD+CP	Medical Indications	0.52	**0.56**	0.59	16	0.39
Llama-3.1-8B	chat	instance	MD+CP	Medical Indications	0.52	**0.55**	0.58	20	0.65
Llama-3.2-3B	chat	instance	MD+CP	Medical Indications	0.48	**0.52**	0.56	14	0.52
Mistral-7B-v0.3	chat	instance	MD+CP	Medical Indications	0.50	**0.54**	0.57	24	0.77
Qwen-2.5-7B	chat	instance	MD+CP	Medical Indications	0.52	**0.55**	0.59	22	0.81
Qwen-2.5-14B	chat	instance	MD+CP	Medical Indications	0.51	**0.54**	0.57	35	0.74
Gemma-7B	chat	instance	MD+CP	Medical Indications	0.44	**0.48**	0.52	17	0.63
Gemma-2-9B	chat	instance	MD+CP	Medical Indications	0.54	**0.57**	0.60	26	0.63
Bio-Medical-Llama	chat	instance	MD+CP	Medical Indications	0.51	**0.55**	0.58	26	0.84
Llama3-Med42-8B	chat	instance	MD+CP	Medical Indications	0.52	**0.56**	0.59	17	0.55
Llama-3-8B	default	instance	MD+CP	Word Definitions	0.38	**0.41**	0.43	17	0.55
Llama-3.2-3B	default	instance	MD+CP	Word Definitions	0.36	**0.38**	0.41	8	0.30
Mistral-7B-v0.3	default	instance	MD+CP	Word Definitions	0.38	**0.40**	0.42	14	0.45
Qwen-2.5-7B	default	instance	MD+CP	Word Definitions	0.40	**0.42**	0.45	14	0.52
Qwen-2.5-14B	default	instance	MD+CP	Word Definitions	0.37	**0.40**	0.42	29	0.62
Gemma-7B	default	instance	MD+CP	Word Definitions	0.39	**0.41**	0.44	15	0.56
Gemma-2-9B	default	instance	MD+CP	Word Definitions	0.39	**0.41**	0.44	14	0.34
Llama-3.1-8B	chat	instance	MD+CP	Word Definitions	0.40	**0.42**	0.45	26	0.84
Llama-3.2-3B	chat	instance	MD+CP	Word Definitions	0.33	**0.36**	0.38	10	0.37
Mistral-7B-v0.3	chat	instance	MD+CP	Word Definitions	0.37	**0.40**	0.43	13	0.42
Qwen-2.5-7B	chat	instance	MD+CP	Word Definitions	0.40	**0.42**	0.44	18	0.67
Qwen-2.5-14B	chat	instance	MD+CP	Word Definitions	0.39	**0.41**	0.43	30	0.64
Gemma-7B	chat	instance	MD+CP	Word Definitions	0.38	**0.41**	0.44	13	0.48
Gemma-2-9B	chat	instance	MD+CP	Word Definitions	0.37	**0.40**	0.43	17	0.41
Bio-Medical-Llama	chat	instance	MD+CP	Word Definitions	0.38	**0.40**	0.42	20	0.65
Llama3-Med42-8B	chat	instance	MD+CP	Word Definitions	0.39	**0.41**	0.43	27	0.87

Table A13: **Classification performance of the mean-difference probe with conformal prediction (MD+CP) probe across datasets and models** (evaluated on the *last token's representation*). This probe is trained and evaluated using the representations of the last tokens. We report Matthews Correlation Coefficient (MCC) with the 95% confidence intervals. Confidence intervals are based on bootstrapping with $n = 1{,}000$ samples. The **bold** values mark MCC with the confidence intervals that exclude zero. The 'Setting' column specifies the evaluation setting. The 'Layer' column specifies the layer at which the probe achieved the highest MCC, and 'Depth' denotes the relative depth of that layer within the model.

Model Name	Type	Setting	Probe	Dataset	$CI_{.025}$	MCC	$CI_{.975}$	Layer	Depth
Llama-3-8B	default	bag	MD+CP	City Locations	0.00	0.00	0.00	2	0.06
Llama-3.2-3B	default	bag	MD+CP	City Locations	0.00	0.00	0.00	2	0.07
Mistral-7B-v0.3	default	bag	MD+CP	City Locations	0.08	**0.11**	0.14	9	0.29
Qwen-2.5-7B	default	bag	MD+CP	City Locations	0.06	**0.08**	0.11	12	0.44
Qwen-2.5-14B	default	bag	MD+CP	City Locations	0.07	**0.10**	0.11	37	0.79
Gemma-7B	default	bag	MD+CP	City Locations	0.29	**0.31**	0.34	7	0.26
Gemma-2-9B	default	bag	MD+CP	City Locations	0.01	**0.04**	0.06	33	0.80
Llama-3.1-8B	chat	bag	MD+CP	City Locations	0.00	0.00	0.00	2	0.06
Llama-3.2-3B	chat	bag	MD+CP	City Locations	0.00	0.00	0.00	2	0.07
Mistral-7B-v0.3	chat	bag	MD+CP	City Locations	0.06	**0.08**	0.10	16	0.52
Qwen-2.5-7B	chat	bag	MD+CP	City Locations	0.00	0.00	0.00	2	0.07
Qwen-2.5-14B	chat	bag	MD+CP	City Locations	0.00	0.00	0.00	4	0.09
Gemma-7B	chat	bag	MD+CP	City Locations	0.18	**0.21**	0.24	8	0.30
Gemma-2-9B	chat	bag	MD+CP	City Locations	0.26	**0.30**	0.34	39	0.95
Bio-Medical-Llama	chat	bag	MD+CP	City Locations	0.00	0.00	0.00	2	0.06
Llama3-Med42-8B	chat	bag	MD+CP	City Locations	0.00	0.00	0.00	2	0.06
Llama-3-8B	default	bag	MD+CP	Medical Indications	0.05	**0.09**	0.12	20	0.65
Llama-3.2-3B	default	bag	MD+CP	Medical Indications	0.21	**0.25**	0.28	11	0.41
Mistral-7B-v0.3	default	bag	MD+CP	Medical Indications	0.00	0.02	0.04	4	0.13
Qwen-2.5-7B	default	bag	MD+CP	Medical Indications	0.10	**0.14**	0.17	24	0.89
Qwen-2.5-14B	default	bag	MD+CP	Medical Indications	0.03	**0.07**	0.10	18	0.38
Gemma-7B	default	bag	MD+CP	Medical Indications	0.00	0.00	0.00	2	0.07
Gemma-2-9B	default	bag	MD+CP	Medical Indications	0.04	**0.08**	0.11	12	0.29
Llama-3.1-8B	chat	bag	MD+CP	Medical Indications	0.00	0.00	0.00	2	0.06
Llama-3.2-3B	chat	bag	MD+CP	Medical Indications	0.00	0.00	0.00	2	0.07
Mistral-7B-v0.3	chat	bag	MD+CP	Medical Indications	0.04	**0.07**	0.11	31	1.00
Qwen-2.5-7B	chat	bag	MD+CP	Medical Indications	0.00	0.00	0.00	2	0.07
Qwen-2.5-14B	chat	bag	MD+CP	Medical Indications	0.14	**0.18**	0.22	47	1.00
Gemma-7B	chat	bag	MD+CP	Medical Indications	0.00	0.00	0.00	2	0.07
Gemma-2-9B	chat	bag	MD+CP	Medical Indications	0.00	0.04	0.06	23	0.56
Bio-Medical-Llama	chat	bag	MD+CP	Medical Indications	0.05	**0.08**	0.10	18	0.58
Llama3-Med42-8B	chat	bag	MD+CP	Medical Indications	0.00	0.00	0.00	2	0.06
Llama-3-8B	default	bag	MD+CP	Word Definitions	0.01	**0.03**	0.06	22	0.71
Llama-3.2-3B	default	bag	MD+CP	Word Definitions	0.09	**0.12**	0.14	13	0.48
Mistral-7B-v0.3	default	bag	MD+CP	Word Definitions	0.02	**0.04**	0.07	20	0.65
Qwen-2.5-7B	default	bag	MD+CP	Word Definitions	-0.01	0.01	0.04	6	0.22
Qwen-2.5-14B	default	bag	MD+CP	Word Definitions	0.04	**0.06**	0.09	27	0.57
Gemma-7B	default	bag	MD+CP	Word Definitions	-0.01	0.02	0.05	24	0.89
Gemma-2-9B	default	bag	MD+CP	Word Definitions	0.05	**0.07**	0.10	19	0.46
Llama-3.1-8B	chat	bag	MD+CP	Word Definitions	0.00	0.00	0.00	2	0.06
Llama-3.2-3B	chat	bag	MD+CP	Word Definitions	0.00	0.00	0.00	2	0.07
Mistral-7B-v0.3	chat	bag	MD+CP	Word Definitions	-0.00	0.03	0.06	14	0.45
Qwen-2.5-7B	chat	bag	MD+CP	Word Definitions	0.00	0.00	0.00	2	0.07
Qwen-2.5-14B	chat	bag	MD+CP	Word Definitions	0.00	0.00	0.00	4	0.09
Gemma-7B	chat	bag	MD+CP	Word Definitions	-0.01	0.03	0.06	23	0.85
Gemma-2-9B	chat	bag	MD+CP	Word Definitions	0.18	**0.22**	0.27	38	0.93
Bio-Medical-Llama	chat	bag	MD+CP	Word Definitions	0.00	0.00	0.00	2	0.06
Llama3-Med42-8B	chat	bag	MD+CP	Word Definitions	0.00	0.00	0.00	2	0.06

Table A14: **Classification performance of the mean-difference probe with conformal prediction (MD+CP) probe across datasets and models** (evaluated on the *full bag*). This probe is trained on the representation of the last tokens and evaluated on the bag representation of the statements. We report Matthews Correlation Coefficient (MCC) with the 95% confidence intervals. Confidence intervals are based on bootstrapping with $n = 1{,}000$ samples. The **bold** values mark MCC with the confidence intervals that exclude zero. The 'Setting' column specifies the evaluation setting. The 'Layer' column specifies the layer at which the probe achieved the highest MCC, and 'Depth' denotes the relative depth of that layer within the model.

Model Name	Type	Setting	Probe	Dataset	$CI_{.025}$	MCC	$CI_{.975}$	Layer	Depth
Llama-3-8B	default	instance	TTPD+CP	City Locations	0.53	**0.55**	0.57	13	0.42
Llama-3.2-3B	default	instance	TTPD+CP	City Locations	0.51	**0.54**	0.56	12	0.44
Mistral-7B-v0.3	default	instance	TTPD+CP	City Locations	0.65	**0.67**	0.69	11	0.35
Qwen-2.5-7B	default	instance	TTPD+CP	City Locations	0.54	**0.57**	0.59	18	0.67
Qwen-2.5-14B	default	instance	TTPD+CP	City Locations	0.55	**0.58**	0.61	28	0.60
Gemma-7B	default	instance	TTPD+CP	City Locations	0.56	**0.58**	0.59	12	0.44
Gemma-2-9B	default	instance	TTPD+CP	City Locations	0.59	**0.61**	0.63	17	0.41
Llama-3.1-8B	chat	instance	TTPD+CP	City Locations	0.83	**0.85**	0.88	17	0.55
Llama-3.2-3B	chat	instance	TTPD+CP	City Locations	0.45	**0.47**	0.50	10	0.37
Mistral-7B-v0.3	chat	instance	TTPD+CP	City Locations	0.77	**0.80**	0.83	14	0.45
Qwen-2.5-7B	chat	instance	TTPD+CP	City Locations	0.69	**0.71**	0.74	15	0.56
Qwen-2.5-14B	chat	instance	TTPD+CP	City Locations	0.75	**0.78**	0.80	36	0.77
Gemma-7B	chat	instance	TTPD+CP	City Locations	0.62	**0.65**	0.67	17	0.63
Gemma-2-9B	chat	instance	TTPD+CP	City Locations	0.58	**0.64**	0.69	13	0.32
Bio-Medical-Llama	chat	instance	TTPD+CP	City Locations	0.67	**0.69**	0.72	13	0.42
Llama3-Med42-8B	chat	instance	TTPD+CP	City Locations	0.93	**0.95**	0.96	16	0.52
Llama-3-8B	default	instance	TTPD+CP	Medical Indications	0.55	**0.59**	0.63	11	0.35
Llama-3.2-3B	default	instance	TTPD+CP	Medical Indications	0.51	**0.54**	0.57	11	0.41
Mistral-7B-v0.3	default	instance	TTPD+CP	Medical Indications	0.62	**0.65**	0.69	13	0.42
Qwen-2.5-7B	default	instance	TTPD+CP	Medical Indications	0.57	**0.60**	0.64	14	0.52
Qwen-2.5-14B	default	instance	TTPD+CP	Medical Indications	0.63	**0.66**	0.70	23	0.49
Gemma-7B	default	instance	TTPD+CP	Medical Indications	0.57	**0.60**	0.64	13	0.48
Gemma-2-9B	default	instance	TTPD+CP	Medical Indications	0.60	**0.64**	0.67	17	0.41
Llama-3.1-8B	chat	instance	TTPD+CP	Medical Indications	0.69	**0.73**	0.77	16	0.52
Llama-3.2-3B	chat	instance	TTPD+CP	Medical Indications	0.54	**0.58**	0.62	11	0.41
Mistral-7B-v0.3	chat	instance	TTPD+CP	Medical Indications	0.62	**0.65**	0.68	15	0.48
Qwen-2.5-7B	chat	instance	TTPD+CP	Medical Indications	0.50	**0.54**	0.59	14	0.52
Qwen-2.5-14B	chat	instance	TTPD+CP	Medical Indications	0.63	**0.67**	0.70	28	0.60
Gemma-7B	chat	instance	TTPD+CP	Medical Indications	0.48	**0.52**	0.56	17	0.63
Gemma-2-9B	chat	instance	TTPD+CP	Medical Indications	0.58	**0.62**	0.66	24	0.59
Bio-Medical-Llama	chat	instance	TTPD+CP	Medical Indications	0.54	**0.58**	0.62	16	0.52
Llama3-Med42-8B	chat	instance	TTPD+CP	Medical Indications	0.68	**0.71**	0.75	16	0.52
Llama-3-8B	default	instance	TTPD+CP	Word Definitions	0.36	**0.39**	0.42	9	0.29
Llama-3.2-3B	default	instance	TTPD+CP	Word Definitions	0.38	**0.41**	0.43	11	0.41
Mistral-7B-v0.3	default	instance	TTPD+CP	Word Definitions	0.40	**0.43**	0.46	15	0.48
Qwen-2.5-7B	default	instance	TTPD+CP	Word Definitions	0.34	**0.37**	0.41	19	0.70
Qwen-2.5-14B	default	instance	TTPD+CP	Word Definitions	0.38	**0.42**	0.46	41	0.87
Gemma-7B	default	instance	TTPD+CP	Word Definitions	0.41	**0.43**	0.46	15	0.56
Gemma-2-9B	default	instance	TTPD+CP	Word Definitions	0.44	**0.46**	0.48	17	0.41
Llama-3.1-8B	chat	instance	TTPD+CP	Word Definitions	0.44	**0.47**	0.50	19	0.61
Llama-3.2-3B	chat	instance	TTPD+CP	Word Definitions	0.32	**0.35**	0.38	12	0.44
Mistral-7B-v0.3	chat	instance	TTPD+CP	Word Definitions	0.39	**0.42**	0.45	13	0.42
Qwen-2.5-7B	chat	instance	TTPD+CP	Word Definitions	0.38	**0.41**	0.44	15	0.56
Qwen-2.5-14B	chat	instance	TTPD+CP	Word Definitions	0.42	**0.45**	0.47	46	0.98
Gemma-7B	chat	instance	TTPD+CP	Word Definitions	0.36	**0.39**	0.42	16	0.59
Gemma-2-9B	chat	instance	TTPD+CP	Word Definitions	0.36	**0.40**	0.43	22	0.54
Bio-Medical-Llama	chat	instance	TTPD+CP	Word Definitions	0.31	**0.35**	0.38	13	0.42
Llama3-Med42-8B	chat	instance	TTPD+CP	Word Definitions	0.41	**0.44**	0.47	16	0.52

Table A15: **Classification performance of the TTPD with conformal prediction (TTPD+CP) probe across datasets and models** (evaluated on the *last token's representation*). This probe is trained and evaluated using the representations of the last tokens. We report Matthews Correlation Coefficient (MCC) with the 95% confidence intervals. Confidence intervals are based on bootstrapping with $n = 1{,}000$ samples. The **bold** values mark MCC with the confidence intervals that exclude zero. The 'Setting' column specifies the evaluation setting. The 'Layer' column specifies the layer at which the probe achieved the highest MCC, and 'Depth' denotes the relative depth of that layer within the model.

