

Probing for Knowledge Attribution in Large Language Models

Ivo Brink^{✉, 📧}, Alexander Boer[📧], Dennis Ulmer[✉]

[📧] KPMG NL, [✉] University of Amsterdam

ivobrink00@gmail.com, boer.alexander@kpmg.nl, dennis.ulmer@mailbox.org

Abstract

Large language model (LLM) hallucinations, meaning fluent but factually incorrect generations, fall into two types: faithfulness violations, where the model misuses provided context, and factuality violations, where answers reflect errors in internal knowledge. Proper mitigation depends on knowing which source drives each answer. We study *contributive attribution*, i.e. the classification of the dominant knowledge source behind each output, and show that a simple linear probe trained on hidden representations can reliably identify it. We introduce ATTRIWIKI, a self-supervised pipeline that automatically generates labeled training data by prompting models to recall withheld entities from memory or read them from context without relying on knowledge conflicts. Probes trained on ATTRIWIKI achieve up to 0.96 Macro- F_1 on Llama-3.1-8B, Mistral-7B, and Qwen2.5-7B, transfer to SQuAD and WebQuestions with 0.92–0.99 accuracy, and generalize zero-shot to Tighidet et al. (2024)’s benchmark, outperforming their probe on conflicting settings without retraining. Furthermore, attribution mismatches raise error rates by up to 70%, though correct attribution does not guarantee correct answers, pointing to the need for broader detection frameworks.

🔗 [IvoBrink/AttriWiki](https://github.com/IvoBrink/AttriWiki)

1 Introduction

Picture yourself rushing to book a last-minute flight for a funeral. You ask an airline’s chatbot whether you can claim a bereavement fare, and the system confidently replies, “Yes, you have 90 days.” The rule, however, never existed, and months later, the refund is denied. This example stems from a real case in which Air Canada was held liable for its chatbot’s misinformation (Cecco, 2024; Lifshitz and Hung, 2024). A single inaccurate response can now impose serious financial, legal, or practical consequences.

Although LLM hallucinations, untrue or ungrounded outputs (Huang et al., 2025), have become a central research focus, existing detection techniques often miss a more fundamental issue: users lack transparency over where a model’s answers come from. The airline could have benefited from knowing the chatbot’s true provenance, as attribution would have enabled verification against official policy. Prior approaches, including LLM-as-a-judge (Manakul et al., 2023; Ravi et al., 2024; Friel and Sanyal, 2023) and uncertainty measures (Tomani et al., 2024; Shelmanov et al., 2025), struggle because models can be confidently wrong (Kuhn et al., 2023; Simhi et al., 2025) and because “hallucination” lacks a consistent operational definition (Qi et al., 2024).

To address this gap, we turn to *contributive attribution* (Worledge et al., 2024), which distinguishes between contextual knowledge (from the prompt or retrieved evidence) and parametric knowledge (stored in the model’s weights). This framing complements the well-known divide between faithfulness and factuality errors (Huang et al., 2025; Qi et al., 2024; Ye et al., 2024) within hallucination taxonomies, but shifts the focus: instead of asking whether the output is true, we ask whether the model relied on the source the user intended. Attribution signals can help users detect when a model defaults to conflicting parametric knowledge or, conversely, when it overly relies on irrelevant context. This is particularly valuable in retrieval-augmented generation systems (Lewis et al., 2020; Guu et al., 2020; Borgeaud et al., 2022; Izacard et al., 2023), where errors often arise not from bad retrieval but from the model ignoring retrieved information (Petroni et al., 2020; Li et al., 2023).

Our contributions are as follows: Using the ATTRIWIKI pipeline involving Wikipedia passages, entity selection, and controlled prompt variations, we create a dataset in which each completion has a clear, verifiable knowledge source (parametric or

contextual). During generation, we record hidden states at key token positions, producing attribution features without manual labels or conflicting sources. We then train lightweight probing classifiers on these representations to distinguish contextual from parametric retrieval and test their generalization on external QA datasets such as WEBQUESTIONS and SQUAD (Berant et al., 2013; Rajpurkar et al., 2016) and compare to existing attribution probing work (Tighidet et al., 2024). We show that contributive attribution is linearly decodable from LLM hidden states, particularly from the middle to upper layers. Attribution mismatches, in which a model answers from the wrong knowledge source, significantly increase error rates, especially in misleading contexts. Yet, correct attribution alone does not guarantee factual correctness. ATTRIWIKI and all code are publicly available.¹

2 Related Work

Knowledge in LLMs. Understanding hallucinations ultimately requires understanding how large language models acquire, store, and retrieve knowledge. Factual knowledge in LLMs is parametric and emergent: no single parameter encodes a fact, but distributed activation patterns collectively give rise to factual recall (Dai et al., 2022; Yao et al., 2024; Wang and Xu, 2025). Early probing work, such as LAMA (Petroni et al., 2019), showed that factual recall in language models is brittle, varying substantially across languages, paraphrases, and prompt formulations (Kassner et al., 2021; Elazar et al., 2021b; Jiang et al., 2020). In addition, recent works formalize knowledge in models by mapping philosophical notions of knowledge onto model behavior: Under the “true-belief” framework (Fierro et al., 2024) for instance, a fact is known by a model if it is reproduced consistently, which we use to make a practical distinction between parametric knowledge and context-derived information.

Attribution and Knowledge Conflicts. Once parametric and contextual knowledge are distinguished, a central challenge is detecting their usage. Knowledge conflicts arise when parametric beliefs contradict retrieved or provided context, a situation common in retrieval-augmented settings (Petroni et al., 2020; Li et al., 2023). Attribution methods aim to trace generated content back to its source. Retrieval-augmented generation approaches (Lewis

et al., 2020; Izacard et al., 2023) expose evidence to the model, but cannot guarantee its use, and models frequently hallucinate citations even when retrieval is correct (Zuccon et al., 2023). Model-based attribution methods target contributive attribution through architectural modifications that encode source identifiers (Khalifa et al., 2024), probing approaches that infer whether outputs derive from memory or context (Tighidet et al., 2024), or component-level analyses that localize the attention heads responsible for each retrieval mode (Kahardipraja et al., 2025). The first requires substantial architectural changes, while the latter two both depend on counterfactual prompts in which context contradicts parametric knowledge, and therefore conflate provenance with conflict resolution. Recent mechanistic work has begun to clarify how conflicting knowledge signals are processed. Zhao et al. (2025) trace entity-level knowledge flows through the model, showing that parametric and contextual knowledge are routed through largely distinct attention circuits and coexist as superposed signals rather than competing via direct suppression. Their entity-conditioned probes analyze how competing signals evolve across a forward pass, making their setup better suited to study conflict resolution dynamics rather than genuine source detection in the absence of conflict. Nonetheless, this perspective provides a mechanistic foundation for attribution: because both knowledge sources are routed through largely distinct circuits and persist as superposed signals, their relative contribution might remain detectable even without an explicit conflict.

Probing LLM Representations. Probing hidden states offers a lightweight approach for studying knowledge usage directly from model internals. This line of work builds on a long tradition of analyzing how linguistic and factual information is encoded across transformer layers (Conneau et al., 2018; Tenney et al., 2019; Hewitt and Manning, 2019). Prior studies show that layers specialize in surface, syntactic, and semantic features (Jawahar et al., 2019), that factual associations emerge in neuron-level key-value structures (Geva et al., 2021), and that hidden states reflect latent capabilities such as lying (Azaria and Mitchell, 2023), causal reasoning (Rohekar et al., 2024), and instruction following (Heo et al., 2025). These findings suggest that signals relevant to attribution and provenance are also present in model activations.

¹The repository is available under <https://github.com/IvoBrink/AttriWiki>

Probing for factual knowledge has already been explored by a series of works (see Youssef et al., 2023 for an overview), but recent work applies probing specifically to hallucination detection. MIND (Su et al., 2024) trains classifiers on final-layer activations, outperforming SelfCheckGPT (Manakul et al., 2023) and GPT-4o critics on several benchmarks. However, it provides limited interpretability regarding why a response is classified as hallucinated.

Overall, prior work offers strong taxonomies and attribution techniques, but a key gap remains: existing methods lack scalable, controlled datasets that cleanly isolate parametric versus contextual knowledge use. This motivates approaches that explicitly construct such contrasts and probe hidden states to reveal the true origin of model outputs.

3 A Self-Supervised Data Pipeline

3.1 Motivation

Although recent work examines the use of parametric-contextual knowledge, existing datasets do not isolate retrieval modes. Many studies expose both sources simultaneously through conflict settings and infer attribution *post hoc* from the model’s answer (Tighidet et al., 2024; Xu et al., 2024; Zhao et al., 2025), conflating provenance with downstream decision-making and conflict resolution.

Therefore, checking whether a fact is present in a model’s parametric memory is necessary to isolate contextual vs. parametric examples. Prior work verifies parametric knowledge via data recency (Zhang et al., 2024) or by querying the model (Cheng et al., 2024). However, these approaches neither explicitly detect attribution nor prevent the simultaneous availability of parametric and contextual knowledge when analyzing their interactions.

To this end, we introduce ATTRIWIKI, a self-supervised data pipeline that generates paired examples which *force* models to draw information either from context or from stored knowledge, established through extensive knowledge testing. By ensuring that only a single knowledge source is available at generation time, ATTRIWIKI enables isolated attribution analysis without manual labeling.

3.2 Data and Models

Data. Wikipedia is part of many modern pre-training corpora (Jiang et al., 2023; Dubey et al., 2024; Qwen Team et al., 2025). We use its dense

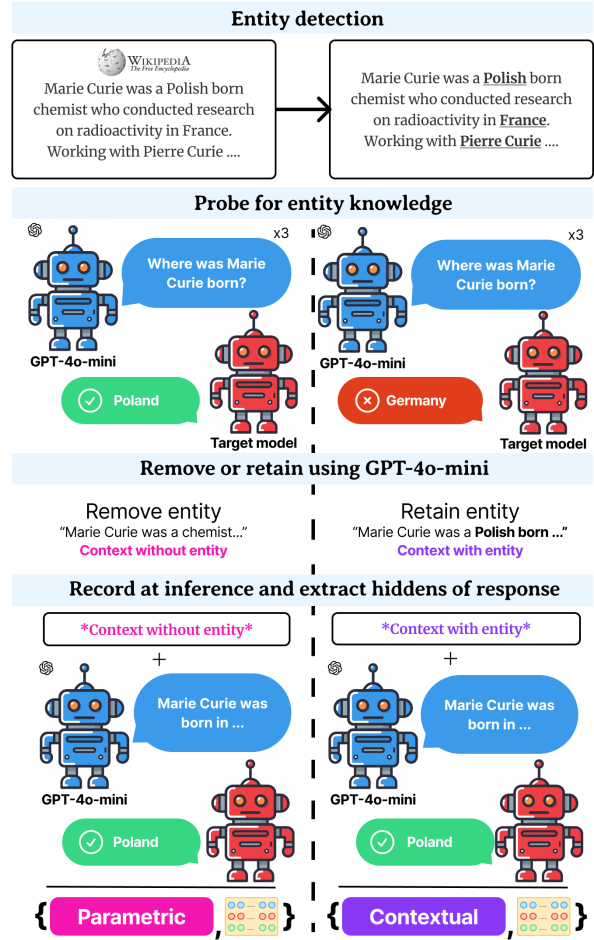


Figure 1: Overview of the data generation pipeline, showing entity selection, knowledge testing, prompt construction, and hidden-state extraction.

factual text to test both retrieval modes naturally. Our knowledge-testing procedure (Section 3.3) supports the claim that mid-sized open language models (approximately 7–8B parameters) only partially memorize Wikipedia facts: roughly half of sampled entities are recalled through our knowledge tests and the rest are not (Table 1), indicating that neither retrieval mode is degenerate for any target model. We sample 20k pages using the MediaWiki API,² discarding snippets shorter than 300 characters.

Data Generation Model. We use GPT-4O-MINI (gpt-4o-mini; OpenAI, 2024) as an instruction-following model to paraphrase passages, remove or retain entities, and rewrite completion sentences during dataset construction (see next section).

Target Models. Attribution probing is performed on open-weight language models from which we

²See <https://en.wikipedia.org/w/api.php> (last accessed May 20, 2026). The Wikipedia text is licensed under CC BY-SA 3.0/4.0, and all use complies with the terms of that license.

extract hidden states during generation. We use LLAMA-3.1-8B (Dubey et al., 2024), MISTRAL-7B (Jiang et al., 2023), and QWEN2.5-7B (Qwen Team et al., 2025) as target models.

3.3 Method

An overview of the data generation pipeline is shown in Figure 1. Because attribution concerns the source of specific factual content, we focus on *named entities* as attribution anchors, identified using spaCy’s transformer-based NER (Honnibal et al., 2020). Entities that overlap, are synonymous, or closely resemble the passage title are found using a RoBERTa (Liu et al., 2019) cross-encoder (CE) similarity model.³ The CE assigns pairwise semantic similarity scores, and entities above a threshold of 0.6 are considered synonyms (we refer to this procedure as *synonym matching*).⁴ In this case, synonyms are excluded. From the remaining entity candidates, GPT-4O-MINI selects up to three representative, non-numeric, non-temporal entities per passage (see Appendix D.1.1).

Each entity–passage pair undergoes *knowledge testing* to determine whether the target model can recall the entity without the passage, serving as an *operational proxy* for parametric memory. We use three prompt formats to test entity knowledge (see Appendix D.1.2). If any test elicits the correct entity (exact or *synonym match*), it is labeled *known*; otherwise, *unknown*. This high-recall labeling avoids false unknown labels, preferring discarding an example over admitting one that could draw on both sources.

We then construct *contextual* and *parametric* variants for each passage. For known entities, all mentions are removed using GPT-4O-MINI (Appendix D.1.3), thereby forcing parametric retrieval; for unknown entities, the name remains visible, while a different entity is removed to prevent the lexical bias this may introduce. GPT-4O-MINI appends a short, natural completion cue such as “The [description] is called . . .,” yielding paired prompts that differ only in the presence of the entity (see Appendix D.1.4).

³cross-encoder/stsb-roberta-base.

⁴The threshold of 0.6 is anchored to the cross-encoder’s training scale: it is trained on STS-B, whose 0–5 human ratings are rescaled to [0,1] (Cer et al., 2017), so 0.6 corresponds to a rating of 3 (roughly equivalent). The value sets how strictly two strings must correspond to denote the same entity, controlling both the permissiveness of our knowledge proxy in accepting similar recalled completions and the similarity a candidate entity may bear to the passage title before exclusion.

ATTRIWIKI	Ratio (P:C)	Syn	Size
Llama-3.1-8B	≈5:3	10.3%	27,244
Mistral-7B-v0.1	≈3:2	10.7%	26,853
Qwen2.5-7B	≈1:1	12.0%	24,336

Table 1: ATTRIWIKI dataset composition for each model. **Ratio (P:C)** denotes the approximate parametric-to-contextual proportion of examples. **Syn** denotes the synonym match rate, and **Size** denotes the total number of examples.

Each prompt is submitted to the target model to produce a completion. During greedy decoding, we extract hidden representations at two key token positions: (i) the first generated token (FTG), which reflects the model’s initial response intent allowing us to compare responses without specific entities (Su et al., 2024), and (ii) the last token of the entity (LTE), capturing the fully integrated entity semantics (Tighidet et al., 2024).⁵ Entity spans are located by exact or *synonym matching*, and the corresponding hidden states are serialized for attribution classification. This automated pipeline produces large volumes of data, with each example having a verifiable knowledge source (see Appendix A for an overview of the resource consumption).

Results. Table 1 summarizes the ATTRIWIKI dataset for Llama-3.1-8B, Mistral-7B-v0.1, and Qwen2.5-7B. Llama and Mistral show parametric-dominant distributions (≈5:3 and 3:2), while Qwen is nearly balanced (≈1:1). Dataset sizes are similar across models. Synonym match rates (percentage of completions that matched semantically but not verbatim) are highest for Qwen (12.0%). Overall, the pipeline behaves consistently.

Latent Space Analysis. To gain insight into how contextual and parametric information is represented internally, we perform a per-layer 2D principal component analysis (PCA; Pearson, 1901) on decoder hidden states extracted at the first token of generation (FTG). As shown in Figure 2 (refer to Appendix C for all the layers), representations from early layers largely overlap, while a clearer separation between contextual and parametric activations emerges from the middle layers onward in all three models. This pattern suggests that attribution-related information becomes increasingly linearly accessible in higher layers, indicating that contextual and parametric signals are likely encoded in

⁵We omit first-token-entity as it coincides with first-token-generation in almost all cases.

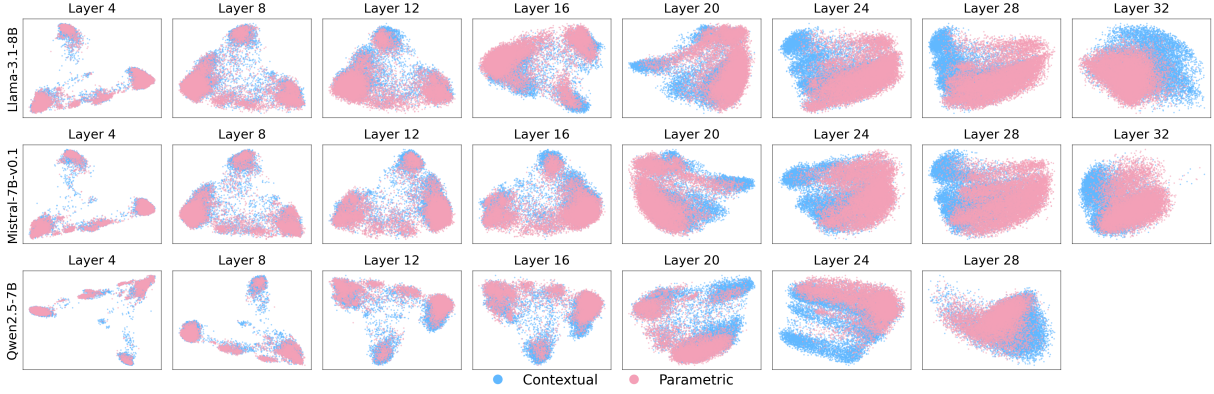


Figure 2: Per-layer PCA of decoder hidden states at the first generated token (FTG). Qwen has 28 layers (vs. 32 in Llama and Mistral). In the mid to upper layers, contextual and parametric activations increasingly diverge, indicating greater separability. Plots for all layers are given in [Appendix C.1](#).

Model	Classifier	F_1
Llama-3.1-8B	BoW	$.652 \pm .008$
	Embedding	$.663 \pm .002$
Mistral-7B-v0.1	BoW	$.660 \pm .005$
	Embedding	$.667 \pm .005$
Qwen2.5-7B	BoW	$.675 \pm .004$
	Embedding	$.648 \pm .006$

Table 2: Bias detection performance using BoW and embedding-based classifiers on ATTRIWIKI with Macro- F_1 (mean $\pm$ cross-validation standard deviation).

separable directions of the hidden state space.

3.4 Evaluating Bias

To see if ATTRIWIKI introduces lexical shortcuts between classes as an unintended effect of our knowledge-proxy,⁶ we train text-only classifiers on the passages. A balanced bag-of-words (BoW) logistic regression and a logistic regression model trained on DeBERTaV3⁷ embeddings (He et al., 2023) are evaluated under five-fold cross-validation (for specifics see [Appendix B](#)).