Model Name	Type	Setting	Probe	Dataset	$CI_{.025}$	MCC	$CI_{.975}$	Layer	Depth
Llama-3-8B	default	bag	TTPD+CP	City Locations	0.52	**0.54**	0.56	13	0.42
Llama-3.2-3B	default	bag	TTPD+CP	City Locations	0.47	**0.54**	0.62	24	0.89
Mistral-7B-v0.3	default	bag	TTPD+CP	City Locations	0.49	**0.51**	0.52	15	0.48
Qwen-2.5-7B	default	bag	TTPD+CP	City Locations	0.49	**0.53**	0.56	17	0.63
Qwen-2.5-14B	default	bag	TTPD+CP	City Locations	0.53	**0.55**	0.58	20	0.43
Gemma-7B	default	bag	TTPD+CP	City Locations	0.56	**0.58**	0.60	16	0.59
Gemma-2-9B	default	bag	TTPD+CP	City Locations	0.50	**0.53**	0.55	21	0.51
Llama-3.1-8B	chat	bag	TTPD+CP	City Locations	0.00	0.00	0.00	2	0.06
Llama-3.2-3B	chat	bag	TTPD+CP	City Locations	0.00	0.00	0.00	2	0.07
Mistral-7B-v0.3	chat	bag	TTPD+CP	City Locations	0.00	0.00	0.00	2	0.06
Qwen-2.5-7B	chat	bag	TTPD+CP	City Locations	0.00	0.00	0.00	2	0.07
Qwen-2.5-14B	chat	bag	TTPD+CP	City Locations	0.55	**0.59**	0.63	35	0.74
Gemma-7B	chat	bag	TTPD+CP	City Locations	0.04	**0.09**	0.15	6	0.22
Gemma-2-9B	chat	bag	TTPD+CP	City Locations	0.00	0.00	0.00	3	0.07
Bio-Medical-Llama	chat	bag	TTPD+CP	City Locations	-0.01	0.04	0.09	4	0.13
Llama3-Med42-8B	chat	bag	TTPD+CP	City Locations	0.11	**0.13**	0.15	4	0.13
Llama-3-8B	default	bag	TTPD+CP	Medical Indications	0.55	**0.61**	0.67	24	0.77
Llama-3.2-3B	default	bag	TTPD+CP	Medical Indications	0.40	**0.44**	0.48	13	0.48
Mistral-7B-v0.3	default	bag	TTPD+CP	Medical Indications	0.07	**0.22**	0.33	17	0.55
Qwen-2.5-7B	default	bag	TTPD+CP	Medical Indications	0.09	**0.17**	0.26	7	0.26
Qwen-2.5-14B	default	bag	TTPD+CP	Medical Indications	-0.03	0.08	0.18	7	0.15
Gemma-7B	default	bag	TTPD+CP	Medical Indications	0.05	**0.13**	0.20	5	0.19
Gemma-2-9B	default	bag	TTPD+CP	Medical Indications	-0.04	0.09	0.22	4	0.10
Llama-3.1-8B	chat	bag	TTPD+CP	Medical Indications	0.10	**0.16**	0.22	5	0.16
Llama-3.2-3B	chat	bag	TTPD+CP	Medical Indications	0.00	0.00	0.00	2	0.07
Mistral-7B-v0.3	chat	bag	TTPD+CP	Medical Indications	0.09	**0.13**	0.18	8	0.26
Qwen-2.5-7B	chat	bag	TTPD+CP	Medical Indications	0.00	0.00	0.00	2	0.07
Qwen-2.5-14B	chat	bag	TTPD+CP	Medical Indications	0.00	0.00	0.00	4	0.09
Gemma-7B	chat	bag	TTPD+CP	Medical Indications	0.00	0.00	0.00	2	0.07
Gemma-2-9B	chat	bag	TTPD+CP	Medical Indications	0.00	0.00	0.00	3	0.07
Bio-Medical-Llama	chat	bag	TTPD+CP	Medical Indications	0.00	0.00	0.00	2	0.06
Llama3-Med42-8B	chat	bag	TTPD+CP	Medical Indications	0.00	0.00	0.00	2	0.06
Llama-3-8B	default	bag	TTPD+CP	Word Definitions	0.00	0.00	0.00	2	0.06
Llama-3.2-3B	default	bag	TTPD+CP	Word Definitions	0.30	**0.36**	0.41	19	0.70
Mistral-7B-v0.3	default	bag	TTPD+CP	Word Definitions	0.14	**0.19**	0.25	17	0.55
Qwen-2.5-7B	default	bag	TTPD+CP	Word Definitions	0.17	**0.21**	0.26	11	0.41
Qwen-2.5-14B	default	bag	TTPD+CP	Word Definitions	0.05	**0.07**	0.10	20	0.43
Gemma-7B	default	bag	TTPD+CP	Word Definitions	0.30	**0.34**	0.38	20	0.74
Gemma-2-9B	default	bag	TTPD+CP	Word Definitions	0.31	**0.35**	0.39	26	0.63
Llama-3.1-8B	chat	bag	TTPD+CP	Word Definitions	0.00	0.00	0.00	2	0.06
Llama-3.2-3B	chat	bag	TTPD+CP	Word Definitions	0.00	0.00	0.00	2	0.07
Mistral-7B-v0.3	chat	bag	TTPD+CP	Word Definitions	0.00	0.00	0.00	2	0.06
Qwen-2.5-7B	chat	bag	TTPD+CP	Word Definitions	0.00	0.00	0.00	2	0.07
Qwen-2.5-14B	chat	bag	TTPD+CP	Word Definitions	0.00	0.00	0.00	4	0.09
Gemma-7B	chat	bag	TTPD+CP	Word Definitions	0.00	0.00	0.00	2	0.07
Gemma-2-9B	chat	bag	TTPD+CP	Word Definitions	0.33	**0.37**	0.40	40	0.98
Bio-Medical-Llama	chat	bag	TTPD+CP	Word Definitions	0.00	0.00	0.00	2	0.06
Llama3-Med42-8B	chat	bag	TTPD+CP	Word Definitions	0.00	0.00	0.00	2	0.06

Table A16: **Classification performance of the TTPD with conformal prediction (TTPD+CP) probe across datasets and models** (evaluated on the *full bag*). This probe is trained on the representation of the last tokens and evaluated on the bag representation of the statements. We report Matthews Correlation Coefficient (MCC) with the 95% confidence intervals. Confidence intervals are based on bootstrapping with $n = 1{,}000$ samples. The **bold** values mark MCC with the confidence intervals that exclude zero. The 'Setting' column specifies the evaluation setting. The 'Layer' column specifies the layer at which the probe achieved the highest MCC, and 'Depth' denotes the relative depth of that layer within the model.

Model Name	Type	Setting	Probe	Dataset	$CI_{.025}$	MCC	$CI_{.975}$	Layer	Depth
Llama-3-8B	default	instance	sPCA+CP	City Locations	0.72	**0.74**	0.77	26	0.84
Llama-3.2-3B	default	instance	sPCA+CP	City Locations	0.65	**0.67**	0.70	25	0.93
Mistral-7B-v0.3	default	instance	sPCA+CP	City Locations	0.70	**0.73**	0.75	19	0.61
Qwen-2.5-7B	default	instance	sPCA+CP	City Locations	0.67	**0.69**	0.72	21	0.78
Qwen-2.5-14B	default	instance	sPCA+CP	City Locations	0.71	**0.73**	0.76	25	0.53
Gemma-7B	default	instance	sPCA+CP	City Locations	0.70	**0.73**	0.75	18	0.67
Gemma-2-9B	default	instance	sPCA+CP	City Locations	0.68	**0.71**	0.73	21	0.51
Llama-3.1-8B	chat	instance	sPCA+CP	City Locations	0.77	**0.80**	0.82	22	0.71
Llama-3.2-3B	chat	instance	sPCA+CP	City Locations	0.60	**0.62**	0.64	12	0.44
Mistral-7B-v0.3	chat	instance	sPCA+CP	City Locations	0.71	**0.73**	0.76	20	0.65
Qwen-2.5-7B	chat	instance	sPCA+CP	City Locations	0.64	**0.66**	0.68	17	0.63
Qwen-2.5-14B	chat	instance	sPCA+CP	City Locations	0.69	**0.71**	0.73	39	0.83
Gemma-7B	chat	instance	sPCA+CP	City Locations	0.63	**0.65**	0.67	19	0.70
Gemma-2-9B	chat	instance	sPCA+CP	City Locations	0.74	**0.77**	0.79	21	0.51
Bio-Medical-Llama	chat	instance	sPCA+CP	City Locations	0.66	**0.69**	0.71	13	0.42
Llama3-Med42-8B	chat	instance	sPCA+CP	City Locations	0.81	**0.84**	0.86	31	1.00
Llama-3-8B	default	instance	sPCA+CP	Medical Indications	0.55	**0.58**	0.61	8	0.26
Llama-3.2-3B	default	instance	sPCA+CP	Medical Indications	0.54	**0.58**	0.61	15	0.56
Mistral-7B-v0.3	default	instance	sPCA+CP	Medical Indications	0.57	**0.60**	0.63	17	0.55
Qwen-2.5-7B	default	instance	sPCA+CP	Medical Indications	0.56	**0.59**	0.62	21	0.78
Qwen-2.5-14B	default	instance	sPCA+CP	Medical Indications	0.59	**0.63**	0.67	42	0.89
Gemma-7B	default	instance	sPCA+CP	Medical Indications	0.58	**0.61**	0.64	15	0.56
Gemma-2-9B	default	instance	sPCA+CP	Medical Indications	0.56	**0.58**	0.61	18	0.44
Llama-3.1-8B	chat	instance	sPCA+CP	Medical Indications	0.55	**0.58**	0.61	16	0.52
Llama-3.2-3B	chat	instance	sPCA+CP	Medical Indications	0.54	**0.58**	0.61	14	0.52
Mistral-7B-v0.3	chat	instance	sPCA+CP	Medical Indications	0.57	**0.60**	0.64	31	1.00
Qwen-2.5-7B	chat	instance	sPCA+CP	Medical Indications	0.52	**0.55**	0.58	26	0.96
Qwen-2.5-14B	chat	instance	sPCA+CP	Medical Indications	0.56	**0.59**	0.61	28	0.60
Gemma-7B	chat	instance	sPCA+CP	Medical Indications	0.47	**0.51**	0.55	23	0.85
Gemma-2-9B	chat	instance	sPCA+CP	Medical Indications	0.60	**0.63**	0.67	27	0.66
Bio-Medical-Llama	chat	instance	sPCA+CP	Medical Indications	0.54	**0.57**	0.60	31	1.00
Llama3-Med42-8B	chat	instance	sPCA+CP	Medical Indications	0.57	**0.60**	0.63	28	0.90
Llama-3-8B	default	instance	sPCA+CP	Word Definitions	0.39	**0.41**	0.43	12	0.39
Llama-3.2-3B	default	instance	sPCA+CP	Word Definitions	0.40	**0.42**	0.45	10	0.37
Mistral-7B-v0.3	default	instance	sPCA+CP	Word Definitions	0.39	**0.41**	0.43	13	0.42
Qwen-2.5-7B	default	instance	sPCA+CP	Word Definitions	0.41	**0.43**	0.46	22	0.81
Qwen-2.5-14B	default	instance	sPCA+CP	Word Definitions	0.39	**0.42**	0.44	39	0.83
Gemma-7B	default	instance	sPCA+CP	Word Definitions	0.38	**0.41**	0.43	11	0.41
Gemma-2-9B	default	instance	sPCA+CP	Word Definitions	0.41	**0.43**	0.46	18	0.44
Llama-3.1-8B	chat	instance	sPCA+CP	Word Definitions	0.40	**0.42**	0.44	13	0.42
Llama-3.2-3B	chat	instance	sPCA+CP	Word Definitions	0.36	**0.38**	0.41	9	0.33
Mistral-7B-v0.3	chat	instance	sPCA+CP	Word Definitions	0.38	**0.40**	0.42	17	0.55
Qwen-2.5-7B	chat	instance	sPCA+CP	Word Definitions	0.39	**0.41**	0.44	11	0.41
Qwen-2.5-14B	chat	instance	sPCA+CP	Word Definitions	0.39	**0.41**	0.43	39	0.83
Gemma-7B	chat	instance	sPCA+CP	Word Definitions	0.42	**0.44**	0.46	16	0.59
Gemma-2-9B	chat	instance	sPCA+CP	Word Definitions	0.40	**0.42**	0.45	19	0.46
Bio-Medical-Llama	chat	instance	sPCA+CP	Word Definitions	0.37	**0.40**	0.42	19	0.61
Llama3-Med42-8B	chat	instance	sPCA+CP	Word Definitions	0.39	**0.41**	0.42	17	0.55

Table A17: **Classification performance of the supervised PCA with conformal prediction (sPCA+CP) probe across datasets and models** (evaluated on the *last token's representation*). This probe is trained and evaluated using the representations of the last tokens. We report Matthews Correlation Coefficient (MCC) with the 95% confidence intervals. Confidence intervals are based on bootstrapping with $n = 1{,}000$ samples. The **bold** values mark MCC with the confidence intervals that exclude zero. The 'Setting' column specifies the evaluation setting. The 'Layer' column specifies the layer at which the probe achieved the highest MCC, and 'Depth' denotes the relative depth of that layer within the model.