Results. Results are shown in [Table 2](#). Bag-of-words and embedding-based classifiers achieve F_1 scores of 0.65–0.67, indicating the presence of lexical cues, though these alone are insufficient for reliable performance. BoW on the data for all models shows similar performance and nearly identical bigram profiles: contextual examples favor

entity-specific terms (e.g., named, born in, towns), whereas parametric examples lean toward broader encyclopedic vocabulary (e.g., countries, nationalities). This suggests that our knowledge proxy introduces a mild and stable lexical bias.

4 Attribution Classification

To assess whether hidden states in ATTRIWIKI truly encode the *source* of a model’s knowledge, we train three probing classifiers of increasing complexity. Each model is trained on 64% of ATTRIWIKI, validated on 16%, and tested on the remaining 20%, then evaluated for out-of-domain generalization. Results are reported as macro- F_1 . Consult [Appendix B.2](#) for more details.

4.1 Classifiers

Since the data for Llama and Mistral shows a mild class imbalance ($\approx 3:2$ parametric–contextual), we compensate through loss weights. To prevent information leakage, we further enforce title-disjoint train–test splits, ensuring that no samples originating from the same article appear in both splits.

Final-Layer Logistic Regression (*Final-LR*).

As a minimal baseline, we train a single logistic unit on the normalized hidden vector $\mathbf{h}_L \in \mathbb{R}^H$ from the final transformer layer:

$$p = \sigma(\mathbf{w}^\top \mathbf{h}_L + b), \quad (1)$$

where $\mathbf{w} \in \mathbb{R}^H$ and $b \in \mathbb{R}$ are the classifier weights and bias, $\sigma(\cdot)$ is the sigmoid function, and training uses an ℓ_2 -regularized logistic loss with inverse-frequency class weighting.

⁶For instance, generic descriptors such as “a country” may occur more frequently in contexts associated with widely known facts, inadvertently biasing a classifier toward predicting parametric knowledge based on surface-level phrasing rather than attribution.

⁷[microsoft/deberta-v3-large](https://github.com/microsoft/deberta-v3-large).

Layer-Weighted Logistic Regression (*Layer-LR*).

To exploit depth information, we learn a set of unconstrained parameters $\theta \in \mathbb{R}^L$, which are transformed via a softmax to obtain aggregation weights $\alpha = \text{softmax}(\theta)$ over transformer layers. These weights are used to form a blended representation

$$\bar{\mathbf{h}} = \sum_{\ell=1}^L \alpha_{\ell} \mathbf{h}_{\ell}. \quad (2)$$

A logistic head then predicts the probability of parametric retrieval from $\bar{\mathbf{h}}$. All parameters $(\theta, \mathbf{w}, b)$ are trained jointly using binary cross-entropy with logits and positive-class reweighting to account for label imbalance, optimized with AdamW (Kingma and Ba, 2015). Hyperparameters and training details are reported in Appendix B.2.

Layer-Weighted MLP (*Layer-MLP*). We extend the previous model with sparsemax-normalized aggregation weights (Martins and Astudillo, 2016), encouraging focus on a small number of salient layers, motivated by the tendency of softmax-aggregated MLP probes to learn diffuse layer weights. The linear classifier is replaced with a compact two-layer MLP operating on the aggregated representation $\bar{\mathbf{h}}$:

$$p = \sigma(\mathbf{w}_2^{\top} \text{GELU}(\mathbf{W}_1 \bar{\mathbf{h}}) + b), \quad (3)$$

We use GELU (Hendrycks and Gimpel, 2016), where $\mathbf{W}_1 \in \mathbb{R}^{m \times H}$ projects the hidden representation of dimension H into a bottleneck of size m , and $\mathbf{w}_2 \in \mathbb{R}^m$ denotes the output-layer weights. Hyperparameters and training details are reported in Appendix B.2.

4.2 Out-of-distribution Datasets

All classifiers are first trained on ATTRIWIKI and then evaluated on 2,000 samples from two external QA datasets. The best classifier is additionally compared against the probe of Tighidet et al. (2024) on conflicting scenarios. All samples include a single demonstration line to enforce output format (see Appendix D.2).

PARAREL (Elazar et al., 2021a). To situate our approach relative to Tighidet et al. (2024), we evaluate on prompts derived from PARAREL, a structured dataset of factual (SUBJECT, RELATION, OBJECT) triples sourced from Wikidata. Each triple is converted into a probing prompt consisting of a counter-factual statement contradicting the model’s

Model	Final-LR	Layer-LR	Layer-MLP
<i>Token = FTG (First Token of Generation)</i>			
Llama-3.1-8B	.841 $\pm$.005	.919 $\pm$.004	.922 $\pm$.004
Mistral-7B-v0.1	.844 $\pm$.005	.922 $\pm$.004	.933 $\pm$.004
Qwen2.5-7B	.835 $\pm$.005	.912 $\pm$.004	.926 $\pm$.004
<i>Token = LTE (Last Token of Entity)</i>			
Llama-3.1-8B	.854 $\pm$.005	.948 $\pm$.003	.955 $\pm$.003
Mistral-7B-v0.1	.904 $\pm$.004	.958 $\pm$.003	.961 $\pm$.003
Qwen2.5-7B	.867 $\pm$.005	.935 $\pm$.004	.949 $\pm$.006

Table 3: Macro- F_1 (mean $\pm$ bootstrap std, $B = 1000$) on ATTRIWIKI-TEST ($n \approx 5,000$).

parametric knowledge, followed by a query over the same subject–relation pair. Labels are assigned based on whether the model’s response matches the planted counter-object (CK) or the true parametric object (PK), other answers are ignored. Relations are grouped into eight semantic categories.

WEBQUESTIONS (Berant et al., 2013). Each factoid question lacks supporting context (e.g., “Who developed general relativity?” $\rightarrow$ *Albert Einstein*); hence, answers must come from the model’s parametric memory. Items are wrapped in a one-shot question-answer prompt for compatibility with decoder-only models.

SQUAD (Rajpurkar et al., 2016). Passages contain answers verbatim, making the task predominantly *contextual*. To filter memorized facts, we query each question without its paragraph; if the model still answers correctly, the item is recorded as prior knowledge and used for a separate experiment (see below). The remaining examples form the contextual test set.

SQuAD Variants. We derive three controlled variants: (a) an *irrelevant context* split pairing each known question with an irrelevant, randomly sampled paragraph to force parametric retrieval; (b) an *answer-string decoy* version where the paragraph contains the correct answer verbatim in an unrelated context, testing reliance on surface overlap; and (c) a *dual-source* version retaining both context and prior-knowledge cases to study retrieval preference. Together, these variants test robustness against degenerate heuristics, e.g., predicting attribution solely from the presence of context.

4.3 Results

All reported macro- F_1 values are accompanied by bootstrap standard deviations estimated from $B = 1,000$ samples with replacement.

Model	SQuAD	WebQ
Llama-3.1-8B	.977 $\pm$.003	.955 $\pm$.004
Mistral-7B-v0.1	.918 $\pm$.005	.996 $\pm$.001
Qwen2.5-7B	.996 $\pm$.005	.991 $\pm$.002

Table 4: Generalization accuracy (mean $\pm$ bootstrap std, $B = 1000$) of Layer-LR (FTG) on out-of-domain QA datasets. WebQ = WebQuestions.

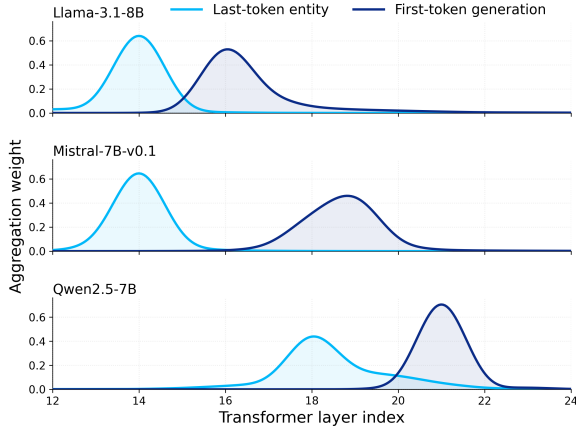


Figure 3: Layer-aggregation weights learned by *Layer-LR* for first-token generation and last-token entity representations. Curves show learned weights across transformer layers, smoothed with a Gaussian kernel for visual clarity. The x-axis is restricted to layers 10–24, as earlier and later layers receive negligible weight.

Attribution Performance. Table 3 reports attribution classification on the held-out ATTRIWIKI split. We compare *Final-LR*, *Layer-LR*, and *Layer-MLP*. Layer aggregation improves F_1 by 7–11 pp over the final-layer baseline, while the MLP yields only marginal gains (≈ 1 pp). Token choice matters: the last token of the entity (*LTE*) performs best (0.95–0.96), while the first token of generation (*FTG*) remains slightly lower but above 0.90. Mistral consistently outperforms Llama and Qwen by one point, though trends are consistent across models, with Macro- F_1 closely matching accuracy.

Generalization. To assess out-of-distribution generalization, we evaluate the classifier on two out-of-domain QA benchmarks: SQuAD (contextual) and WEBQUESTIONS (parametric). As shown in Table 4, it generalizes near perfectly without fine-tuning, achieving 0.92–0.99 accuracy across Llama, Mistral, and Qwen on both benchmarks. Similar results are observed for the Layer-MLP in the appendix, in Table 12. We report only on first-token generation (FTG), because it allows us to analyze incorrect responses in later sections.

Model	Probe	Macro- F_1
Mistral-7B-v0.1	Tighidet et al. (L10)	.745 $\pm$.014
	ATTRIWIKI	.806 $\pm$.012
Llama-3.1-8B	Tighidet et al. (L11)	.640 $\pm$.019
	ATTRIWIKI	.771 $\pm$.016
Qwen2.5-7B	Tighidet et al. (L22)	.757 $\pm$.007
	ATTRIWIKI	.793 $\pm$.006

Table 5: Comparison of Tighidet et al. and ATTRIWIKI probe (Layer-LR on FTG) across models with Macro- F_1 (mean $\pm$ bootstrap std, $B = 1000$). Layer number in parentheses indicates the best-performing layer.

Which Layers Does the Classifier Rely on?

As shown in Figure 3, the classifier concentrates weight in the lower-middle transformer layers across models when using the last token of the entity. For first-token generation, peak contributions occur at layers 16 (Llama-3.1-8B), 19 (Mistral-7B-v0.1), and 21 (Qwen2.5-7B), while last-token entity representations peak earlier at layers 14 (Llama, Mistral) and 18 (Qwen). Notably, Qwen assigns weight to relatively deeper layers given its shallower overall depth (28 versus 32). This indicates that using the last token of the entity might make the probe focused on entity-specific, lexical information, while the first-token generation hidden states push the probe to focus on potentially high-level information processed in later layers.

Comparison to Tighidet et al. (2024). Table 5 compares our ATTRIWIKI-trained probe Layer-LR on FTG against the best-performing layer of Tighidet et al. (2024) evaluated on PARAREL under the same exact scheme. The ATTRIWIKI probe consistently outperforms across all three models on the never-seen before data, with gains ranging from 3 pp (Qwen2.5-7B) to 13 pp (Llama-3.1-8B). Unlike Tighidet et al. (2024), our probe requires no conflicting context during training, extracts hidden states at the first generated token rather than the last input token, and learns to aggregate across all layers jointly rather than selecting a single layer post-hoc. We include the per-category results in Appendix C.

4.4 Ablation

Ablations are performed on Layer-LR at FTG, as it is robust and does not rely on entity mentions.

Degenerate Attribution. We evaluate whether attribution probes rely on degenerate heuristics rather than genuine provenance signals. Firstly,

we test the answer-string decoy variant of SQuAD, where the gold answer appears verbatim in an unrelated context—performance drops for all classifiers. Nevertheless, *Layer-LR* remains well above chance (0.724)⁸. In contrast, the MLP collapses to near-random performance (0.541), indicating that the MLP overfits to lexical shortcuts such as entity repetition, whereas *Layer-LR* captures a more robust attribution signal (see Table 11 in Appendix C). Secondly, in the irrelevant-context setting—where known facts are paired with unrelated passages—92% of answers are still attributed as parametric, ruling out a trivial heuristic that equates the mere presence of context with contextual attribution. Thirdly, when both parametric memory and context suffice to answer a question, models overwhelmingly default to the contextual channel (87.1%). This rules out the explanation that the probe merely detects whether a fact is stored in the model’s parameters, rather than the knowledge source.

Error and Attribution Mismatch. To analyze the relationship between attribution and answer correctness, we consider two complementary conditions: (i) present parametric knowledge paired with irrelevant context, and (ii) absent parametric knowledge paired with relevant context. We use SQuAD with random contexts in case of parametric knowledge and original context otherwise, recording answer correctness. Source alignment (parametric when required, or contextual when required) and correctness form a 2×2 contingency table, analyzed using Fisher’s exact test (Fisher, 2018), with relative risk as the effect size. Attribution mismatches strongly correlate with errors ($p \ll 0.001$), with asymmetric effects: relying on misleading context increases errors by up to $\approx 70\%$ when parametric knowledge is required, whereas defaulting to parametric memory in contextual settings increases errors by only $\approx 30\%$.

5 Discussion

Our findings indicate that *contributive attribution*, whether a completion is driven by contextual evidence or parametric memory, is a representation-level property that is accessible from hidden states, offering a diagnostic lens on knowledge channel usage that complements hallucination taxonomies distinguishing faithfulness from factuality errors (Huang et al., 2025; Qi et al., 2024; Ye et al., 2024).

⁸Semantic irrelevance is not explicitly verified; we use this ablation to uncover potential model shortcuts.

Our mismatch analysis motivates this: using the wrong knowledge source substantially increases error risk, particularly when misleading context overrides parametric knowledge. Yet many errors persist despite correct source alignment, indicating that attribution is one diagnostic axis among others rather than a complete error detector.

Moreover, attribution is learnable even without explicit knowledge conflicts, as evidenced by our strong out-of-distribution performance including on unseen conflict settings. This challenges conflict-based approaches (Tighidet et al., 2024; Xu et al., 2024), which conflate attribution with conflict resolution by design. Where prior mechanistic work shows that contextual and parametric signals coexist as superposed activations routed through largely distinct circuits (Zhao et al., 2025), our results go further: their relative contribution is quantitatively detectable at generation time without ever exposing contradictory information, suggesting conflict-free training is not merely sufficient but actually preferable.

Two further results confirm prior work. The limited gains from an MLP head and its degradation under the answer-string decoy suggest that added model complexity encourages shortcut learning, while linear probes more reliably predict provenance—consistent with the knowledge source being another property encoded linearly in hidden representations (Park et al., 2024; Jiang et al., 2024; Merullo et al., 2025). The preference of the probe for upper-middle layers aligns with prior findings that mid-layer activations are most informative for knowledge attribution (Tighidet et al., 2024), and with broader evidence that later layers encode decision-relevant abstractions beyond surface form (Jawahar et al., 2019; Tenney et al., 2019).

6 Conclusions

We study contributive attribution in large language models using a self-supervised, conflict-free pipeline (ATTRIWIKI). Linear probes trained on ATTRIWIKI achieve up to 0.96 F_1 on Llama, Mistral, and Qwen models and generalize strongly to SQuAD, WebQuestions and unseen conflict settings. Attribution mismatches increase error risk by 30–70%, particularly under misleading contexts, although correct attribution alone does not guarantee correctness. Together, these findings establish attribution as a meaningful diagnostic signal.

Future work should explore how attribution sig-

nals can be integrated into downstream systems. In retrieval-augmented generation (RAG; [Guu et al., 2020](#); [Borgeaud et al., 2022](#)), attribution probes could provide online signals to detect ignored evidence and trigger selective verification or re-retrieval. In LLM-powered chatbots, exposing attribution may also elicit more critical user engagement by clarifying whether responses rely on retrieved evidence or prior knowledge, similar to [Deng et al. \(2023\)](#). Such applications call for human evaluation, both to validate attribution judgments independently of our automated proxy and to establish whether users find the signal actionable in practice. Other works have identified potential interventions through steering model activations along a knowledge-source direction ([Basu et al., 2025](#); [Bi et al., 2025](#)).

Limitations

Data. ATTRIWIKI’s controlled Wikipedia domain may limit generalization; testing attribution signals on legal text, dialogue, and noisy web data is essential. Extension to multilingual settings also poses challenges to ATTRIWIKI since knowledge varies across languages ([Kassner et al., 2021](#)). Additionally, shortcut learning poses a risk, inputs might be lexically associated with parametric knowledge as demonstrated in [Section 3.4](#). In future work, such shortcuts could be mitigated at the dataset level through adversarial filtering, which iteratively removes examples solved by a lexical-only classifier, such as our BoW baseline ([Bras et al., 2020](#)). Furthermore, distribution shift remains a concern: PARAREL covers Wikidata subjects not present in ATTRIWIKI, and per-category results in [Appendix C](#) show that probe performance varies greatly across categories, suggesting sensitivity to unseen domains.

Attribution Classification. The classifier would benefit from moving beyond token-level attribution to sequence-level approaches. While our current method performs well on factual questions, it struggles in more natural settings where no clear entity span exists. Aggregation strategies, such as attention-based span scoring, could provide more accurate attribution in these cases. Another limitation is model dependence: each LLM requires a new classifier, and retraining the model often necessitates retraining the classifier and regenerating data if the model’s knowledge evolves.

Ablation. Although informative, error analysis remains challenging. Phenomena such as contradictions, exaggerations, and refusal to answer are often grouped under hallucinations, yet our focus here is on forced factual retrieval. Assessing answer correctness is equally difficult. Minor variations, such as one-letter acronym slips (e.g., GDP → GPD), can bypass the cross-encoder (*synonym matching*), while valid polarity with low lexical overlap may also lead to misclassifications. For example, the answer “smaller” correctly responds to “Were the houses bigger or smaller?” but fails against the gold label “smaller houses.”

Ethics Considerations

This work only uses publicly available datasets and does not involve human subjects or personal data. One API-based LLM is used solely for data generation. We do not identify significant ethical risks beyond those common to interpretability research.

Acknowledgments

This work is supported by the Dutch National Science Foundation (NWO Vici VI.C.212.053) and KPMG NL. We would also like to express our gratitude to Ivan Titov for his valuable guidance and feedback throughout this work.

References

- Amos Azaria and Tom M. Mitchell. 2023. [The internal state of an LLM knows when it’s lying](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 967–976. Association for Computational Linguistics.
- Samyadeep Basu, Vlad Morariu, Zichao Wang, Ryan Rossi, Cherry Zhao, Soheil Feizi, and Varun Manjunatha. 2025. On mechanistic circuits for extractive question-answering. *arXiv preprint arXiv:2502.08059*.
- Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. [Semantic parsing on Freebase from question-answer pairs](#). In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1533–1544, Seattle, Washington, USA. Association for Computational Linguistics.
- Baolong Bi, Shenghua Liu, Yiwei Wang, Yilong Xu, Junfeng Fang, Lingrui Mei, and Xueqi Cheng. 2025. Parameters vs. context: Fine-grained control of knowledge reliance in language models. *arXiv preprint arXiv:2503.15888*.

- Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego De Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, and 9 others. 2022. [Improving language models by retrieving from trillions of tokens](#). In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 2206–2240. PMLR.
- Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew E. Peters, Ashish Sabharwal, and Yejin Choi. 2020. [Adversarial filters of dataset biases](#). In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 1078–1088. PMLR.
- Leyland Cecco. 2024. [Air canada ordered to pay customer who was misled by airline’s chatbot](https://www.theguardian.com/world/2024/feb/16/air-canada-chatbot-lawsuit). <https://www.theguardian.com/world/2024/feb/16/air-canada-chatbot-lawsuit>.
- Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. [SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation](#). In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.
- Sitao Cheng, Liangming Pan, Xunjian Yin, Xinyi Wang, and William Yang Wang. 2024. [Understanding the interplay between parametric and contextual knowledge for large language models](#). *CoRR*, abs/2410.08414.
- Alexis Conneau, Germán Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. [What you can cram into a single \$\\$ \& ! \# *\$ vector: Probing sentence embeddings for linguistic properties](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2126–2136. Association for Computational Linguistics.
- Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502.
- Yang Deng, Lizi Liao, Liang Chen, Hongru Wang, Wenqiang Lei, and Tat-Seng Chua. 2023. Prompting and evaluating large language models for proactive dialogues: Clarification, target-guided, and non-collaboration. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10602–10621.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 82 others. 2024. [The llama 3 herd of models](#). *CoRR*, abs/2407.21783.
- Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. 2021a. [Measuring and improving consistency in pretrained language models](#). *Transactions of the Association for Computational Linguistics*, 9:1012–1031.
- Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard H. Hovy, Hinrich Schütze, and Yoav Goldberg. 2021b. [Measuring and improving consistency in pretrained language models](#). *Trans. Assoc. Comput. Linguistics*, 9:1012–1031.
- Constanza Fierro, Ruchira Dhar, Filippos Stamatiou, Nicolas Garneau, and Anders Søgaard. 2024. [Defining knowledge: Bridging epistemology and large language models](#). In *Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 16096–16111. Association for Computational Linguistics.
- R. A. Fisher. 1918. [The logic of inductive inference](#). *Journal of the Royal Statistical Society*, 98(1):39–54.
- Robert Friel and Atindriyo Sanyal. 2023. [Chainpoll: A high efficacy method for llm hallucination detection](#). *Preprint*, arXiv:2310.18344.
- Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. [Transformer feed-forward layers are key-value memories](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. [Retrieval augmented language model pre-training](#). In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3929–3938. PMLR.
- Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. [Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing](#). In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.
- Dan Hendrycks and Kevin Gimpel. 2016. [Gaussian Error Linear Units \(GELUs\)](#).
- Juyeon Heo, Christina Heinze-Deml, Oussama Elachqar, Kwan Ho Ryan Chan, Shirley You Ren, Andrew C. Miller, Udhyakumar Nallasamy, and Jaya Narain.

2025. [Do llms "know" internally when they follow instructions?](#) In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.
- John Hewitt and Christopher D. Manning. 2019. [A structural probe for finding syntax in word representations](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.
- Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. [spaCy: Industrial-strength Natural Language Processing in Python](#).
- 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):1–55.
- Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2023. [Atlas: Few-shot learning with retrieval augmented language models](#). *J. Mach. Learn. Res.*, 24:251:1–251:43.
- Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. [What does BERT learn about the structure of language?](#) In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. 2023. [Mistral 7b](#). *CoRR*, abs/2310.06825.
- Yibo Jiang, Goutham Rajendran, Pradeep Kumar Ravikumar, Bryon Aragam, and Victor Veitch. 2024. On the origins of linear representations in large language models. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, Proceedings of Machine Learning Research, pages 21879–21911. PMLR / OpenReview.net.
- Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. [How can we know what language models know?](#) *Transactions of the Association for Computational Linguistics*, 8:423–438.
- Patrick Kahardipraja, Reduan Achtibat, Thomas Wiegand, Wojciech Samek, and Sebastian Lapuschkin. 2025. [The atlas of in-context learning: How attention heads shape in-context retrieval augmentation](#). In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*.
- Nora Kassner, Philipp Dufter, and Hinrich Sch  tze. 2021. [Multilingual LAMA: Investigating knowledge in multilingual pretrained language models](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3250–3258, Online. Association for Computational Linguistics.
- Muhammad Khalifa, David Wadden, Emma Strubell, Honglak Lee, Lu Wang, Iz Beltagy, and Hao Peng. 2024. [Source-aware training enables knowledge attribution in language models](#). *Preprint*, arXiv:2404.01019.
- Diederik P. Kingma and Jimmy Ba. 2015. [Adam: A method for stochastic optimization](#). In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.
- Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. [Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation](#). In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.
- Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. 2019. [Quantifying the carbon emissions of machine learning](#). *CoRR*, abs/1910.09700.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich K  ttler, Mike Lewis, Wen-tau Yih, Tim Rock  tschel, Sebastian Riedel, and Douwe Kiela. 2020. [Retrieval-augmented generation for knowledge-intensive NLP tasks](#). In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.
- Daliang Li, Ankit Singh Rawat, Manzil Zaheer, Xin Wang, Michal Lukasik, Andreas Veit, Felix Yu, and Sanjiv Kumar. 2023. [Large language models with controllable working memory](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1774–1793, Toronto, Canada. Association for Computational Linguistics.
- Lisa R. Lifshitz and Roland Hung. 2024. [Bc tribunal confirms companies remain liable for information provided by ai chatbot](#).
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [Roberta: A robustly optimized BERT pretraining approach](#). *CoRR*, abs/1907.11692.

- Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. [Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 9004–9017. Association for Computational Linguistics.
- André F. T. Martins and Ramón Fernandez Astudillo. 2016. [From softmax to sparsemax: A sparse model of attention and multi-label classification](#). In *Proceedings of the 33rd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 1614–1623. JMLR.org.
- Jack Merullo, Noah Smith, Sarah Wiegrefe, and Yanai Elazar. 2025. On linear representations and pretraining data frequency in language models. In *International Conference on Learning Representations*, volume 2025, pages 71963–71987.
- OpenAI. 2024. [Gpt-4o mini: advancing cost-efficient intelligence](#). Accessed 2025-06-20.
- Kiho Park, Yo Joong Choe, and Victor Veitch. 2024. The linear representation hypothesis and the geometry of large language models. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, Proceedings of Machine Learning Research, pages 39643–39666. PMLR / OpenReview.net.
- Karl Pearson. 1901. [Liii. on lines and planes of closest fit to systems of points in space](#). *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572.
- Fabio Petroni, Patrick Lewis, Aleksandra Piktus, Tim Rocktäschel, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2020. [How context affects language models’ factual predictions](#). In *Conference on Automated Knowledge Base Construction, AKBC 2020, Virtual, June 22-24, 2020*.
- Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander H. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 2463–2473. Association for Computational Linguistics.
- Siya Qi, Yulan He, and Zheng Yuan. 2024. [Can we catch the elephant? a survey of the evolvement of hallucination evaluation on natural language generation](#). *Preprint*, arXiv:2404.12041.
- Qwen Team, 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, Jiayi Yang, Jingren Zhou, Junyang Lin, and 24 others. 2025. [Qwen2.5 technical report](#). *Preprint*, arXiv:2412.15115.
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. [Squad: 100, 000+ questions for machine comprehension of text](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 2383–2392. The Association for Computational Linguistics.
- Selvan Sunitha Ravi, Bartosz Mielczarek, Anand Kannappan, Douwe Kiela, and Rebecca Qian. 2024. [Lynx: An open source hallucination evaluation model](#). *Preprint*, arXiv:2407.08488.
- Raanan Y. Rohekar, Yaniv Gurwicz, Sungduk Yu, and Vasudev Lal. 2024. [Causal world representation in the gpt model](#). *Preprint*, arXiv:2412.07446.
- Artem Shelmanov, Ekaterina Fadeeva, Akim Tsvigun, Ivan Tsvigun, Zhuohan Xie, Igor Kiselev, Nico Daheim, Caiqi Zhang, Artem Vazhentsev, Mrinmaya Sachan, and 1 others. 2025. A head to predict and a head to question: Pre-trained uncertainty quantification heads for hallucination detection in llm outputs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 35700–35719.
- Adi Simhi, Itay Itzhak, Fazl Barez, Gabriel Stanovsky, and Yonatan Belinkov. 2025. Trust me, i’m wrong: High-certainty hallucinations in llms. *arXiv e-prints*, pages arXiv–2502.
- Weihang Su, Changyue Wang, Qingyao Ai, Yiran Hu, Zhijing Wu, Yujia Zhou, and Yiqun Liu. 2024. [Un-supervised real-time hallucination detection based on the internal states of large language models](#). In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 14379–14391. Association for Computational Linguistics.
- Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick. 2019. [What do you learn from context? probing for sentence structure in contextualized word representations](#). In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.
- Zineddine Tighidet, Andrea Mogini, Jiali Mei, Benjamin Piwowarski, and Patrick Gallinari. 2024. [Probing language models on their knowledge source](#). *CoRR*, abs/2410.05817.
- Christian Tomani, Kamalika Chaudhuri, Ivan Evtimov, Daniel Cremers, and Mark Ibrahim. 2024. Uncertainty-based abstention in llms improves safety and reduces hallucinations. *arXiv preprint arXiv:2404.10960*.