Model Name	Type	Setting	Probe	Dataset	$CI_{.025}$	MCC	$CI_{.975}$	Layer	Depth
Llama-3-8B	default	bag	sPCA+CP	City Locations	0.38	**0.40**	0.43	8	0.26
Llama-3.2-3B	default	bag	sPCA+CP	City Locations	0.55	**0.57**	0.60	27	1.00
Mistral-7B-v0.3	default	bag	sPCA+CP	City Locations	0.00	0.02	0.04	6	0.19
Qwen-2.5-7B	default	bag	sPCA+CP	City Locations	0.44	**0.47**	0.49	15	0.56
Qwen-2.5-14B	default	bag	sPCA+CP	City Locations	0.47	**0.51**	0.55	37	0.79
Gemma-7B	default	bag	sPCA+CP	City Locations	0.01	**0.04**	0.07	6	0.22
Gemma-2-9B	default	bag	sPCA+CP	City Locations	0.11	**0.14**	0.17	32	0.78
Llama-3.1-8B	chat	bag	sPCA+CP	City Locations	0.74	**0.78**	0.82	16	0.52
Llama-3.2-3B	chat	bag	sPCA+CP	City Locations	0.00	0.00	0.00	2	0.07
Mistral-7B-v0.3	chat	bag	sPCA+CP	City Locations	0.71	**0.73**	0.76	20	0.65
Qwen-2.5-7B	chat	bag	sPCA+CP	City Locations	0.00	0.00	0.00	2	0.07
Qwen-2.5-14B	chat	bag	sPCA+CP	City Locations	0.00	0.00	0.00	4	0.09
Gemma-7B	chat	bag	sPCA+CP	City Locations	0.62	**0.64**	0.66	17	0.63
Gemma-2-9B	chat	bag	sPCA+CP	City Locations	0.00	0.00	0.00	3	0.07
Bio-Medical-Llama	chat	bag	sPCA+CP	City Locations	0.00	0.00	0.00	2	0.06
Llama3-Med42-8B	chat	bag	sPCA+CP	City Locations	0.00	0.00	0.00	2	0.06
Llama-3-8B	default	bag	sPCA+CP	Medical Indications	0.54	**0.57**	0.60	11	0.35
Llama-3.2-3B	default	bag	sPCA+CP	Medical Indications	0.47	**0.51**	0.55	13	0.48
Mistral-7B-v0.3	default	bag	sPCA+CP	Medical Indications	0.54	**0.58**	0.61	15	0.48
Qwen-2.5-7B	default	bag	sPCA+CP	Medical Indications	0.13	**0.16**	0.20	19	0.70
Qwen-2.5-14B	default	bag	sPCA+CP	Medical Indications	0.37	**0.40**	0.44	34	0.72
Gemma-7B	default	bag	sPCA+CP	Medical Indications	0.05	**0.11**	0.18	6	0.22
Gemma-2-9B	default	bag	sPCA+CP	Medical Indications	0.26	**0.30**	0.33	15	0.37
Llama-3.1-8B	chat	bag	sPCA+CP	Medical Indications	0.00	0.00	0.00	2	0.06
Llama-3.2-3B	chat	bag	sPCA+CP	Medical Indications	0.00	0.00	0.00	2	0.07
Mistral-7B-v0.3	chat	bag	sPCA+CP	Medical Indications	0.00	0.00	0.00	2	0.06
Qwen-2.5-7B	chat	bag	sPCA+CP	Medical Indications	0.00	0.00	0.00	2	0.07
Qwen-2.5-14B	chat	bag	sPCA+CP	Medical Indications	0.55	**0.58**	0.61	45	0.96
Gemma-7B	chat	bag	sPCA+CP	Medical Indications	0.00	0.00	0.00	2	0.07
Gemma-2-9B	chat	bag	sPCA+CP	Medical Indications	0.00	0.00	0.00	3	0.07
Bio-Medical-Llama	chat	bag	sPCA+CP	Medical Indications	0.00	0.00	0.00	2	0.06
Llama3-Med42-8B	chat	bag	sPCA+CP	Medical Indications	0.20	**0.23**	0.26	22	0.71
Llama-3-8B	default	bag	sPCA+CP	Word Definitions	0.01	**0.03**	0.04	9	0.29
Llama-3.2-3B	default	bag	sPCA+CP	Word Definitions	0.03	**0.06**	0.08	16	0.59
Mistral-7B-v0.3	default	bag	sPCA+CP	Word Definitions	0.14	**0.17**	0.19	19	0.61
Qwen-2.5-7B	default	bag	sPCA+CP	Word Definitions	0.01	**0.04**	0.06	6	0.22
Qwen-2.5-14B	default	bag	sPCA+CP	Word Definitions	0.02	**0.05**	0.07	14	0.30
Gemma-7B	default	bag	sPCA+CP	Word Definitions	0.06	**0.08**	0.11	20	0.74
Gemma-2-9B	default	bag	sPCA+CP	Word Definitions	0.05	**0.08**	0.10	28	0.68
Llama-3.1-8B	chat	bag	sPCA+CP	Word Definitions	0.00	0.00	0.00	2	0.06
Llama-3.2-3B	chat	bag	sPCA+CP	Word Definitions	0.00	0.00	0.00	2	0.07
Mistral-7B-v0.3	chat	bag	sPCA+CP	Word Definitions	0.00	0.03	0.06	7	0.23
Qwen-2.5-7B	chat	bag	sPCA+CP	Word Definitions	0.34	**0.37**	0.39	25	0.93
Qwen-2.5-14B	chat	bag	sPCA+CP	Word Definitions	0.00	0.00	0.00	4	0.09
Gemma-7B	chat	bag	sPCA+CP	Word Definitions	0.17	**0.19**	0.22	20	0.74
Gemma-2-9B	chat	bag	sPCA+CP	Word Definitions	0.09	**0.12**	0.14	24	0.59
Bio-Medical-Llama	chat	bag	sPCA+CP	Word Definitions	0.03	**0.06**	0.09	7	0.23
Llama3-Med42-8B	chat	bag	sPCA+CP	Word Definitions	0.00	0.00	0.00	2	0.06

Table A18: **Classification performance of the supervised PCA with conformal prediction (sPCA+CP) probe across datasets and models** (evaluated on the *full bag*). This probe is trained on the representation of the last tokens and evaluated on the bag representation of the statements. We report Matthews Correlation Coefficient (MCC) with the 95% confidence intervals. Confidence intervals are based on bootstrapping with $n = 1{,}000$ samples. The **bold** values mark MCC with the confidence intervals that exclude zero. The 'Setting' column specifies the evaluation setting. The 'Layer' column specifies the layer at which the probe achieved the highest MCC, and 'Depth' denotes the relative depth of that layer within the model.

Model Name	Type	Setting	Probe	Dataset	$CI_{.025}$	MCC	$CI_{.975}$	Layer	Depth
Llama-3.2-3B	default	instance	SVM	City Locations	0.95	**0.96**	0.97	9	0.33
Mistral-7B-v0.3	default	instance	SVM	City Locations	0.96	**0.97**	0.98	12	0.39
Qwen-2.5-7B	default	instance	SVM	City Locations	0.96	**0.97**	0.98	18	0.67
Qwen-2.5-14B	default	instance	SVM	City Locations	0.97	**0.97**	0.98	24	0.51
Gemma-7B	default	instance	SVM	City Locations	0.97	**0.98**	0.99	15	0.56
Gemma-2-9B	default	instance	SVM	City Locations	0.97	**0.98**	0.99	21	0.51
Llama-3.1-8B	chat	instance	SVM	City Locations	0.97	**0.98**	0.99	12	0.39
Llama-3.2-3B	chat	instance	SVM	City Locations	0.94	**0.95**	0.97	13	0.48
Mistral-7B-v0.3	chat	instance	SVM	City Locations	0.97	**0.98**	0.98	21	0.68
Qwen-2.5-7B	chat	instance	SVM	City Locations	0.95	**0.96**	0.97	20	0.74
Qwen-2.5-14B	chat	instance	SVM	City Locations	0.97	**0.98**	0.98	46	0.98
Gemma-7B	chat	instance	SVM	City Locations	0.95	**0.96**	0.97	15	0.56
Gemma-2-9B	chat	instance	SVM	City Locations	0.98	**0.99**	0.99	21	0.51
Bio-Medical-Llama	chat	instance	SVM	City Locations	0.96	**0.97**	0.98	14	0.45
Llama3-Med42-8B	chat	instance	SVM	City Locations	0.97	**0.98**	0.99	27	0.87
Llama-3-8B	default	instance	SVM	Medical Indications	0.82	**0.84**	0.86	14	0.45
Llama-3.2-3B	default	instance	SVM	Medical Indications	0.75	**0.78**	0.81	12	0.44
Mistral-7B-v0.3	default	instance	SVM	Medical Indications	0.81	**0.84**	0.86	11	0.35
Qwen-2.5-7B	default	instance	SVM	Medical Indications	0.80	**0.82**	0.84	17	0.63
Qwen-2.5-14B	default	instance	SVM	Medical Indications	0.83	**0.85**	0.87	24	0.51
Gemma-7B	default	instance	SVM	Medical Indications	0.80	**0.82**	0.85	16	0.59
Gemma-2-9B	default	instance	SVM	Medical Indications	0.82	**0.84**	0.86	21	0.51
Llama-3.1-8B	chat	instance	SVM	Medical Indications	0.81	**0.83**	0.86	16	0.52
Llama-3.2-3B	chat	instance	SVM	Medical Indications	0.74	**0.76**	0.79	12	0.44
Mistral-7B-v0.3	chat	instance	SVM	Medical Indications	0.79	**0.82**	0.84	14	0.45
Qwen-2.5-7B	chat	instance	SVM	Medical Indications	0.78	**0.81**	0.83	19	0.70
Qwen-2.5-14B	chat	instance	SVM	Medical Indications	0.83	**0.85**	0.88	30	0.64
Gemma-7B	chat	instance	SVM	Medical Indications	0.71	**0.74**	0.77	16	0.59
Gemma-2-9B	chat	instance	SVM	Medical Indications	0.80	**0.82**	0.84	20	0.49
Bio-Medical-Llama	chat	instance	SVM	Medical Indications	0.78	**0.81**	0.83	14	0.45
Llama3-Med42-8B	chat	instance	SVM	Medical Indications	0.81	**0.83**	0.85	25	0.81
Llama-3-8B	default	instance	SVM	Word Definitions	0.89	**0.90**	0.92	12	0.39
Llama-3.2-3B	default	instance	SVM	Word Definitions	0.86	**0.88**	0.90	12	0.44
Mistral-7B-v0.3	default	instance	SVM	Word Definitions	0.89	**0.90**	0.92	13	0.42
Qwen-2.5-7B	default	instance	SVM	Word Definitions	0.89	**0.90**	0.92	16	0.59
Qwen-2.5-14B	default	instance	SVM	Word Definitions	0.90	**0.92**	0.93	23	0.49
Gemma-7B	default	instance	SVM	Word Definitions	0.87	**0.89**	0.90	15	0.56
Gemma-2-9B	default	instance	SVM	Word Definitions	0.89	**0.91**	0.92	18	0.44
Llama-3.1-8B	chat	instance	SVM	Word Definitions	0.90	**0.91**	0.93	13	0.42
Llama-3.2-3B	chat	instance	SVM	Word Definitions	0.82	**0.84**	0.85	11	0.41
Mistral-7B-v0.3	chat	instance	SVM	Word Definitions	0.88	**0.89**	0.91	12	0.39
Qwen-2.5-7B	chat	instance	SVM	Word Definitions	0.89	**0.90**	0.92	18	0.67
Qwen-2.5-14B	chat	instance	SVM	Word Definitions	0.90	**0.91**	0.93	23	0.49
Gemma-7B	chat	instance	SVM	Word Definitions	0.87	**0.88**	0.90	17	0.63
Gemma-2-9B	chat	instance	SVM	Word Definitions	0.87	**0.89**	0.90	24	0.59
Bio-Medical-Llama	chat	instance	SVM	Word Definitions	0.86	**0.88**	0.90	13	0.42
Llama3-Med42-8B	chat	instance	SVM	Word Definitions	0.90	**0.91**	0.93	19	0.61

Table A19: **Classification performance of the multiclass SVM probe across datasets and models** (evaluated on the *last token's representation*). This probe is trained and evaluated using the representations of the last tokens. We report Matthews Correlation Coefficient (MCC) with the 95% confidence intervals. Confidence intervals are based on bootstrapping with $n = 1{,}000$ samples. The **bold** values mark MCC with the confidence intervals that exclude zero. The 'Setting' column specifies the evaluation setting. The 'Layer' column specifies the layer at which the probe achieved the highest MCC, and 'Depth' denotes the relative depth of that layer within the model.

Model Name	Type	Setting	Probe	Dataset	$CI_{.025}$	MCC	$CI_{.975}$	Layer	Depth
Llama-3-8B	default	bag	SVM	City Locations	0.52	**0.54**	0.57	8	0.26
Llama-3.2-3B	default	bag	SVM	City Locations	0.71	**0.73**	0.75	27	1.00
Mistral-7B-v0.3	default	bag	SVM	City Locations	0.86	**0.88**	0.89	10	0.32
Qwen-2.5-7B	default	bag	SVM	City Locations	0.50	**0.52**	0.54	26	0.96
Qwen-2.5-14B	default	bag	SVM	City Locations	0.38	**0.41**	0.43	40	0.85
Gemma-7B	default	bag	SVM	City Locations	0.38	**0.40**	0.42	11	0.41
Gemma-2-9B	default	bag	SVM	City Locations	0.84	**0.86**	0.88	17	0.41
Llama-3.1-8B	chat	bag	SVM	City Locations	0.00	0.00	0.00	2	0.06
Llama-3.2-3B	chat	bag	SVM	City Locations	0.49	**0.52**	0.54	25	0.93
Mistral-7B-v0.3	chat	bag	SVM	City Locations	0.76	**0.79**	0.81	21	0.68
Qwen-2.5-7B	chat	bag	SVM	City Locations	0.00	0.00	0.00	2	0.07
Qwen-2.5-14B	chat	bag	SVM	City Locations	0.36	**0.38**	0.40	22	0.47
Gemma-7B	chat	bag	SVM	City Locations	0.42	**0.44**	0.46	16	0.59
Gemma-2-9B	chat	bag	SVM	City Locations	0.91	**0.92**	0.94	26	0.63
Bio-Medical-Llama	chat	bag	SVM	City Locations	0.06	**0.09**	0.12	6	0.19
Llama3-Med42-8B	chat	bag	SVM	City Locations	0.34	**0.36**	0.38	31	1.00
Llama-3-8B	default	bag	SVM	Medical Indications	0.44	**0.48**	0.51	20	0.65
Llama-3.2-3B	default	bag	SVM	Medical Indications	0.32	**0.34**	0.36	4	0.15
Mistral-7B-v0.3	default	bag	SVM	Medical Indications	0.65	**0.68**	0.71	11	0.35
Qwen-2.5-7B	default	bag	SVM	Medical Indications	0.18	**0.21**	0.25	27	1.00
Qwen-2.5-14B	default	bag	SVM	Medical Indications	0.23	**0.26**	0.28	12	0.26
Gemma-7B	default	bag	SVM	Medical Indications	0.38	**0.40**	0.42	23	0.85
Gemma-2-9B	default	bag	SVM	Medical Indications	0.56	**0.58**	0.61	18	0.44
Llama-3.1-8B	chat	bag	SVM	Medical Indications	0.29	**0.32**	0.35	18	0.58
Llama-3.2-3B	chat	bag	SVM	Medical Indications	0.35	**0.38**	0.41	6	0.22
Mistral-7B-v0.3	chat	bag	SVM	Medical Indications	0.59	**0.62**	0.66	21	0.68
Qwen-2.5-7B	chat	bag	SVM	Medical Indications	0.00	0.00	0.00	2	0.07
Qwen-2.5-14B	chat	bag	SVM	Medical Indications	0.40	**0.43**	0.45	46	0.98
Gemma-7B	chat	bag	SVM	Medical Indications	0.37	**0.40**	0.43	9	0.33
Gemma-2-9B	chat	bag	SVM	Medical Indications	0.60	**0.63**	0.66	19	0.46
Bio-Medical-Llama	chat	bag	SVM	Medical Indications	0.27	**0.30**	0.32	4	0.13
Llama3-Med42-8B	chat	bag	SVM	Medical Indications	0.52	**0.54**	0.56	31	1.00
Llama-3-8B	default	bag	SVM	Word Definitions	0.63	**0.65**	0.66	6	0.19
Llama-3.2-3B	default	bag	SVM	Word Definitions	0.50	**0.52**	0.54	9	0.33
Mistral-7B-v0.3	default	bag	SVM	Word Definitions	0.62	**0.63**	0.65	17	0.55
Qwen-2.5-7B	default	bag	SVM	Word Definitions	0.54	**0.56**	0.58	5	0.19
Qwen-2.5-14B	default	bag	SVM	Word Definitions	0.63	**0.65**	0.66	22	0.47
Gemma-7B	default	bag	SVM	Word Definitions	0.60	**0.62**	0.64	8	0.30
Gemma-2-9B	default	bag	SVM	Word Definitions	0.77	**0.79**	0.82	27	0.66
Llama-3.1-8B	chat	bag	SVM	Word Definitions	0.65	**0.67**	0.69	12	0.39
Llama-3.2-3B	chat	bag	SVM	Word Definitions	0.37	**0.39**	0.42	12	0.44
Mistral-7B-v0.3	chat	bag	SVM	Word Definitions	0.74	**0.76**	0.78	23	0.74
Qwen-2.5-7B	chat	bag	SVM	Word Definitions	-0.00	0.03	0.07	26	0.96
Qwen-2.5-14B	chat	bag	SVM	Word Definitions	0.26	**0.29**	0.31	35	0.74
Gemma-7B	chat	bag	SVM	Word Definitions	0.56	**0.58**	0.60	25	0.93
Gemma-2-9B	chat	bag	SVM	Word Definitions	0.81	**0.83**	0.85	26	0.63
Bio-Medical-Llama	chat	bag	SVM	Word Definitions	0.34	**0.36**	0.38	10	0.32
Llama3-Med42-8B	chat	bag	SVM	Word Definitions	0.28	**0.30**	0.32	22	0.71

Table A20: **Classification performance of the multiclass SVM probe across datasets and models** (evaluated on the *full bag*). This probe is trained on the representation of the last tokens and evaluated on the bag representation of the statements. We report Matthews Correlation Coefficient (MCC) with the 95% confidence intervals. Confidence intervals are based on bootstrapping with $n = 1{,}000$ samples. The **bold** values mark MCC with the confidence intervals that exclude zero. The 'Setting' column specifies the evaluation setting. The 'Layer' column specifies the layer at which the probe achieved the highest MCC, and 'Depth' denotes the relative depth of that layer within the model.

Evaluation Setting	Probe	Training Dataset	MCC	Stand. Err.
Bag-Level	MD+CP	City Locations	0.02	0.00
Bag-Level	MD+CP	Medical Indications	0.03	0.01
Bag-Level	MD+CP	Word Definitions	0.04	0.02
Bag-Level	SVM	City Locations	0.24	0.03
Bag-Level	SVM	Medical Indications	0.37	0.04
Bag-Level	SVM	Word Definitions	0.40	0.04
Bag-Level	TTPD+CP	City Locations	0.11	0.03
Bag-Level	TTPD+CP	Medical Indications	0.09	0.03
Bag-Level	TTPD+CP	Word Definitions	0.13	0.04
Bag-Level	sAwMIL	City Locations	0.82	0.02
Bag-Level	sAwMIL	Medical Indications	0.88	0.01
Bag-Level	sAwMIL	Word Definitions	0.86	0.01
Bag-Level	sPCA+CP	City Locations	0.11	0.02
Bag-Level	sPCA+CP	Medical Indications	0.13	0.03
Bag-Level	sPCA+CP	Word Definitions	0.11	0.03
Instance-Level	MD+CP	City Locations	0.22	0.02
Instance-Level	MD+CP	Medical Indications	0.21	0.02
Instance-Level	MD+CP	Word Definitions	0.43	0.02
Instance-Level	SVM	City Locations	0.61	0.02
Instance-Level	SVM	Medical Indications	0.74	0.02
Instance-Level	SVM	Word Definitions	0.77	0.02
Instance-Level	TTPD+CP	City Locations	0.36	0.03
Instance-Level	TTPD+CP	Medical Indications	0.50	0.04
Instance-Level	TTPD+CP	Word Definitions	0.62	0.03
Instance-Level	sAwMIL	City Locations	0.82	0.02
Instance-Level	sAwMIL	Medical Indications	0.88	0.01
Instance-Level	sAwMIL	Word Definitions	0.86	0.01
Instance-Level	sPCA+CP	City Locations	0.38	0.03
Instance-Level	sPCA+CP	Medical Indications	0.42	0.03
Instance-Level	sPCA+CP	Word Definitions	0.54	0.01

Table A21: **Aggregated generalization performance.** Each probe is trained on one dataset and evaluated on the remaining two (e.g., a probe trained on *City Locations* is evaluated on *Medical Indications* and *Word Definitions*). The performance is averaged over all large language models (we pick only the performance from the best-performing layers). The performance is measured by the Matthews Correlation Coefficient (MCC) with the standard error. The 'Evaluation Setting' column indicates whether the probe was evaluated using only the representation of the last token in the statement (*Instance-Level*) or the full bag of tokens (*Bag-Level*). Multiclass probes (i.e., SVM and sAwMIL) provide higher generalization performance in both settings, and the multiclass sAwMIL achieves the highest generalization performance.

Model Name	Training Dataset	Test Dataset	$CI_{.025}$	MCC	$CI_{.975}$	Rel. Depth
Gemma-7B-it	City Locations	City Locations	0.93	**0.94**	0.95	0.59
Gemma-7B-it	City Locations	Medical Indications	0.43	**0.46**	0.49	0.63
Gemma-7B-it	City Locations	Word Definitions	0.44	**0.47**	0.50	0.63
Gemma-2-9B-it	City Locations	City Locations	0.96	**0.97**	0.98	0.56
Gemma-2-9B-it	City Locations	Medical Indications	0.67	**0.70**	0.73	0.44
Gemma-2-9B-it	City Locations	Word Definitions	0.69	**0.71**	0.74	0.49
Llama-3.2-3B-Instruct	City Locations	City Locations	0.95	**0.96**	0.97	0.52
Llama-3.2-3B-Instruct	City Locations	Medical Indications	0.59	**0.62**	0.65	0.56
Llama-3.2-3B-Instruct	City Locations	Word Definitions	0.62	**0.65**	0.67	0.48
Llama3-Med42-8B	City Locations	City Locations	0.96	**0.97**	0.98	0.42
Llama3-Med42-8B	City Locations	Medical Indications	0.75	**0.78**	0.81	0.90
Llama3-Med42-8B	City Locations	Word Definitions	0.75	**0.78**	0.80	0.45
Llama-3.1-8B-Instruct	City Locations	City Locations	0.96	**0.97**	0.98	0.48
Llama-3.1-8B-Instruct	City Locations	Medical Indications	0.73	**0.75**	0.78	0.52
Llama-3.1-8B-Instruct	City Locations	Word Definitions	0.79	**0.81**	0.83	0.42
Bio-Medical-Llama-3-8B	City Locations	City Locations	0.96	**0.97**	0.98	0.97
Bio-Medical-Llama-3-8B	City Locations	Medical Indications	0.59	**0.62**	0.65	0.42
Bio-Medical-Llama-3-8B	City Locations	Word Definitions	0.53	**0.56**	0.59	0.26
Mistral-7B-Instruct-v0.3	City Locations	City Locations	0.95	**0.96**	0.97	0.48
Mistral-7B-Instruct-v0.3	City Locations	Medical Indications	0.72	**0.75**	0.78	0.55
Mistral-7B-Instruct-v0.3	City Locations	Word Definitions	0.69	**0.71**	0.74	0.35
Qwen-2.5-7B-Instruct	City Locations	City Locations	0.94	**0.95**	0.96	0.67
Qwen-2.5-7B-Instruct	City Locations	Medical Indications	0.71	**0.74**	0.76	0.70
Qwen-2.5-7B-Instruct	City Locations	Word Definitions	0.62	**0.64**	0.67	0.70
Qwen-2.5-14B-Instruct	City Locations	City Locations	0.96	**0.97**	0.98	0.62
Qwen-2.5-14B-Instruct	City Locations	Medical Indications	0.74	**0.77**	0.80	0.64
Qwen-2.5-14B-Instruct	City Locations	Word Definitions	0.73	**0.75**	0.77	0.60
Gemma-7B	City Locations	City Locations	0.96	**0.97**	0.98	0.74
Gemma-7B	City Locations	Medical Indications	0.55	**0.58**	0.60	0.41
Gemma-7B	City Locations	Word Definitions	0.61	**0.63**	0.66	0.52
Gemma-2-9B	City Locations	City Locations	0.97	**0.98**	0.99	0.63
Gemma-2-9B	City Locations	Medical Indications	0.52	**0.55**	0.58	0.44
Gemma-2-9B	City Locations	Word Definitions	0.54	**0.56**	0.58	0.27
Llama-3.2-3B	City Locations	City Locations	0.95	**0.96**	0.97	0.37
Llama-3.2-3B	City Locations	Medical Indications	0.35	**0.38**	0.41	0.33
Llama-3.2-3B	City Locations	Word Definitions	0.48	**0.50**	0.52	0.30
Llama-3-8B	City Locations	City Locations	0.96	**0.97**	0.98	0.32
Llama-3-8B	City Locations	Medical Indications	0.59	**0.62**	0.65	0.35
Llama-3-8B	City Locations	Word Definitions	0.56	**0.58**	0.61	0.26
Mistral-7B-v0.3	City Locations	City Locations	0.96	**0.97**	0.98	0.42
Mistral-7B-v0.3	City Locations	Medical Indications	0.50	**0.53**	0.55	0.39
Mistral-7B-v0.3	City Locations	Word Definitions	0.58	**0.61**	0.63	0.39
Qwen-2.5-7B	City Locations	City Locations	0.94	**0.95**	0.96	0.70
Qwen-2.5-7B	City Locations	Medical Indications	0.48	**0.52**	0.55	0.59
Qwen-2.5-7B	City Locations	Word Definitions	0.51	**0.54**	0.57	0.59
Qwen-2.5-14B	City Locations	City Locations	0.95	**0.96**	0.97	0.79
Qwen-2.5-14B	City Locations	Medical Indications	0.59	**0.62**	0.65	0.45
Qwen-2.5-14B	City Locations	Word Definitions	0.62	**0.64**	0.67	0.43

Table A22: **Generalization performance of the multiclass `sAwMIL` trained on the *City Locations* dataset**. The performance is measured by the Matthews Correlation Coefficient (MCC) with 95% confidence intervals, based on bootstrapping with $n = 1{,}000$ samples. The **bold** values mark MCC with the confidence intervals that exclude zero. The 'Rel. Depth' column specifies the relative depth of the layer where the multiclass `sAwMIL` probe achieves the best MCC score.

Model Name	Training Dataset	Test Dataset	$CI_{.025}$	MCC	$CI_{.975}$	Rel. Depth
Gemma-7B-it	Medical Indications	City Locations	0.90	**0.91**	0.93	0.70
Gemma-7B-it	Medical Indications	Medical Indications	0.65	**0.68**	0.72	0.59
Gemma-7B-it	Medical Indications	Word Definitions	0.64	**0.67**	0.69	0.59
Gemma-2-9B-it	Medical Indications	City Locations	0.91	**0.92**	0.94	0.63
Gemma-2-9B-it	Medical Indications	Medical Indications	0.81	**0.83**	0.85	0.49
Gemma-2-9B-it	Medical Indications	Word Definitions	0.78	**0.80**	0.82	0.61
Llama-3.2-3B-Instruct	Medical Indications	City Locations	0.80	**0.82**	0.85	0.41
Llama-3.2-3B-Instruct	Medical Indications	Medical Indications	0.73	**0.76**	0.79	0.48
Llama-3.2-3B-Instruct	Medical Indications	Word Definitions	0.65	**0.67**	0.70	0.48
Llama3-Med42-8B	Medical Indications	City Locations	0.95	**0.96**	0.97	0.52
Llama3-Med42-8B	Medical Indications	Medical Indications	0.80	**0.83**	0.85	0.45
Llama3-Med42-8B	Medical Indications	Word Definitions	0.76	**0.78**	0.80	0.45
Llama-3.1-8B-Instruct	Medical Indications	City Locations	0.92	**0.93**	0.94	0.55
Llama-3.1-8B-Instruct	Medical Indications	Medical Indications	0.81	**0.83**	0.85	0.55
Llama-3.1-8B-Instruct	Medical Indications	Word Definitions	0.77	**0.79**	0.81	0.42
Bio-Medical-Llama-3-8B	Medical Indications	City Locations	0.85	**0.86**	0.88	0.39
Bio-Medical-Llama-3-8B	Medical Indications	Medical Indications	0.78	**0.81**	0.83	0.81
Bio-Medical-Llama-3-8B	Medical Indications	Word Definitions	0.70	**0.73**	0.75	0.32
Mistral-7B-Instruct-v0.3	Medical Indications	City Locations	0.93	**0.94**	0.95	0.39
Mistral-7B-Instruct-v0.3	Medical Indications	Medical Indications	0.78	**0.80**	0.83	0.45
Mistral-7B-Instruct-v0.3	Medical Indications	Word Definitions	0.74	**0.77**	0.79	0.45
Qwen-2.5-7B-Instruct	Medical Indications	City Locations	0.78	**0.80**	0.83	0.52
Qwen-2.5-7B-Instruct	Medical Indications	Medical Indications	0.74	**0.77**	0.79	0.67
Qwen-2.5-7B-Instruct	Medical Indications	Word Definitions	0.64	**0.67**	0.69	0.63
Qwen-2.5-14B-Instruct	Medical Indications	City Locations	0.90	**0.92**	0.93	0.49
Qwen-2.5-14B-Instruct	Medical Indications	Medical Indications	0.79	**0.82**	0.84	0.57
Qwen-2.5-14B-Instruct	Medical Indications	Word Definitions	0.76	**0.79**	0.81	0.57
Gemma-7B	Medical Indications	City Locations	0.60	**0.63**	0.66	0.56
Gemma-7B	Medical Indications	Medical Indications	0.75	**0.78**	0.80	0.63
Gemma-7B	Medical Indications	Word Definitions	0.55	**0.57**	0.59	0.70
Gemma-2-9B	Medical Indications	City Locations	0.83	**0.85**	0.87	0.56
Gemma-2-9B	Medical Indications	Medical Indications	0.77	**0.80**	0.82	0.44
Gemma-2-9B	Medical Indications	Word Definitions	0.64	**0.67**	0.69	0.39
Llama-3.2-3B	Medical Indications	City Locations	0.49	**0.51**	0.53	0.41
Llama-3.2-3B	Medical Indications	Medical Indications	0.74	**0.76**	0.79	0.44
Llama-3.2-3B	Medical Indications	Word Definitions	0.59	**0.62**	0.64	0.52
Llama-3-8B	Medical Indications	City Locations	0.75	**0.77**	0.80	0.45
Llama-3-8B	Medical Indications	Medical Indications	0.77	**0.80**	0.83	0.39
Llama-3-8B	Medical Indications	Word Definitions	0.63	**0.65**	0.68	0.26
Mistral-7B-v0.3	Medical Indications	City Locations	0.58	**0.60**	0.63	0.42
Mistral-7B-v0.3	Medical Indications	Medical Indications	0.77	**0.80**	0.82	0.42
Mistral-7B-v0.3	Medical Indications	Word Definitions	0.63	**0.65**	0.68	0.42
Qwen-2.5-7B	Medical Indications	City Locations	0.80	**0.82**	0.84	0.67
Qwen-2.5-7B	Medical Indications	Medical Indications	0.76	**0.78**	0.81	0.63
Qwen-2.5-7B	Medical Indications	Word Definitions	0.59	**0.62**	0.65	0.74
Qwen-2.5-14B	Medical Indications	City Locations	0.80	**0.82**	0.84	0.70
Qwen-2.5-14B	Medical Indications	Medical Indications	0.76	**0.79**	0.82	0.60
Qwen-2.5-14B	Medical Indications	Word Definitions	0.70	**0.72**	0.75	0.60

Table A23: **Generalization performance of the multiclass `sAwMIL` trained on the *Medical Indications* dataset.** The performance is measured by the Matthews Correlation Coefficient (MCC) with 95% confidence intervals, based on bootstrapping with $n = 1{,}000$ samples. The **bold** values mark MCC with the confidence intervals that exclude zero. The 'Rel. Depth' column specifies the relative depth of the layer where a multiclass `sAwMIL` probe achieves the best MCC score.