- Zijian Wang and Chang Xu. 2025. Functional abstraction of knowledge recall in large language models. *arXiv preprint arXiv:2504.14496*.
- Theodora Worledge, Judy Hanwen Shen, Nicole Meister, Caleb Winston, and Carlos Guestrin. 2024. [Unifying corroborative and contributive attributions in large language models](#). In *IEEE Conference on Secure and Trustworthy Machine Learning, SaTML 2024, Toronto, ON, Canada, April 9-11, 2024*, pages 665–683. IEEE.
- Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024. [Knowledge conflicts for llms: A survey](#). *Preprint*, arXiv:2403.08319.
- Yunzhi Yao, Ningyu Zhang, Zekun Xi, Mengru Wang, Ziwen Xu, Shumin Deng, and Huajun Chen. 2024. Knowledge circuits in pretrained transformers. *Advances in neural information processing systems*, 37:118571–118602.
- Hongbin Ye, Tong Liu, Aijia Zhang, Wei Hua, and Weiqiang Jia. 2024. [Cognitive mirage: A review of hallucinations in large language models](#). In *Proceedings of the First International OpenKG Workshop: Large Knowledge-Enhanced Models co-located with The International Joint Conference on Artificial Intelligence (IJCAI 2024), Jeju Island, South Korea, August 3, 2024*, volume 3818 of *CEUR Workshop Proceedings*, pages 14–36. CEUR-WS.org.
- Paul Youssef, Osman Alperen Koras, Meijie Li, Jörg Schlöterer, and Christin Seifert. 2023. Give me the facts! A survey on factual knowledge probing in pre-trained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, Findings of ACL, pages 15588–15605. Association for Computational Linguistics.
- Hao Zhang, Yuyang Zhang, Xiaoguang Li, Wenxuan Shi, Haonan Xu, Huanshuo Liu, Yasheng Wang, Lifeng Shang, Qun Liu, Yong Liu, and Ruiming Tang. 2024. [Evaluating the external and parametric knowledge fusion of large language models](#). *Preprint*, arXiv:2405.19010.
- Jun Zhao, Yongzhuo Yang, Xiang Hu, Jingqi Tong, Yi Lu, Wei Wu, Tao Gui, Qi Zhang, and Xuanjing Huang. 2025. [Understanding parametric and contextual knowledge reconciliation within large language models](#). In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.
- Guido Zuccon, Bevan Koopman, and Razia Shaik. 2023. [Chatgpt hallucinates when attributing answers](#). In *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region, SIGIR-AP 2023, Beijing, China, November 26-28, 2023*, pages 46–51. ACM.

A Resources

This appendix reports the computational cost of the work presented here, including exploratory and discarded runs. We distinguish three quantities: the total cost of the research program, the cost of reproducing the pipeline, and the marginal cost of extending it to an additional target model.

A.1 Data Generation (API)

Total GPT-4O-MINI usage across the project was approximately 165M input tokens and 17M output tokens ($\approx$ \$35). This figure covers the complete research process: exploratory runs, prompt iteration, discarded configurations, and an initial sample of Wikipedia pages that was later superseded. It is substantially larger than the cost of the final pipeline and should be read as the cost of *arriving at* ATTRIWIKI, not of building it. Two narrower figures are more informative for practitioners.

Extending to a New Target Model. Because entity selection and knowledge-test prompt generation are model-independent, adding a target model to an existing entity set requires only the model-specific stages, namely entity removal and completion-cue generation, which are conditioned on which entities the model recalls. We measured this directly when adding QWEN2.5-7B: 7.9M input and 1.9M output tokens, or \$2.33 at the pricing in effect at the time. We additionally reused removals across models wherever their known/unknown labels coincided, so a model whose parametric coverage overlaps less with the existing set would incur a somewhat higher cost.

Building From a New Corpus. A clean build on a new data source must re-run the model-independent stages. We estimate this at roughly 80M input tokens and 10M output tokens for a corpus comparable to our 20k Wikipedia passages. The estimate is dominated by knowledge-test generation, which passes a full passage as input for every passage and every candidate entity and a knowledge question as output; the model-specific stages contribute under 10M tokens by the measurement above.

A.2 Target Model Inference and Probe Training (GPU)

Knowledge testing, hidden-state extraction, and probe training were run on a single NVIDIA A100.

Knowledge testing dominates GPU cost, as it evaluates three prompt formats for every candidate entity across the sampled passages, taking up to 7 hours per target model on a single GPU. Hidden-state extraction over the rewritten passages is cheaper, requiring up to 3 hours per model despite capturing full-depth hidden states at two token positions, as it involves a single pass over the constructed prompts. Total inference cost is therefore bounded at roughly 10 GPU-hours per target model, or 30 GPU-hours across the three models. Probe training is negligible by comparison: a single *Layer-LR* probe converges in under a minute.

Following [Lacoste et al. \(2019\)](#), 30 GPU-hours on the A100 draws approximately 12 kWh, which at a carbon intensity of 0.268 kg/kWh for the Dutch grid and a data center PUE of 1.19 corresponds to roughly 3.8 kg CO₂eq.

A.3 Cost Profile of the Method

The cost structure of attribution probing is favorable, and we consider this a substantive property of the approach rather than an incidental one. The expensive stages are incurred once: dataset construction and hidden-state extraction are one-off, and the model-independent portion is amortized across every target model subsequently probed. What remains at deployment is a linear probe over hidden states that are already computed during the forward pass, adding negligible inference overhead. This contrasts with sampling-based hallucination detection such as SelfCheckGPT ([Manakul et al., 2023](#)), which requires multiple full generations per query, and with source-aware training ([Khalifa et al., 2024](#)), which requires retraining the model itself.

The principal recurring cost is model dependence, as noted in the Limitations: each target model requires its own probe, and a model update requires regenerating the model-specific data and retraining. The measurement above bounds this at a few dollars of API usage plus several GPU-hours per model, which we regard as acceptable for a diagnostic that is otherwise free at inference time.

B Hyperparameters

This appendix documents all hyperparameters used in our experiments. We separate configurations for *bias evaluation* models and *attribution classifiers*.

B.1 Bias evaluation

We evaluate bias using two complementary classifiers: a sparse BoW model and an embedding-based model. The former serves as a lightweight lexical baseline, while the latter tests whether bias signals persist in contextualized representations. The corresponding hyperparameters given in [Tables 6](#) and [7](#), respectively.

Component	Configuration
Text representation	TF-IDF ngram_range = (1,2) max_features = 5000
Classifier	Logistic regression class_weight = balanced max_iter = 1000 random_state = 42

Table 6: Hyperparameters for the **TF-IDF logistic regression** bias classifier.

Component	Configuration
Text representation	Transformer embeddings model = microsoft/deberta-v3-large max_length = 128 padding = max_length truncation = true repr. = [CLS] dtype = float16 eval mode, no gradients
Classifier	Logistic regression max_iter = 1000 random_state = 42

Table 7: Hyperparameters for the **embedding-based** bias classifier.

B.2 Attribution classifiers

Attribution classifiers are trained to probe internal model representations. Hyperparameters for **Layer-LR** and **Layer-MLP** were selected via grid search over dropout p ($\{0, 0.1, 0.2\}$), weight decay λ ($\{0, 5 \times 10^{-4}, 10^{-3}, 2 \times 10^{-3}\}$), learning rate η ($\{5 \times 10^{-4}, 10^{-3}, 2 \times 10^{-3}\}$), and (for Layer-MLP) bottleneck size m ($\{64, 128\}$), using a title-aware 15% validation split. For the grid search, models were trained for 10 epochs; the selected configuration was then trained for up to 100 epochs with early stopping on validation macro- F_1 (tie-breaker: validation accuracy). Selected values are reported in [Tables 8](#) to [10](#).

Component	Configuration
Input representation	Last transformer layer ($X[:, -1, :]$)
Preprocessing	StandardScaler (default)
Classifier	Logistic regression class_weight = balanced solver = lbfgs penalty = ℓ_2 max_iter = 1000

Table 8: Hyperparameters for **Final-LR** (last-layer logistic regression).

Component	Configuration
Input representation	Layer stack $X \in \mathbb{R}^{N \times L \times H}$ weighted aggregation
Layer weighting	Softmax(θ), $\sum_{\ell} \alpha_{\ell} = 1$
Normalization	ℓ_2 per layer
Dropout	$p = 0.1$
Classifier head	Linear($H \rightarrow 1$)
Loss	BCEWithLogitsLoss pos_weight = #neg/#pos
Optimizer	AdamW learning rate $\eta = 2 \times 10^{-3}$ weight decay $\lambda = 10^{-3}$
Training	batch size = 64 max epochs = 100 early stopping (macro- F_1 , patience = 3) title-aware validation split (0.15, seed = 42)

Table 9: Hyperparameters for **Layer-LR** (layer-weighted linear classifier).

C Additional Results

C.1 Latent Space Analysis

Here we report the full latent space plots from Figure 2 for all models, with Figure 4 for Llama-3.1-8B, Figure 5 for Mistral-7B-v0.1, and Figure 6 for Qwen2.5-7B.

C.2 Out-of-distribution and Generalization Results

This section reports additional experimental results that support the main findings in the paper, including ablation studies and out-of-domain generalization performance. Table 11 shows the accuracy for the answer-string decoy variant for SQUAD, where we can see that the MLP variant of the probe almost collapses to random accuracy, suggesting potential overfitting. Table 12 shows the results for the MLP probe on SQUAD and WEBQUESTIONS.

Comparison to Tighidet et al. (2024). In the following Tables 13 to 15 we compare the per-group

Component	Configuration
Input representation	Layer stack $X \in \mathbb{R}^{N \times L \times H}$ weighted aggregation
Layer weighting	Sparsemax(θ)
Normalization	ℓ_2 per layer
MLP head	Linear($H \rightarrow m$), $m = 64$ GELU, Dropout($p = 0.1$) Linear($m \rightarrow 1$)
Loss	BCEWithLogitsLoss pos_weight = #neg/#pos
Optimizer	AdamW learning rate $\eta = 10^{-3}$ weight decay $\lambda = 10^{-3}$
Training	batch size = 64 max epochs = 100 early stopping (macro- F_1 , patience = 3) title-aware validation split (0.15, seed = 42)

Table 10: Hyperparameters for **Layer-MLP** (layer-weighted MLP classifier).

Classifier	Accuracy
Layer-LR	.724
Layer-MLP	.541

Table 11: Answer-string decoy ablation on SQUAD using Llama-3.1-8B. While performance drops for all classifiers, the linear probe remains well above chance, whereas the MLP collapses to near-random accuracy, indicating reliance on entity repetition.

macro- F_1 of Tighidet et al. (2024)’s best-layer probe against the ATTRIWIKI Layer-LR probe on FTG across all three models. Relation groups well-represented in encyclopedic Wikipedia text, such as Geographic and Media, transfer well, with ATTRIWIKI consistently matching or exceeding Tighidet et al.’s performance. The Corporate group is a consistent outlier. This reflects the fact that ATTRIWIKI’s training data contains essentially no product-ownership facts. These results confirm that domain shift does affect the ATTRIWIKI probe, but only in a targeted and predictable way: performance degrades specifically for themes absent from

Model	SQuAD	WebQ
Llama-3.1-8B	.972	.959
Mistral-7B-v0.1	.958	.948
Qwen2.5-7B	.998	.980

Table 12: Generalization accuracy of Layer-MLP (FTG) on out-of-domain QA datasets. WebQ = WebQuestions.