Model Name	Training Dataset	Test Dataset	$CI_{.025}$	MCC	$CI_{.975}$	Rel. Depth
Gemma-7B-it	Word Definitions	City Locations	0.90	**0.92**	0.93	0.70
Gemma-7B-it	Word Definitions	Medical Indications	0.53	**0.56**	0.60	0.67
Gemma-7B-it	Word Definitions	Word Definitions	0.78	**0.80**	0.82	0.67
Gemma-2-9B-it	Word Definitions	City Locations	0.94	**0.96**	0.97	0.56
Gemma-2-9B-it	Word Definitions	Medical Indications	0.65	**0.69**	0.71	0.41
Gemma-2-9B-it	Word Definitions	Word Definitions	0.88	**0.90**	0.91	0.54
Llama-3.2-3B-Instruct	Word Definitions	City Locations	0.84	**0.86**	0.88	0.48
Llama-3.2-3B-Instruct	Word Definitions	Medical Indications	0.60	**0.63**	0.66	0.44
Llama-3.2-3B-Instruct	Word Definitions	Word Definitions	0.85	**0.86**	0.88	0.44
Llama3-Med42-8B	Word Definitions	City Locations	0.94	**0.95**	0.97	0.35
Llama3-Med42-8B	Word Definitions	Medical Indications	0.76	**0.79**	0.81	0.45
Llama3-Med42-8B	Word Definitions	Word Definitions	0.87	**0.89**	0.91	0.45
Llama-3.1-8B-Instruct	Word Definitions	City Locations	0.95	**0.96**	0.97	0.45
Llama-3.1-8B-Instruct	Word Definitions	Medical Indications	0.72	**0.75**	0.77	0.32
Llama-3.1-8B-Instruct	Word Definitions	Word Definitions	0.90	**0.91**	0.93	0.45
Bio-Medical-Llama-3-8B	Word Definitions	City Locations	0.78	**0.81**	0.83	0.35
Bio-Medical-Llama-3-8B	Word Definitions	Medical Indications	0.70	**0.73**	0.76	0.32
Bio-Medical-Llama-3-8B	Word Definitions	Word Definitions	0.87	**0.89**	0.90	0.39
Mistral-7B-Instruct-v0.3	Word Definitions	City Locations	0.93	**0.94**	0.96	0.48
Mistral-7B-Instruct-v0.3	Word Definitions	Medical Indications	0.74	**0.77**	0.80	0.52
Mistral-7B-Instruct-v0.3	Word Definitions	Word Definitions	0.87	**0.88**	0.90	0.35
Qwen-2.5-7B-Instruct	Word Definitions	City Locations	0.85	**0.87**	0.88	0.63
Qwen-2.5-7B-Instruct	Word Definitions	Medical Indications	0.74	**0.76**	0.79	0.67
Qwen-2.5-7B-Instruct	Word Definitions	Word Definitions	0.85	**0.87**	0.88	0.63
Qwen-2.5-14B-Instruct	Word Definitions	City Locations	0.95	**0.96**	0.97	0.66
Qwen-2.5-14B-Instruct	Word Definitions	Medical Indications	0.77	**0.79**	0.82	0.66
Qwen-2.5-14B-Instruct	Word Definitions	Word Definitions	0.88	**0.90**	0.92	0.51
Gemma-7B	Word Definitions	City Locations	0.88	**0.90**	0.91	0.56
Gemma-7B	Word Definitions	Medical Indications	0.59	**0.63**	0.66	0.48
Gemma-7B	Word Definitions	Word Definitions	0.81	**0.83**	0.85	0.56
Gemma-2-9B	Word Definitions	City Locations	0.92	**0.93**	0.94	0.56
Gemma-2-9B	Word Definitions	Medical Indications	0.70	**0.73**	0.75	0.46
Gemma-2-9B	Word Definitions	Word Definitions	0.86	**0.88**	0.90	0.41
Llama-3.2-3B	Word Definitions	City Locations	0.76	**0.78**	0.81	0.41
Llama-3.2-3B	Word Definitions	Medical Indications	0.70	**0.73**	0.76	0.41
Llama-3.2-3B	Word Definitions	Word Definitions	0.79	**0.81**	0.83	0.41
Llama-3-8B	Word Definitions	City Locations	0.84	**0.86**	0.88	0.39
Llama-3-8B	Word Definitions	Medical Indications	0.74	**0.76**	0.79	0.39
Llama-3-8B	Word Definitions	Word Definitions	0.85	**0.87**	0.89	0.39
Mistral-7B-v0.3	Word Definitions	City Locations	0.77	**0.79**	0.81	0.52
Mistral-7B-v0.3	Word Definitions	Medical Indications	0.73	**0.76**	0.78	0.39
Mistral-7B-v0.3	Word Definitions	Word Definitions	0.85	**0.87**	0.89	0.45
Qwen-2.5-7B	Word Definitions	City Locations	0.90	**0.92**	0.93	0.67
Qwen-2.5-7B	Word Definitions	Medical Indications	0.72	**0.74**	0.77	0.59
Qwen-2.5-7B	Word Definitions	Word Definitions	0.85	**0.87**	0.88	0.59
Qwen-2.5-14B	Word Definitions	City Locations	0.92	**0.94**	0.95	0.45
Qwen-2.5-14B	Word Definitions	Medical Indications	0.72	**0.75**	0.78	0.66
Qwen-2.5-14B	Word Definitions	Word Definitions	0.85	**0.87**	0.88	0.47

Table A24: **Generalization performance of the multiclass `sAwMIL` trained on the *Word Definitions* dataset.** The performance is measured by the Matthews Correlation Coefficient (MCC) with 95% confidence intervals, based on bootstrapping with $n = 1{,}000$ samples. The **bold** values mark MCC with the confidence intervals that exclude zero. The 'Rel. Depth' column specifies the relative depth of the layer where a multiclass `sAwMIL` probe achieves the best MCC score.

Model ↓	Ground-truth label → Predicted → Dataset ↓	True				False				Neither			
		True	False	Neither	Abstain	True	False	Neither	Abstain	True	False	Neither	Abstain
Bio-Medical-Llama-3-8B	City Locations	0.89	0.11	0.00	0.00	0.06	0.94	0.00	0.00	0.18	0.82	0.00	0.00
	Medical Indications	0.79	0.21	0.00	0.00	0.32	0.68	0.00	0.00	0.27	0.73	0.00	0.00
	Word Definitions	0.61	0.39	0.00	0.00	0.37	0.63	0.00	0.00	0.05	0.95	0.00	0.00
Gemma-2-9B	City Locations	0.50	0.10	0.39	0.01	0.03	0.53	0.43	0.01	0.05	0.00	0.93	0.02
	Medical Indications	0.70	0.12	0.18	0.00	0.34	0.51	0.14	0.00	0.57	0.19	0.24	0.00
	Word Definitions	0.36	0.20	0.40	0.05	0.17	0.26	0.53	0.04	0.14	0.12	0.72	0.01
Gemma-2-9B-it	City Locations	0.98	0.02	0.00	0.00	0.03	0.97	0.00	0.00	0.06	0.45	0.49	0.00
	Medical Indications	0.87	0.12	0.01	0.00	0.25	0.75	0.00	0.00	0.23	0.59	0.18	0.00
	Word Definitions	0.76	0.16	0.09	0.00	0.25	0.70	0.06	0.00	0.10	0.65	0.25	0.00
Gemma-7B	City Locations	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00
	Medical Indications	0.16	0.00	0.00	0.84	0.09	0.00	0.00	0.91	0.03	0.00	0.00	0.96
	Word Definitions	0.03	0.00	0.00	0.97	0.03	0.01	0.00	0.96	0.00	0.00	0.00	1.00
Gemma-7B-it	City Locations	0.76	0.23	0.01	0.00	0.05	0.95	0.00	0.00	0.04	0.64	0.32	0.00
	Medical Indications	0.69	0.30	0.01	0.00	0.27	0.73	0.00	0.00	0.37	0.61	0.02	0.00
	Word Definitions	0.27	0.63	0.09	0.01	0.09	0.89	0.02	0.00	0.04	0.77	0.15	0.04
Llama-3-8B	City Locations	0.35	0.65	0.00	0.00	0.22	0.78	0.00	0.00	0.47	0.52	0.00	0.01
	Medical Indications	0.33	0.67	0.00	0.00	0.08	0.92	0.00	0.00	0.19	0.81	0.00	0.00
	Word Definitions	0.45	0.55	0.00	0.00	0.33	0.67	0.00	0.00	0.37	0.63	0.00	0.00
Llama-3.1-8B-Instruct	City Locations	0.95	0.05	0.00	0.00	0.03	0.97	0.00	0.00	0.08	0.65	0.27	0.00
	Medical Indications	0.54	0.46	0.00	0.00	0.07	0.93	0.00	0.00	0.13	0.86	0.01	0.00
	Word Definitions	0.54	0.46	0.00	0.00	0.21	0.79	0.00	0.00	0.06	0.94	0.00	0.00
Llama-3.2-3B	City Locations	0.29	0.71	0.00	0.00	0.11	0.89	0.00	0.00	0.43	0.57	0.00	0.00
	Medical Indications	0.46	0.54	0.00	0.00	0.34	0.66	0.00	0.00	0.50	0.50	0.00	0.00
	Word Definitions	0.48	0.52	0.00	0.00	0.44	0.56	0.00	0.00	0.46	0.54	0.00	0.00
Llama-3.2-3B-Instruct	City Locations	0.93	0.07	0.00	0.00	0.15	0.85	0.00	0.00	0.04	0.84	0.13	0.00
	Medical Indications	0.35	0.65	0.00	0.00	0.07	0.93	0.00	0.00	0.15	0.85	0.00	0.00
	Word Definitions	0.60	0.40	0.00	0.00	0.46	0.54	0.00	0.00	0.06	0.93	0.01	0.00
Llama3-Med42-8B	City Locations	0.96	0.04	0.00	0.00	0.04	0.96	0.00	0.00	0.22	0.64	0.14	0.00
	Medical Indications	0.69	0.31	0.00	0.00	0.13	0.87	0.00	0.00	0.17	0.82	0.00	0.00
	Word Definitions	0.47	0.52	0.01	0.00	0.18	0.81	0.01	0.00	0.08	0.89	0.03	0.00
Mistral-7B-Instruct-v0.3	City Locations	0.88	0.00	0.09	0.03	0.04	0.09	0.50	0.37	0.08	0.00	0.90	0.02
	Medical Indications	0.47	0.04	0.49	0.00	0.07	0.09	0.85	0.00	0.03	0.00	0.97	0.00
	Word Definitions	0.63	0.02	0.33	0.02	0.40	0.01	0.55	0.04	0.25	0.00	0.73	0.02
Mistral-7B-v0.3	City Locations	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00
	Medical Indications	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00
	Word Definitions	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00
Qwen-2.5-14B	City Locations	0.95	0.05	0.00	0.00	0.02	0.98	0.00	0.00	0.04	0.47	0.49	0.00
	Medical Indications	0.53	0.46	0.01	0.00	0.04	0.95	0.00	0.00	0.00	0.77	0.23	0.00
	Word Definitions	0.50	0.41	0.09	0.00	0.05	0.90	0.05	0.00	0.00	0.53	0.47	0.00
Qwen-2.5-14B-Instruct	City Locations	0.93	0.07	0.00	0.00	0.02	0.98	0.00	0.00	0.00	0.28	0.71	0.00
	Medical Indications	0.63	0.23	0.14	0.00	0.04	0.87	0.09	0.00	0.00	0.46	0.54	0.00
	Word Definitions	0.55	0.32	0.13	0.00	0.06	0.89	0.06	0.00	0.01	0.37	0.62	0.00
Qwen-2.5-7B	City Locations	0.92	0.07	0.01	0.00	0.03	0.96	0.01	0.00	0.09	0.68	0.23	0.00
	Medical Indications	0.56	0.44	0.00	0.00	0.11	0.89	0.00	0.00	0.25	0.75	0.00	0.00
	Word Definitions	0.60	0.39	0.01	0.00	0.31	0.67	0.02	0.00	0.35	0.64	0.01	0.00
Qwen-2.5-7B-Instruct	City Locations	0.89	0.11	0.00	0.00	0.02	0.98	0.00	0.00	0.02	0.57	0.37	0.03
	Medical Indications	0.41	0.56	0.02	0.00	0.04	0.94	0.02	0.00	0.01	0.92	0.06	0.01
	Word Definitions	0.56	0.39	0.04	0.01	0.27	0.70	0.03	0.01	0.25	0.61	0.12	0.03

Table A25: **Row-wise confusion matrices for zero-shot prompting across all ⟨model, dataset⟩ pairs** (evaluated on the full bag). Each row corresponds to a specific model and a dataset. Columns are grouped by the ground-truth labels (*True*, *False*, *Neither*) with groups of subcolumns that specify the distribution of predictions (*true*, *false*, *neither*, *abstain*). For each statement in a dataset, the predicted class is the class with the highest probability (as estimated by zero-shot prompting). The values in each group of four subcolumns sum to 1 because they are normalized counts. For example, in the first row under the *True* ground-truth label, we see that *true* predictions have the value of 0.89; that is, 89% of all the true statements are classified as true.