Wikipedia-style factual text.

Relation group	n	Tighidet et al. (L10)	ATTRIWIKI
Corporate	100	.333 $\pm$.023	.369 $\pm$.035
Geographic	920	.780 $\pm$.014	.837 $\pm$.012
Media	16	.329 $\pm$.057	.800 $\pm$.104

Table 13: Per-group macro- F_1 (mean $\pm$ bootstrap std, $B = 1000$) for **Mistral 7B**. **Tighidet et al.** uses the best overall layer (L10).

Relation group	n	Tighidet et al. (L11)	ATTRIWIKI
Corporate	52	.411 $\pm$.061	.541 $\pm$.070
Geographic	530	.670 $\pm$.021	.763 $\pm$.019
Media	50	.566 $\pm$.075	.878 $\pm$.046
Occupy-position	16	.328 $\pm$.058	1.000 $\pm$.000
Play-instrument	20	.589 $\pm$.113	.897 $\pm$.070

Table 14: Per-group macro- F_1 (mean $\pm$ bootstrap std, $B = 1000$) for **Llama 3.1 8B**. **Tighidet et al.** uses the best overall layer (L11).

Relation group	n	Tighidet et al. (L22)	ATTRIWIKI
Corporate	640	.762 $\pm$.017	.633 $\pm$.019
Geographic	2928	.754 $\pm$.008	.812 $\pm$.007
Hierarchy	46	.686 $\pm$.073	.768 $\pm$.062
Media	178	.798 $\pm$.029	.886 $\pm$.024
Occupy-position	136	.624 $\pm$.042	.837 $\pm$.031
Play-instrument	20	.329 $\pm$.053	.303 $\pm$.054
Religion	208	.794 $\pm$.029	.840 $\pm$.025

Table 15: Per-group macro- F_1 (mean $\pm$ bootstrap std, $B = 1000$) for **Qwen2.5 7B**. **Tighidet et al.** uses the best overall layer (L22).

Per-layer PCA of Hidden State Activations — Llama-3.1-8B (layers 1-32)

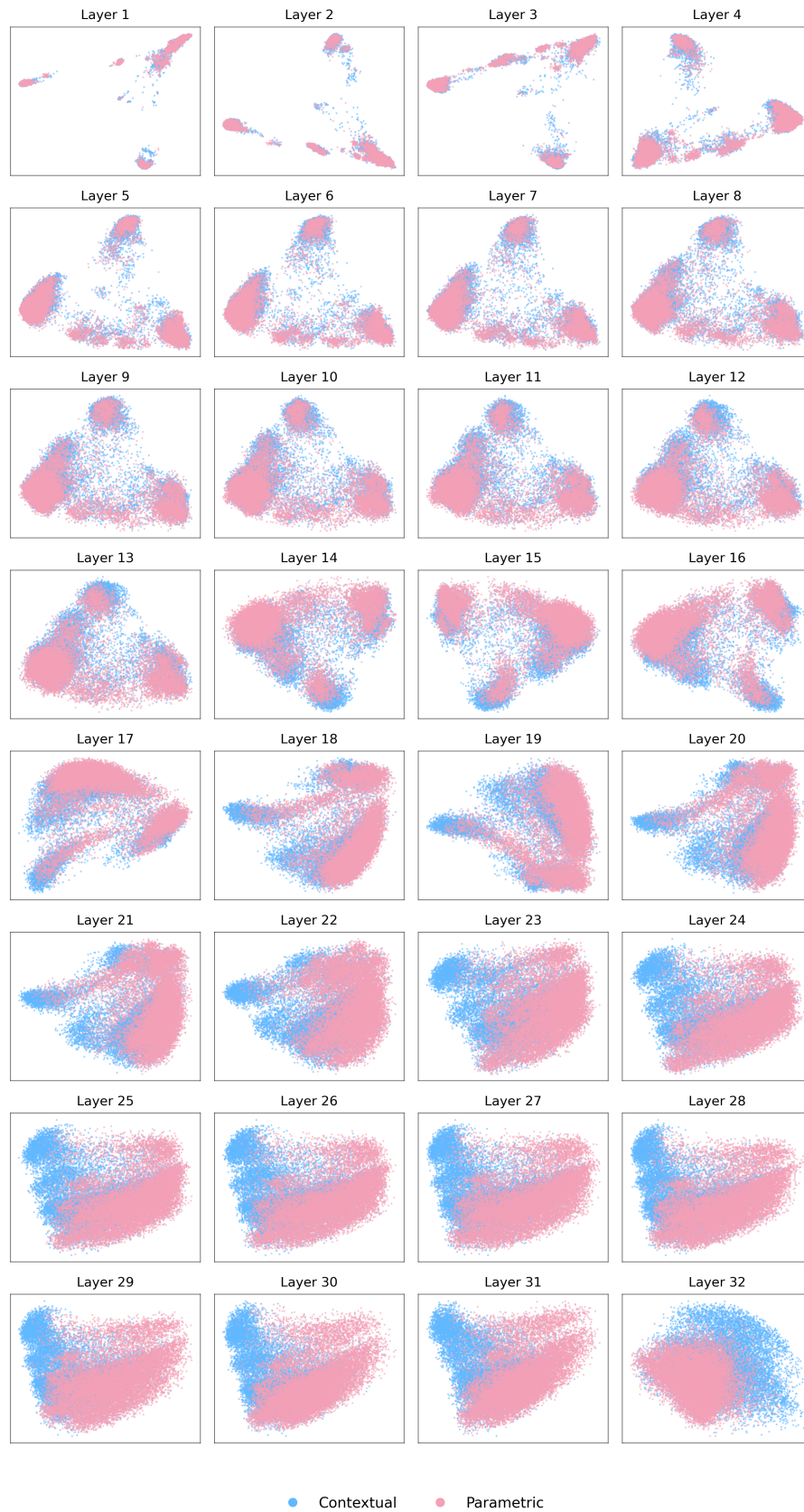


Figure 4: Per-layer PCA of hidden states for Llama-3.1-8B.

Per-layer PCA of Hidden State Activations — Mistral-7B-v0.1 (layers 1–32)

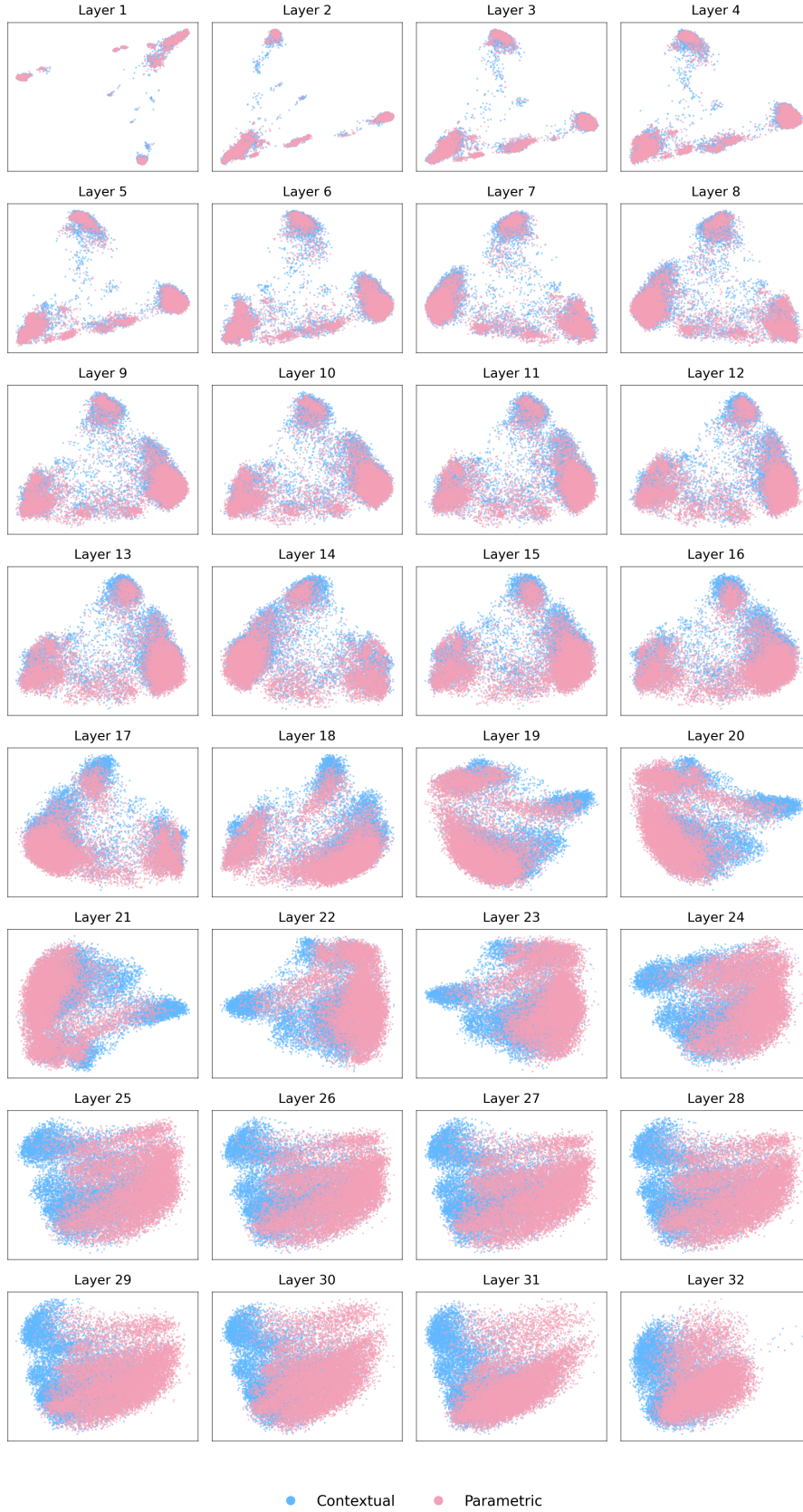


Figure 5: Per-layer PCA of hidden states for Mistral-7B-v0.1.