Model ↓	True Labels → Predicted → Dataset ↓	True				False				Neither			
		True	False	Neither	Abstain	True	False	Neither	Abstain	True	False	Neither	Abstain
Bio-Medical-Llama-3-8B	City Locations	0.89	0.01	0.10	0.00	0.01	0.92	0.07	0.00	0.34	0.23	0.42	0.00
	Medical Indications	0.68	0.13	0.19	0.00	0.09	0.75	0.16	0.00	0.21	0.58	0.20	0.00
	Word Definitions	0.73	0.09	0.18	0.00	0.09	0.72	0.18	0.00	0.42	0.36	0.22	0.00
Gemma-2-9B	City Locations	0.90	0.02	0.09	0.00	0.02	0.91	0.07	0.00	0.07	0.66	0.27	0.00
	Medical Indications	0.78	0.12	0.10	0.00	0.13	0.77	0.10	0.00	0.15	0.77	0.09	0.00
	Word Definitions	0.80	0.11	0.10	0.00	0.10	0.81	0.09	0.00	0.47	0.39	0.14	0.00
Gemma-2-9B-it	City Locations	0.89	0.09	0.02	0.00	0.05	0.93	0.02	0.00	0.79	0.19	0.02	0.00
	Medical Indications	0.70	0.11	0.19	0.00	0.11	0.73	0.17	0.00	0.42	0.38	0.20	0.00
	Word Definitions	0.75	0.11	0.14	0.00	0.11	0.75	0.14	0.00	0.36	0.42	0.23	0.00
Gemma-7B	City Locations	0.90	0.01	0.08	0.00	0.02	0.92	0.06	0.00	0.04	0.79	0.16	0.00
	Medical Indications	0.71	0.10	0.18	0.00	0.09	0.72	0.19	0.00	0.19	0.53	0.28	0.00
	Word Definitions	0.70	0.10	0.20	0.00	0.08	0.75	0.18	0.00	0.37	0.36	0.28	0.00
Gemma-7B-it	City Locations	0.94	0.03	0.03	0.00	0.06	0.90	0.04	0.00	0.32	0.54	0.14	0.00
	Medical Indications	0.53	0.11	0.36	0.00	0.12	0.61	0.27	0.00	0.19	0.47	0.34	0.00
	Word Definitions	0.75	0.11	0.13	0.00	0.11	0.78	0.11	0.00	0.31	0.48	0.22	0.00
Llama-3-8B	City Locations	0.90	0.01	0.08	0.00	0.00	0.92	0.08	0.00	0.66	0.17	0.17	0.00
	Medical Indications	0.71	0.12	0.17	0.00	0.11	0.75	0.14	0.00	0.40	0.39	0.21	0.00
	Word Definitions	0.73	0.09	0.17	0.00	0.07	0.74	0.19	0.00	0.48	0.31	0.22	0.00
Llama-3.1-8B-Instruct	City Locations	0.89	0.03	0.08	0.00	0.04	0.89	0.06	0.00	0.68	0.27	0.05	0.00
	Medical Indications	0.74	0.14	0.13	0.00	0.10	0.79	0.12	0.00	0.22	0.67	0.11	0.00
	Word Definitions	0.80	0.10	0.10	0.00	0.07	0.83	0.09	0.00	0.45	0.46	0.09	0.00
Llama-3.2-3B	City Locations	0.89	0.03	0.08	0.00	0.03	0.90	0.07	0.00	0.25	0.39	0.35	0.00
	Medical Indications	0.72	0.12	0.16	0.00	0.12	0.71	0.18	0.00	0.19	0.64	0.17	0.00
	Word Definitions	0.72	0.09	0.18	0.00	0.11	0.71	0.18	0.00	0.31	0.44	0.26	0.00
Llama-3.2-3B-Instruct	City Locations	0.87	0.05	0.08	0.00	0.04	0.90	0.06	0.00	0.85	0.06	0.09	0.00
	Medical Indications	0.59	0.11	0.30	0.00	0.11	0.62	0.27	0.00	0.33	0.38	0.30	0.00
	Word Definitions	0.59	0.10	0.31	0.00	0.09	0.62	0.29	0.00	0.37	0.40	0.24	0.00
Llama3-Med42-8B	City Locations	0.91	0.01	0.08	0.00	0.01	0.91	0.09	0.00	0.04	0.73	0.23	0.00
	Medical Indications	0.77	0.14	0.09	0.00	0.10	0.79	0.11	0.00	0.21	0.70	0.09	0.00
	Word Definitions	0.83	0.09	0.08	0.00	0.08	0.88	0.05	0.00	0.59	0.32	0.09	0.00
Mistral-7B-Instruct-v0.3	City Locations	0.88	0.01	0.11	0.00	0.01	0.90	0.09	0.00	0.45	0.40	0.16	0.00
	Medical Indications	0.68	0.11	0.21	0.00	0.13	0.71	0.16	0.00	0.29	0.50	0.20	0.00
	Word Definitions	0.77	0.13	0.10	0.00	0.11	0.81	0.08	0.00	0.24	0.60	0.15	0.00
Mistral-7B-v0.3	City Locations	0.88	0.02	0.10	0.00	0.02	0.91	0.07	0.00	0.31	0.45	0.24	0.00
	Medical Indications	0.69	0.12	0.19	0.00	0.10	0.73	0.17	0.00	0.23	0.54	0.22	0.00
	Word Definitions	0.72	0.08	0.20	0.00	0.08	0.77	0.15	0.00	0.42	0.35	0.24	0.00
Qwen-2.5-14B	City Locations	0.89	0.04	0.07	0.00	0.02	0.93	0.06	0.00	0.21	0.48	0.30	0.00
	Medical Indications	0.72	0.15	0.13	0.00	0.11	0.76	0.14	0.00	0.54	0.25	0.21	0.00
	Word Definitions	0.80	0.09	0.11	0.00	0.11	0.79	0.10	0.00	0.39	0.46	0.15	0.00
Qwen-2.5-14B-Instruct	City Locations	0.88	0.09	0.03	0.00	0.06	0.92	0.02	0.00	0.38	0.59	0.03	0.00
	Medical Indications	0.77	0.12	0.11	0.00	0.11	0.79	0.10	0.00	0.46	0.44	0.10	0.00
	Word Definitions	0.82	0.08	0.10	0.00	0.08	0.85	0.07	0.00	0.30	0.62	0.08	0.00
Qwen-2.5-7B	City Locations	0.91	0.03	0.06	0.00	0.03	0.91	0.06	0.00	0.34	0.55	0.12	0.00
	Medical Indications	0.71	0.11	0.17	0.00	0.10	0.77	0.14	0.00	0.34	0.43	0.23	0.00
	Word Definitions	0.74	0.09	0.16	0.00	0.10	0.74	0.16	0.00	0.32	0.43	0.25	0.00
Qwen-2.5-7B-Instruct	City Locations	0.90	0.03	0.07	0.00	0.02	0.92	0.06	0.00	0.62	0.24	0.14	0.00
	Medical Indications	0.75	0.09	0.15	0.00	0.10	0.78	0.12	0.00	0.67	0.20	0.13	0.00
	Word Definitions	0.78	0.10	0.12	0.00	0.09	0.83	0.08	0.00	0.43	0.44	0.13	0.00

Table A26: **Row-wise confusion matrices for mean-difference probe with conformal prediction (`MD+CP`) across all ⟨model, dataset⟩ pairs** (evaluated on the *last token representation*). Each row corresponds to a specific model and a dataset. Columns are grouped by the ground-truth labels (*True*, *False*, *Neither*) with groups of subcolumns that specify the distribution of predictions (*true*, *false*, *neither*, *abstain*). For each statement in a dataset, the predicted class is the class with the highest probability (as estimated by `MD+CP`). The values in each group of four subcolumns sum to 1 because they are normalized counts. For example, in the first row under the *True* ground-truth label, we see that *true* predictions have the value of 0.89; that is, 89% of all the true statements are classified as true.

Model ↓	True Labels → Predicted → Dataset ↓	True				False				Neither			
		True	False	Neither	Abstain	True	False	Neither	Abstain	True	False	Neither	Abstain
Bio-Medical-Llama-3-8B	City Locations	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
	Medical Indications	0.99	0.00	0.01	0.00	0.95	0.03	0.02	0.00	0.98	0.01	0.01	0.00
	Word Definitions	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
Gemma-2-9B	City Locations	1.00	0.00	0.00	0.00	0.97	0.01	0.02	0.00	0.99	0.00	0.01	0.00
	Medical Indications	0.95	0.01	0.04	0.00	0.88	0.04	0.09	0.00	0.98	0.01	0.01	0.00
	Word Definitions	0.80	0.02	0.18	0.00	0.65	0.05	0.29	0.00	0.63	0.06	0.31	0.00
Gemma-2-9B-it	City Locations	0.87	0.01	0.12	0.00	0.30	0.15	0.55	0.00	0.42	0.07	0.51	0.00
	Medical Indications	1.00	0.00	0.00	0.00	0.98	0.00	0.01	0.00	0.99	0.00	0.01	0.00
	Word Definitions	0.57	0.02	0.41	0.00	0.32	0.13	0.55	0.00	0.20	0.12	0.68	0.00
Gemma-7B	City Locations	0.95	0.03	0.02	0.00	0.48	0.46	0.06	0.00	0.76	0.21	0.02	0.00
	Medical Indications	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
	Word Definitions	0.93	0.00	0.06	0.00	0.90	0.01	0.09	0.00	0.94	0.00	0.05	0.00
Gemma-7B-it	City Locations	0.83	0.06	0.11	0.00	0.59	0.31	0.11	0.00	0.81	0.12	0.07	0.00
	Medical Indications	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
	Word Definitions	0.93	0.00	0.07	0.00	0.88	0.01	0.11	0.00	0.96	0.00	0.04	0.00
Llama-3-8B	City Locations	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
	Medical Indications	0.68	0.02	0.30	0.00	0.59	0.05	0.36	0.00	0.69	0.04	0.27	0.00
	Word Definitions	0.96	0.00	0.03	0.00	0.94	0.01	0.05	0.00	0.91	0.01	0.08	0.00
Llama-3.1-8B-Instruct	City Locations	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
	Medical Indications	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
	Word Definitions	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
Llama-3.2-3B	City Locations	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
	Medical Indications	0.83	0.03	0.14	0.00	0.56	0.22	0.23	0.00	0.95	0.02	0.03	0.00
	Word Definitions	0.92	0.01	0.07	0.00	0.75	0.09	0.16	0.00	0.92	0.02	0.07	0.00
Llama-3.2-3B-Instruct	City Locations	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
	Medical Indications	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
	Word Definitions	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
Llama3-Med42-8B	City Locations	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
	Medical Indications	0.93	0.00	0.07	0.00	0.93	0.00	0.06	0.00	0.94	0.00	0.06	0.00
	Word Definitions	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
Mistral-7B-Instruct-v0.3	City Locations	0.97	0.00	0.03	0.00	0.88	0.02	0.10	0.00	0.97	0.00	0.03	0.00
	Medical Indications	0.96	0.01	0.03	0.00	0.91	0.04	0.05	0.00	0.90	0.04	0.06	0.00
	Word Definitions	1.00	0.00	0.00	0.00	0.97	0.01	0.02	0.00	1.00	0.00	0.00	0.00
Mistral-7B-v0.3	City Locations	0.97	0.01	0.02	0.00	0.86	0.08	0.06	0.00	0.90	0.05	0.06	0.00
	Medical Indications	0.97	0.00	0.03	0.00	0.96	0.00	0.04	0.00	0.97	0.00	0.03	0.00
	Word Definitions	0.94	0.00	0.06	0.00	0.86	0.02	0.12	0.00	0.92	0.02	0.06	0.00
Qwen-2.5-14B	City Locations	1.00	0.00	0.00	0.00	0.89	0.03	0.07	0.00	0.99	0.00	0.01	0.00
	Medical Indications	0.91	0.01	0.07	0.00	0.88	0.04	0.08	0.00	0.96	0.01	0.03	0.00
	Word Definitions	0.98	0.00	0.01	0.00	0.93	0.03	0.04	0.00	0.97	0.01	0.02	0.00
Qwen-2.5-14B-Instruct	City Locations	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
	Medical Indications	0.90	0.01	0.09	0.00	0.66	0.10	0.24	0.00	0.60	0.04	0.36	0.00
	Word Definitions	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
Qwen-2.5-7B	City Locations	0.99	0.00	0.01	0.00	0.95	0.03	0.02	0.00	1.00	0.00	0.00	0.00
	Medical Indications	0.93	0.01	0.06	0.00	0.76	0.08	0.16	0.00	0.87	0.05	0.09	0.00
	Word Definitions	0.92	0.01	0.07	0.00	0.90	0.01	0.09	0.00	0.90	0.01	0.08	0.00
Qwen-2.5-7B-Instruct	City Locations	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
	Medical Indications	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
	Word Definitions	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00

Table A27: **Row-wise confusion matrices for mean-difference probe with conformal prediction (MD+CP) across all ⟨model, dataset⟩ pairs** (evaluated on the *bag*). Each row corresponds to a specific model and a dataset. Columns are grouped by the ground-truth labels (*True*, *False*, *Neither*) with groups of subcolumns that specify the distribution of predictions (*true*, *false*, *neither*, *abstain*). For each statement in a dataset, the predicted class is the class with the highest probability (as estimated by MD+CP). The values in each group of four subcolumns sum to 1 because they are normalized counts. For example, in the first row under the *True* ground-truth label, we see that *true* predictions have the value of 1.00; that is, 100% of all the true statements are classified as true.

Model ↓	True Labels → Predicted → Dataset ↓	True				False				Neither			
		True	False	Neither	Abstain	True	False	Neither	Abstain	True	False	Neither	Abstain
Bio-Medical-Llama-3-8B	City Locations	0.87	0.08	0.05	0.00	0.04	0.90	0.05	0.00	0.09	0.78	0.14	0.00
	Medical Indications	0.59	0.08	0.33	0.00	0.09	0.64	0.27	0.00	0.03	0.61	0.36	0.00
	Word Definitions	0.44	0.11	0.46	0.00	0.08	0.50	0.41	0.00	0.00	0.69	0.31	0.00
Gemma-2-9B	City Locations	0.91	0.09	0.00	0.00	0.14	0.85	0.00	0.00	0.00	1.00	0.00	0.00
	Medical Indications	0.71	0.09	0.20	0.00	0.11	0.70	0.19	0.00	0.11	0.73	0.16	0.00
	Word Definitions	0.68	0.08	0.23	0.00	0.08	0.74	0.18	0.00	0.08	0.55	0.36	0.00
Gemma-2-9B-it	City Locations	0.86	0.12	0.02	0.00	0.04	0.94	0.01	0.00	0.98	0.02	0.01	0.00
	Medical Indications	0.70	0.07	0.23	0.00	0.11	0.70	0.19	0.00	0.14	0.34	0.51	0.00
	Word Definitions	0.53	0.11	0.36	0.00	0.07	0.59	0.34	0.00	0.03	0.58	0.39	0.00
Gemma-7B	City Locations	0.90	0.05	0.06	0.00	0.08	0.88	0.03	0.00	0.03	0.95	0.02	0.00
	Medical Indications	0.68	0.08	0.25	0.00	0.09	0.68	0.23	0.00	0.10	0.52	0.38	0.00
	Word Definitions	0.75	0.09	0.16	0.00	0.09	0.79	0.12	0.00	0.13	0.66	0.21	0.00
Gemma-7B-it	City Locations	0.88	0.04	0.08	0.00	0.05	0.91	0.03	0.00	0.90	0.01	0.09	0.00
	Medical Indications	0.54	0.07	0.39	0.00	0.13	0.52	0.35	0.00	0.45	0.13	0.41	0.00
	Word Definitions	0.60	0.12	0.28	0.00	0.09	0.64	0.26	0.00	0.17	0.37	0.45	0.00
Llama-3-8B	City Locations	0.88	0.10	0.02	0.00	0.07	0.89	0.04	0.00	0.02	0.98	0.01	0.00
	Medical Indications	0.65	0.10	0.25	0.00	0.11	0.70	0.19	0.00	0.02	0.62	0.36	0.00
	Word Definitions	0.72	0.11	0.17	0.00	0.09	0.79	0.13	0.00	0.41	0.42	0.17	0.00
Llama-3.1-8B-Instruct	City Locations	0.84	0.03	0.13	0.00	0.03	0.92	0.05	0.00	0.00	0.84	0.16	0.00
	Medical Indications	0.68	0.06	0.26	0.00	0.10	0.65	0.25	0.00	0.00	0.74	0.26	0.00
	Word Definitions	0.80	0.08	0.11	0.00	0.09	0.81	0.11	0.00	0.15	0.55	0.30	0.00
Llama-3.2-3B	City Locations	0.88	0.10	0.02	0.00	0.07	0.90	0.02	0.00	0.02	0.98	0.01	0.00
	Medical Indications	0.71	0.08	0.21	0.00	0.10	0.69	0.21	0.00	0.01	0.86	0.12	0.00
	Word Definitions	0.64	0.09	0.27	0.00	0.08	0.66	0.25	0.00	0.10	0.55	0.34	0.00
Llama-3.2-3B-Instruct	City Locations	0.85	0.11	0.04	0.00	0.10	0.87	0.02	0.00	0.73	0.23	0.04	0.00
	Medical Indications	0.51	0.13	0.36	0.00	0.08	0.60	0.32	0.00	0.06	0.41	0.53	0.00
	Word Definitions	0.44	0.12	0.43	0.00	0.07	0.49	0.44	0.00	0.13	0.48	0.40	0.00
Llama3-Med42-8B	City Locations	0.89	0.04	0.07	0.00	0.05	0.92	0.02	0.00	0.22	0.66	0.12	0.00
	Medical Indications	0.69	0.09	0.22	0.00	0.10	0.74	0.16	0.00	0.06	0.07	0.87	0.00
	Word Definitions	0.71	0.10	0.19	0.00	0.07	0.78	0.15	0.00	0.02	0.82	0.17	0.00
Mistral-7B-Instruct-v0.3	City Locations	0.88	0.11	0.00	0.00	0.07	0.93	0.00	0.00	0.30	0.70	0.00	0.00
	Medical Indications	0.73	0.08	0.19	0.00	0.10	0.72	0.19	0.00	0.06	0.54	0.40	0.00
	Word Definitions	0.71	0.10	0.19	0.00	0.09	0.72	0.19	0.00	0.16	0.57	0.27	0.00
Mistral-7B-v0.3	City Locations	0.88	0.08	0.04	0.00	0.06	0.92	0.03	0.00	0.31	0.59	0.10	0.00
	Medical Indications	0.72	0.07	0.21	0.00	0.10	0.73	0.17	0.00	0.15	0.36	0.49	0.00
	Word Definitions	0.72	0.10	0.17	0.00	0.11	0.75	0.15	0.00	0.20	0.50	0.31	0.00
Qwen-2.5-14B	City Locations	0.91	0.07	0.01	0.00	0.08	0.91	0.01	0.00	0.10	0.88	0.02	0.00
	Medical Indications	0.69	0.06	0.25	0.00	0.15	0.65	0.20	0.00	0.36	0.14	0.50	0.00
	Word Definitions	0.60	0.12	0.28	0.00	0.09	0.61	0.30	0.00	0.13	0.49	0.37	0.00
Qwen-2.5-14B-Instruct	City Locations	0.88	0.10	0.03	0.00	0.07	0.91	0.02	0.00	0.39	0.54	0.07	0.00
	Medical Indications	0.80	0.08	0.13	0.00	0.13	0.75	0.11	0.00	0.03	0.55	0.42	0.00
	Word Definitions	0.72	0.10	0.19	0.00	0.11	0.75	0.15	0.00	0.04	0.71	0.25	0.00
Qwen-2.5-7B	City Locations	0.92	0.03	0.05	0.00	0.06	0.89	0.05	0.00	0.01	0.90	0.09	0.00
	Medical Indications	0.67	0.07	0.25	0.00	0.10	0.70	0.20	0.00	0.33	0.19	0.48	0.00
	Word Definitions	0.52	0.11	0.37	0.00	0.07	0.56	0.37	0.00	0.11	0.52	0.37	0.00
Qwen-2.5-7B-Instruct	City Locations	0.90	0.03	0.07	0.00	0.07	0.87	0.06	0.00	0.33	0.10	0.57	0.00
	Medical Indications	0.73	0.09	0.17	0.00	0.15	0.72	0.13	0.00	0.02	0.88	0.10	0.00
	Word Definitions	0.71	0.11	0.18	0.00	0.10	0.71	0.19	0.00	0.11	0.60	0.30	0.00

Table A28: **Row-wise confusion matrices for TTPD probe with conformal prediction (TTPD+CP) across all ⟨model, dataset⟩ pairs** (evaluated on the *last token representation*). Each row corresponds to a specific model and a dataset. Columns are grouped by the ground-truth labels (*True*, *False*, *Neither*) with groups of subcolumns that specify the distribution of predictions (*true*, *false*, *neither*, *abstain*). For each statement in a dataset, the predicted class is the class with the highest probability (as estimated by TTPD+CP). The values in each group of four subcolumns sum to 1 because they are normalized counts. For example, in the first row under the *True* ground-truth label, we see that *true* predictions have the value of 0.87; that is, 87% of all the true statements are classified as true.

Model ↓	True Labels → / Predicted → / Dataset ↓	True				False				Neither			
		True	False	Neither	Abstain	True	False	Neither	Abstain	True	False	Neither	Abstain
Bio-Medical-Llama-3-8B	City Locations	0.19	0.00	0.81	0.00	0.16	0.01	0.83	0.00	0.03	0.24	0.72	0.00
	Medical Indications	0.08	0.08	0.84	0.00	0.09	0.05	0.86	0.00	0.08	0.02	0.91	0.00
	Word Definitions	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
Gemma-2-9B	City Locations	0.92	0.08	0.00	0.00	0.14	0.85	0.00	0.00	0.00	1.00	0.00	0.00
	Medical Indications	0.12	0.03	0.85	0.00	0.08	0.03	0.89	0.00	0.00	0.02	0.98	0.00
	Word Definitions	0.30	0.08	0.62	0.00	0.11	0.16	0.72	0.00	0.00	0.79	0.21	0.00
Gemma-2-9B-it	City Locations	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
	Medical Indications	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
	Word Definitions	0.48	0.01	0.51	0.00	0.11	0.20	0.69	0.00	0.01	0.40	0.59	0.00
Gemma-7B	City Locations	0.93	0.04	0.04	0.00	0.05	0.88	0.07	0.00	0.01	0.98	0.01	0.00
	Medical Indications	0.10	0.08	0.82	0.00	0.07	0.13	0.80	0.00	0.00	0.26	0.74	0.00
	Word Definitions	0.38	0.10	0.53	0.00	0.09	0.27	0.64	0.00	0.01	0.59	0.40	0.00
Gemma-7B-it	City Locations	0.23	0.05	0.72	0.00	0.19	0.13	0.67	0.00	0.01	0.51	0.48	0.00
	Medical Indications	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
	Word Definitions	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
Llama-3-8B	City Locations	0.88	0.11	0.00	0.00	0.11	0.88	0.01	0.00	0.02	0.98	0.00	0.00
	Medical Indications	0.54	0.08	0.37	0.00	0.10	0.54	0.36	0.00	0.01	0.66	0.32	0.00
	Word Definitions	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
Llama-3.1-8B-Instruct	City Locations	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
	Medical Indications	0.18	0.10	0.73	0.00	0.08	0.14	0.78	0.00	0.20	0.32	0.48	0.00
	Word Definitions	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
Llama-3.2-3B	City Locations	0.91	0.07	0.02	0.00	0.07	0.90	0.03	0.00	0.02	0.98	0.01	0.00
	Medical Indications	0.64	0.05	0.31	0.00	0.10	0.18	0.72	0.00	0.01	0.56	0.43	0.00
	Word Definitions	0.38	0.03	0.59	0.00	0.09	0.08	0.82	0.00	0.01	0.16	0.83	0.00
Llama-3.2-3B-Instruct	City Locations	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
	Medical Indications	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
	Word Definitions	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
Llama3-Med42-8B	City Locations	0.82	0.01	0.17	0.00	0.41	0.07	0.52	0.00	0.51	0.47	0.02	0.00
	Medical Indications	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
	Word Definitions	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
Mistral-7B-Instruct-v0.3	City Locations	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
	Medical Indications	0.35	0.05	0.60	0.00	0.24	0.11	0.65	0.00	0.23	0.32	0.45	0.00
	Word Definitions	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
Mistral-7B-v0.3	City Locations	0.88	0.06	0.06	0.00	0.24	0.56	0.20	0.00	0.34	0.50	0.16	0.00
	Medical Indications	0.11	0.14	0.75	0.00	0.06	0.13	0.81	0.00	0.00	0.39	0.61	0.00
	Word Definitions	0.76	0.00	0.24	0.00	0.31	0.01	0.68	0.00	0.34	0.08	0.58	0.00
Qwen-2.5-14B	City Locations	0.90	0.04	0.06	0.00	0.05	0.87	0.08	0.00	0.05	0.88	0.08	0.00
	Medical Indications	0.11	0.02	0.87	0.00	0.09	0.02	0.89	0.00	0.01	0.07	0.93	0.00
	Word Definitions	0.52	0.00	0.47	0.00	0.29	0.01	0.70	0.00	0.29	0.12	0.59	0.00
Qwen-2.5-14B-Instruct	City Locations	0.87	0.00	0.13	0.00	0.02	0.17	0.81	0.00	0.26	0.01	0.73	0.00
	Medical Indications	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
	Word Definitions	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
Qwen-2.5-7B	City Locations	0.91	0.06	0.03	0.00	0.11	0.85	0.05	0.00	0.28	0.66	0.06	0.00
	Medical Indications	0.10	0.09	0.81	0.00	0.07	0.08	0.85	0.00	0.00	0.13	0.87	0.00
	Word Definitions	0.39	0.01	0.60	0.00	0.11	0.03	0.86	0.00	0.09	0.11	0.81	0.00
Qwen-2.5-7B-Instruct	City Locations	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
	Medical Indications	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
	Word Definitions	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00

Table A29: **Row-wise confusion matrices for TTPD probe with conformal prediction (TTPD+CP) across all ⟨model-dataset pairs⟩** (evaluated on the *bag*). Each row corresponds to a specific model and a dataset. Columns are grouped by the ground-truth labels (*True*, *False*, *Neither*) with groups of subcolumns that specify the distribution of predictions (*true*, *false*, *neither*, *abstain*). For each statement in a dataset, the predicted class is the class with the highest probability (as estimated by TTPD+CP). The values in each group of four subcolumns sum to 1 because they are normalized counts. For example, in the first row under the *True* ground-truth label, we see that *true* predictions have the value of 0.19; that is, 19% of all the true statements are classified as true.

Model ↓	True Labels → Predicted → Dataset ↓	True				False				Neither			
		True	False	Neither	Abstain	True	False	Neither	Abstain	True	False	Neither	Abstain
Bio-Medical-Llama-3-8B	City Locations	0.90	0.02	0.09	0.00	0.01	0.92	0.07	0.00	0.04	0.75	0.21	0.00
	Medical Indications	0.74	0.10	0.16	0.00	0.09	0.79	0.13	0.00	0.06	0.79	0.15	0.00
	Word Definitions	0.80	0.11	0.09	0.00	0.10	0.83	0.07	0.00	0.39	0.54	0.07	0.00
Gemma-2-9B	City Locations	0.89	0.03	0.08	0.00	0.01	0.93	0.06	0.00	0.00	0.98	0.02	0.00
	Medical Indications	0.83	0.11	0.06	0.00	0.09	0.84	0.07	0.00	0.03	0.86	0.11	0.00
	Word Definitions	0.87	0.10	0.03	0.00	0.08	0.89	0.04	0.00	0.42	0.50	0.08	0.00
Gemma-2-9B-it	City Locations	0.86	0.10	0.03	0.00	0.05	0.93	0.02	0.00	0.47	0.49	0.04	0.00
	Medical Indications	0.79	0.10	0.10	0.00	0.07	0.81	0.12	0.00	0.41	0.28	0.31	0.00
	Word Definitions	0.79	0.13	0.08	0.00	0.09	0.86	0.05	0.00	0.38	0.52	0.10	0.00
Gemma-7B	City Locations	0.89	0.00	0.11	0.00	0.00	0.91	0.09	0.00	0.14	0.29	0.57	0.00
	Medical Indications	0.78	0.09	0.13	0.00	0.08	0.81	0.11	0.00	0.23	0.49	0.27	0.00
	Word Definitions	0.84	0.08	0.08	0.00	0.09	0.83	0.07	0.00	0.42	0.42	0.16	0.00
Gemma-7B-it	City Locations	0.86	0.04	0.10	0.00	0.02	0.92	0.07	0.00	0.10	0.60	0.30	0.00
	Medical Indications	0.62	0.09	0.28	0.00	0.10	0.66	0.24	0.00	0.46	0.46	0.07	0.00
	Word Definitions	0.80	0.08	0.12	0.00	0.07	0.84	0.10	0.00	0.33	0.49	0.18	0.00
Llama-3-8B	City Locations	0.88	0.02	0.10	0.00	0.01	0.91	0.08	0.00	0.70	0.12	0.19	0.00
	Medical Indications	0.84	0.12	0.05	0.00	0.11	0.85	0.04	0.00	0.35	0.56	0.09	0.00
	Word Definitions	0.85	0.09	0.07	0.00	0.08	0.86	0.07	0.00	0.49	0.43	0.08	0.00
Llama-3.1-8B-Instruct	City Locations	0.88	0.12	0.01	0.00	0.07	0.93	0.00	0.00	0.52	0.48	0.00	0.00
	Medical Indications	0.83	0.11	0.06	0.00	0.08	0.84	0.08	0.00	0.25	0.59	0.15	0.00
	Word Definitions	0.89	0.07	0.03	0.00	0.07	0.91	0.02	0.00	0.48	0.48	0.04	0.00
Llama-3.2-3B	City Locations	0.92	0.04	0.04	0.00	0.03	0.93	0.04	0.00	0.39	0.58	0.03	0.00
	Medical Indications	0.73	0.11	0.16	0.00	0.09	0.79	0.13	0.00	0.07	0.72	0.21	0.00
	Word Definitions	0.77	0.07	0.16	0.00	0.09	0.81	0.10	0.00	0.28	0.49	0.22	0.00
Llama-3.2-3B-Instruct	City Locations	0.88	0.12	0.00	0.00	0.09	0.91	0.00	0.00	0.89	0.11	0.00	0.00
	Medical Indications	0.67	0.10	0.23	0.00	0.09	0.70	0.20	0.00	0.26	0.49	0.25	0.00
	Word Definitions	0.69	0.11	0.20	0.00	0.07	0.76	0.17	0.00	0.37	0.43	0.21	0.00
Llama3-Med42-8B	City Locations	0.91	0.01	0.08	0.00	0.00	0.92	0.08	0.00	0.26	0.16	0.57	0.00
	Medical Indications	0.80	0.12	0.07	0.00	0.08	0.87	0.05	0.00	0.24	0.65	0.11	0.00
	Word Definitions	0.91	0.09	0.00	0.00	0.07	0.92	0.00	0.00	0.51	0.49	0.00	0.00
Mistral-7B-Instruct-v0.3	City Locations	0.90	0.04	0.06	0.00	0.04	0.89	0.06	0.00	0.46	0.50	0.03	0.00
	Medical Indications	0.79	0.11	0.10	0.00	0.09	0.81	0.10	0.00	0.48	0.46	0.07	0.00
	Word Definitions	0.86	0.13	0.01	0.00	0.11	0.88	0.01	0.00	0.43	0.56	0.02	0.00
Mistral-7B-v0.3	City Locations	0.86	0.00	0.14	0.00	0.00	0.91	0.08	0.00	0.17	0.45	0.38	0.00
	Medical Indications	0.80	0.10	0.10	0.00	0.10	0.84	0.07	0.00	0.42	0.46	0.12	0.00
	Word Definitions	0.86	0.09	0.05	0.00	0.09	0.87	0.03	0.00	0.43	0.48	0.09	0.00
Qwen-2.5-14B	City Locations	0.92	0.01	0.07	0.00	0.01	0.93	0.06	0.00	0.13	0.47	0.40	0.00
	Medical Indications	0.85	0.11	0.04	0.00	0.08	0.88	0.05	0.00	0.37	0.56	0.07	0.00
	Word Definitions	0.89	0.10	0.01	0.00	0.10	0.89	0.01	0.00	0.48	0.49	0.03	0.00
Qwen-2.5-14B-Instruct	City Locations	0.93	0.04	0.03	0.00	0.05	0.92	0.03	0.00	0.51	0.49	0.01	0.00
	Medical Indications	0.87	0.11	0.02	0.00	0.09	0.89	0.02	0.00	0.23	0.71	0.06	0.00
	Word Definitions	0.90	0.10	0.00	0.00	0.09	0.91	0.00	0.00	0.35	0.65	0.00	0.00
Qwen-2.5-7B	City Locations	0.89	0.06	0.06	0.00	0.03	0.92	0.04	0.00	0.52	0.46	0.02	0.00
	Medical Indications	0.80	0.11	0.10	0.00	0.08	0.83	0.09	0.00	0.39	0.47	0.14	0.00
	Word Definitions	0.87	0.10	0.03	0.00	0.08	0.89	0.03	0.00	0.40	0.53	0.07	0.00
Qwen-2.5-7B-Instruct	City Locations	0.89	0.02	0.09	0.00	0.02	0.91	0.07	0.00	0.50	0.43	0.06	0.00
	Medical Indications	0.80	0.11	0.08	0.00	0.12	0.82	0.06	0.00	0.44	0.49	0.06	0.00
	Word Definitions	0.85	0.10	0.05	0.00	0.10	0.88	0.03	0.00	0.45	0.51	0.04	0.00

Table A30: **Row-wise confusion matrices for supervised PCA probe with conformal prediction (sPCA+CP) across all ⟨model-dataset pairs⟩** (evaluated on the *last token representation*). Each row corresponds to a specific model and a dataset. Columns are grouped by the ground-truth labels (*True*, *False*, *Neither*) with groups of subcolumns that specify the distribution of predictions (*true*, *false*, *neither*, *abstain*). For each statement in a dataset, the predicted class is the class with the highest probability (as estimated by sPCA+CP). The values in each group of four subcolumns sum to 1 because they are normalized counts. For example, in the first row under the *True* ground-truth label, we see that *true* predictions have the value of 0.90; that is, 90% of all the true statements are classified as true.

Model ↓	True Labels → / Predicted → / Dataset ↓	True				False				Neither			
		True	False	Neither	Abstain	True	False	Neither	Abstain	True	False	Neither	Abstain
Bio-Medical-Llama-3-8B	City Locations	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
	Medical Indications	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
	Word Definitions	0.81	0.03	0.17	0.00	0.72	0.07	0.21	0.00	0.68	0.07	0.25	0.00
Gemma-2-9B	City Locations	0.98	0.00	0.02	0.00	0.54	0.04	0.41	0.00	0.06	0.24	0.70	0.00
	Medical Indications	0.86	0.03	0.11	0.00	0.44	0.33	0.23	0.00	0.00	0.84	0.15	0.00
	Word Definitions	0.96	0.01	0.03	0.00	0.88	0.06	0.07	0.00	0.71	0.16	0.13	0.00
Gemma-2-9B-it	City Locations	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
	Medical Indications	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
	Word Definitions	0.95	0.00	0.05	0.00	0.61	0.09	0.31	0.00	0.55	0.15	0.30	0.00
Gemma-7B	City Locations	1.00	0.00	0.00	0.00	0.97	0.01	0.02	0.00	1.00	0.00	0.00	0.00
	Medical Indications	0.27	0.01	0.72	0.00	0.18	0.03	0.79	0.00	0.40	0.05	0.55	0.00
	Word Definitions	0.82	0.00	0.17	0.00	0.53	0.03	0.44	0.00	0.53	0.03	0.44	0.00
Gemma-7B-it	City Locations	0.90	0.01	0.09	0.00	0.02	0.82	0.16	0.00	0.11	0.51	0.37	0.00
	Medical Indications	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
	Word Definitions	0.88	0.01	0.11	0.00	0.37	0.15	0.48	0.00	0.52	0.14	0.35	0.00
Llama-3-8B	City Locations	0.96	0.03	0.02	0.00	0.42	0.49	0.09	0.00	0.54	0.40	0.06	0.00
	Medical Indications	0.82	0.10	0.08	0.00	0.12	0.78	0.10	0.00	0.44	0.45	0.11	0.00
	Word Definitions	0.95	0.01	0.04	0.00	0.86	0.02	0.12	0.00	0.78	0.06	0.16	0.00
Llama-3.1-8B-Instruct	City Locations	0.91	0.00	0.09	0.00	0.00	0.76	0.24	0.00	0.05	0.35	0.59	0.00
	Medical Indications	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
	Word Definitions	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
Llama-3.2-3B	City Locations	0.91	0.04	0.05	0.00	0.05	0.87	0.08	0.00	0.24	0.53	0.23	0.00
	Medical Indications	0.75	0.05	0.20	0.00	0.12	0.41	0.47	0.00	0.18	0.48	0.34	0.00
	Word Definitions	0.75	0.00	0.24	0.00	0.43	0.02	0.55	0.00	0.41	0.03	0.56	0.00
Llama-3.2-3B-Instruct	City Locations	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
	Medical Indications	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
	Word Definitions	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
Llama3-Med42-8B	City Locations	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
	Medical Indications	0.94	0.01	0.05	0.00	0.67	0.17	0.16	0.00	0.89	0.04	0.07	0.00
	Word Definitions	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
Mistral-7B-Instruct-v0.3	City Locations	0.88	0.01	0.11	0.00	0.01	0.86	0.13	0.00	0.26	0.16	0.59	0.00
	Medical Indications	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
	Word Definitions	0.84	0.01	0.15	0.00	0.76	0.03	0.21	0.00	0.79	0.02	0.19	0.00
Mistral-7B-v0.3	City Locations	1.00	0.00	0.00	0.00	0.98	0.00	0.02	0.00	1.00	0.00	0.00	0.00
	Medical Indications	0.80	0.06	0.15	0.00	0.11	0.66	0.23	0.00	0.24	0.49	0.27	0.00
	Word Definitions	0.87	0.02	0.10	0.00	0.49	0.19	0.32	0.00	0.42	0.30	0.28	0.00
Qwen-2.5-14B	City Locations	0.98	0.00	0.02	0.00	0.45	0.32	0.24	0.00	0.53	0.37	0.10	0.00
	Medical Indications	0.83	0.06	0.11	0.00	0.26	0.37	0.37	0.00	0.17	0.49	0.34	0.00
	Word Definitions	0.96	0.00	0.03	0.00	0.87	0.03	0.10	0.00	0.89	0.02	0.09	0.00
Qwen-2.5-14B-Instruct	City Locations	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
	Medical Indications	0.84	0.08	0.09	0.00	0.10	0.82	0.09	0.00	0.15	0.71	0.14	0.00
	Word Definitions	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
Qwen-2.5-7B	City Locations	0.93	0.01	0.06	0.00	0.19	0.54	0.28	0.00	0.22	0.51	0.27	0.00
	Medical Indications	0.36	0.05	0.59	0.00	0.32	0.08	0.59	0.00	0.23	0.17	0.60	0.00
	Word Definitions	0.97	0.00	0.03	0.00	0.92	0.01	0.06	0.00	0.97	0.01	0.03	0.00
Qwen-2.5-7B-Instruct	City Locations	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
	Medical Indications	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
	Word Definitions	0.85	0.09	0.06	0.00	0.16	0.77	0.07	0.00	0.36	0.53	0.11	0.00

Table A31: **Row-wise confusion matrices for supervised PCA probe with conformal prediction (sPCA+CP) across all ⟨model-dataset pairs⟩** (evaluated on the *bag*). Each row corresponds to a specific model and a dataset. Columns are grouped by the ground-truth labels (*True*, *False*, *Neither*) with groups of subcolumns that specify the distribution of predictions (*true*, *false*, *neither*, *abstain*). For each statement in a dataset, the predicted class is the class with the highest probability (as estimated by sPCA+CP). The values in each group of four subcolumns sum to 1 because they are normalized counts. For example, in the first row under the *True* ground-truth label, we see that *true* predictions have the value of 1.00; that is, 100% of all the true statements are classified as true.

Model ↓	True Labels → Predicted → Dataset ↓	True				False				Neither			
		True	False	Neither	Abstain	True	False	Neither	Abstain	True	False	Neither	Abstain
Bio-Medical-Llama-3-8B	City Locations	0.80	0.01	0.00	0.18	0.00	0.88	0.00	0.11	0.00	0.00	1.00	0.00
	Medical Indications	0.77	0.10	0.00	0.12	0.10	0.81	0.01	0.08	0.00	0.00	1.00	0.00
	Word Definitions	0.86	0.09	0.01	0.03	0.11	0.87	0.01	0.02	0.00	0.01	0.98	0.01
Gemma-2-9B	City Locations	0.87	0.00	0.00	0.13	0.00	0.85	0.00	0.15	0.00	0.00	1.00	0.00
	Medical Indications	0.81	0.08	0.01	0.11	0.10	0.73	0.01	0.16	0.01	0.00	0.97	0.02
	Word Definitions	0.86	0.10	0.00	0.04	0.11	0.85	0.01	0.03	0.01	0.01	0.98	0.00
Gemma-2-9B-it	City Locations	0.85	0.02	0.02	0.12	0.02	0.87	0.00	0.11	0.00	0.00	0.98	0.02
	Medical Indications	0.76	0.10	0.01	0.13	0.07	0.84	0.01	0.09	0.01	0.00	0.99	0.01
	Word Definitions	0.83	0.08	0.00	0.08	0.07	0.88	0.01	0.05	0.00	0.01	0.98	0.01
Gemma-7B	City Locations	0.86	0.01	0.00	0.13	0.01	0.89	0.00	0.11	0.00	0.00	1.00	0.00
	Medical Indications	0.75	0.09	0.00	0.17	0.08	0.73	0.01	0.18	0.01	0.01	0.95	0.04
	Word Definitions	0.82	0.11	0.04	0.02	0.15	0.81	0.01	0.03	0.02	0.01	0.95	0.01
Gemma-7B-it	City Locations	0.86	0.02	0.01	0.11	0.02	0.87	0.00	0.11	0.00	0.00	0.99	0.01
	Medical Indications	0.58	0.05	0.01	0.35	0.15	0.50	0.00	0.35	0.00	0.01	0.94	0.05
	Word Definitions	0.69	0.16	0.04	0.11	0.12	0.77	0.03	0.07	0.01	0.01	0.96	0.01
Llama-3-8B	City Locations	0.87	0.01	0.00	0.12	0.01	0.85	0.00	0.14	0.00	0.00	1.00	0.00
	Medical Indications	0.78	0.09	0.01	0.12	0.11	0.77	0.01	0.11	0.00	0.01	0.98	0.01
	Word Definitions	0.87	0.12	0.00	0.00	0.13	0.85	0.01	0.00	0.01	0.01	0.97	0.00
Llama-3.1-8B-Instruct	City Locations	0.87	0.02	0.00	0.11	0.02	0.86	0.00	0.13	0.00	0.00	1.00	0.00
	Medical Indications	0.80	0.12	0.00	0.09	0.09	0.85	0.00	0.06	0.00	0.00	1.00	0.00
	Word Definitions	0.87	0.06	0.00	0.07	0.07	0.86	0.00	0.06	0.00	0.01	0.98	0.01
Llama-3.2-3B	City Locations	0.84	0.03	0.00	0.12	0.04	0.86	0.01	0.09	0.00	0.01	0.99	0.01
	Medical Indications	0.71	0.10	0.00	0.19	0.09	0.70	0.01	0.20	0.00	0.00	0.98	0.02
	Word Definitions	0.74	0.14	0.02	0.10	0.11	0.77	0.02	0.10	0.01	0.01	0.96	0.02
Llama-3.2-3B-Instruct	City Locations	0.86	0.02	0.00	0.11	0.02	0.82	0.00	0.16	0.00	0.00	1.00	0.00
	Medical Indications	0.67	0.12	0.00	0.21	0.11	0.72	0.00	0.18	0.00	0.00	0.99	0.01
	Word Definitions	0.86	0.11	0.03	0.00	0.14	0.85	0.01	0.00	0.01	0.01	0.98	0.00
Llama3-Med42-8B	City Locations	0.87	0.01	0.00	0.12	0.01	0.84	0.00	0.15	0.00	0.00	0.99	0.01
	Medical Indications	0.81	0.10	0.00	0.10	0.11	0.82	0.00	0.07	0.00	0.00	1.00	0.00
	Word Definitions	0.86	0.06	0.01	0.06	0.08	0.85	0.00	0.07	0.01	0.01	0.97	0.02
Mistral-7B-Instruct-v0.3	City Locations	0.86	0.01	0.00	0.13	0.01	0.79	0.00	0.19	0.00	0.00	1.00	0.00
	Medical Indications	0.77	0.09	0.00	0.14	0.10	0.76	0.00	0.14	0.00	0.01	0.99	0.01
	Word Definitions	0.84	0.08	0.03	0.05	0.09	0.87	0.01	0.03	0.01	0.01	0.98	0.01
Mistral-7B-v0.3	City Locations	0.86	0.01	0.00	0.13	0.01	0.86	0.00	0.13	0.00	0.00	1.00	0.00
	Medical Indications	0.75	0.08	0.00	0.17	0.09	0.76	0.00	0.15	0.00	0.00	0.99	0.01
	Word Definitions	0.87	0.11	0.02	0.00	0.14	0.85	0.01	0.00	0.01	0.01	0.98	0.00
Qwen-2.5-14B	City Locations	0.87	0.01	0.00	0.12	0.01	0.86	0.00	0.13	0.00	0.00	0.99	0.01
	Medical Indications	0.78	0.12	0.00	0.10	0.11	0.76	0.00	0.13	0.00	0.00	0.99	0.00
	Word Definitions	0.86	0.12	0.02	0.00	0.13	0.86	0.01	0.00	0.01	0.02	0.98	0.00
Qwen-2.5-14B-Instruct	City Locations	0.82	0.01	0.00	0.17	0.01	0.87	0.00	0.11	0.00	0.00	1.00	0.00
	Medical Indications	0.80	0.12	0.01	0.07	0.07	0.86	0.01	0.07	0.00	0.00	0.99	0.00
	Word Definitions	0.86	0.07	0.01	0.05	0.06	0.86	0.00	0.08	0.00	0.01	0.98	0.01
Qwen-2.5-7B	City Locations	0.84	0.01	0.00	0.16	0.01	0.84	0.00	0.14	0.00	0.00	1.00	0.00
	Medical Indications	0.74	0.10	0.01	0.15	0.10	0.77	0.01	0.13	0.00	0.00	0.98	0.01
	Word Definitions	0.83	0.12	0.02	0.03	0.10	0.87	0.01	0.03	0.01	0.01	0.97	0.01
Qwen-2.5-7B-Instruct	City Locations	0.82	0.02	0.01	0.15	0.02	0.85	0.00	0.13	0.00	0.00	0.99	0.01
	Medical Indications	0.78	0.08	0.01	0.13	0.14	0.72	0.02	0.12	0.01	0.01	0.98	0.01
	Word Definitions	0.83	0.13	0.02	0.02	0.09	0.87	0.02	0.02	0.01	0.01	0.97	0.01

Table A32: **Row-wise confusion matrices for the multiclass `sAwMIL` across all ⟨model-dataset pairs⟩** (evaluated on the *bag*). Each row corresponds to a specific model and a dataset. Columns are grouped by the ground-truth labels (*True*, *False*, *Neither*) with groups of subcolumns that specify the distribution of predictions (*true*, *false*, *neither*, *abstain*). For each statement in a dataset, the predicted class is the class with the highest probability (as estimated by multiclass `sAwMIL`). The values in each group of four subcolumns sum to 1 because they are normalized counts. For example, in the first row under the *True* ground-truth label, we see that *true* predictions have the value of 0.80; that is, 80% of all the true statements are classified as true.

M Related Resources

Code

The code associated with this manuscript is publicly available on GitHub at carlomarxdk/trilemma-of-truth (release version 1.0.0).

Data

The data used in this study are publicly available on Hugging Face at carlomarxdk/trilemma-of-truth (DOI: 10.57967/hf/5900).