Per-layer PCA of Hidden State Activations — Qwen2.5-7B (layers 1-28)

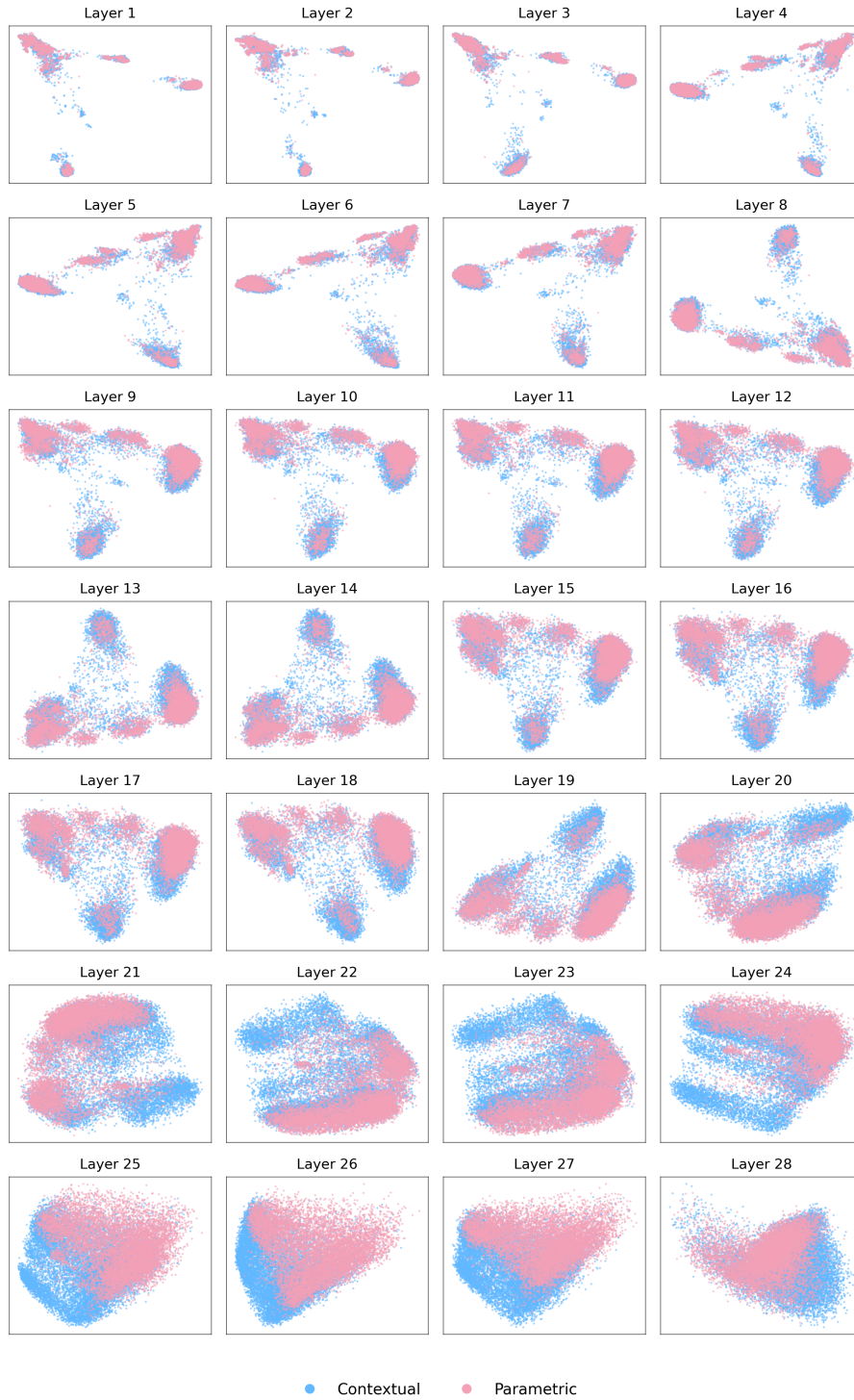


Figure 6: Per-layer PCA of hidden states for Qwen2.5-7B.

D Prompts

D.1 ATTRIWIKI

D.1.1 Entity selection.

This prompt was designed to select up to three entities to prevent longer passages containing more entities from being overrepresented in the data.

Prompt: Entity Selection

Given a Wikipedia passage and a list of mentioned entities, you are tasked to narrow down the list of entities to a maximum of three using the following criteria.

1. Remove entities that refer to a number, e.g., “first”, “last”, “twelve”, “fourth”.
2. Remove entities that refer to a very specific date, e.g., a range, months, days etc. Just a year is acceptable.
3. Remove entities that refer to any quantity.
4. The resulting entities should contain both well-known examples and lesser-known examples.

The string of the remaining entities must be exactly the same.

The passage:

{text}

The entities:

{entities}

Maximum of three selected entities:

D.1.2 Knowledge testing.

Given a Wikipedia passage and an entity, we present examples of the three knowledge tests.

1. Alice: I can't remember exactly who was the king of England in 1265 during the Battle of Evesham. I can't remember.
Bob: Actually, I know. It was **{King Henry III.}**
2. Q: Who was the king of England during the Battle of Evesham in the 13th century?
A: **{King Henry III}**
3. The Battle of Evesham (4 August 1265) was one of the two main battles of 13th century England 's Second Barons ' War . It marked the defeat of Simon de Montfort, Earl of Leicester , and the rebellious barons by Prince Edward - later King Edward I - who led the forces of his father , **{King Henry III.}**

1. Dialogue

Prompt: Knowledge Test (Alice-Bob Dialogue)

You will receive a Wikipedia passage of an arbitrary topic and an entity that is mentioned somewhere within the passage. You will create a dialogue between Alice and Bob. **Alice can't think of the name of** [entity]. She describes it perfectly using the Wikipedia passage. **Bob is all-knowing**, and tells Alice the name of the entity.

- Alice is **not allowed** to say the name of the entity or part of the entity.
- Alice **must only use information** provided in the Wikipedia passage to describe said entity.
- Bob **must say the exact name** of the entity.

Here is an example.

Wikipedia passage:

< The Battle of Evesham (4 August 1265) was one of the two main battles of 13th century England 's Second Barons ' War . It marked the defeat of Simon de Montfort , Earl of Leicester , and the rebellious barons by Prince Edward - later King Edward I - who led the forces of his father , King Henry III . It took place on 4 August 1265 , near the town of Evesham , Worcestershire . >

Entity:

< King Henry III >

Alice-Bob conversation:

< Alice: I can't remember exactly who was the king of England in 1265 during the Battle of Evesham. I can't remember.
Bob: Actually, I know. It was King Henry III. >

Now it is your turn:

Wikipedia passage:

< {passage} >

Entity:

< {entity} >

Alice-Bob conversation (ending with >):

<

2. QA-style

Prompt: Knowledge Test (Question-Answer Style)

You will receive a Wikipedia passage of an arbitrary topic and an entity that is mentioned somewhere within the text. Like so:

Wikipedia passage:

< The Battle of Evesham (4 August 1265) was one of the two main battles of 13th century England 's Second Barons ' War. It marked the defeat of Simon de Montfort, Earl of Leicester, and the rebellious barons by Prince Edward - later King Edward I - who led the forces of his father, King Henry III. It took place on 4 August 1265, near the town of Evesham, Worcestershire. >

Entity:

< King Henry III >

Question-Answer pair:

< Q: Who was the king of England during the Battle of Evesham in the 13th century?
A: King Henry III >

Follow the same pattern using a new Wikipedia passage. You generate a question (Q) to which the answer is the entity (A).

Now it is your turn:

Wikipedia passage:

< {passage} >

Entity:

< {entity} >

Question-Answer (ending with >):

<

Note that the trailing entity mentions are removed using an exact string match in a later phase.

3. Truncated passage. The truncated passage is an exact string match, removing the first mention of the entity in the Wikipedia passage and removing all text succeeding and including the entity.

D.1.3 Entity removal.

Prompt: Entity Removal

Remove every explicit mention of the entity and any variant (abbreviation, nickname, unambiguous pronoun) from the passage while preserving factual content and fluency.

One illustrative example:

Entity:

United Kingdom

Original passage:

Winston Churchill was a British statesman, soldier, and writer who served as Prime Minister of the United Kingdom from 1940 to 1945 and again from 1951 to 1955. He led Britain to victory in the Second World War. Among the British public, he is widely considered the greatest Briton of all time. He was born to an aristocratic family in Oxfordshire, England.

Rewritten passage:

Winston Churchill was a statesman, soldier, and writer who served as Prime Minister from 1940 to 1945 and again from 1951 to 1955. He led the nation to victory in the Second World War. Among the public, he is widely considered one of the greatest leaders of all time. He was born to an aristocratic family in Oxfordshire.

Constraints:

- Do **not** insert placeholder tokens such as "_____" or "[ENTITY]".
- **Forbidden filler words:** another, other, elsewhere, someplace, something, someone, notable, major, region, and global.
- If deleting the entity breaks a sentence, repair the grammar so the passage would pass professional copy-edit.
- Keep all dates, numbers, and named entities that do **not** refer to the target entity.
- If the entity acts as an adjective (e.g., "UK policy"), rewrite only that phrase so the head noun remains (e.g., "the policy").
- **Meta-instruction:** Do **not** mimic specific wording from the example; use whatever phrasing fits the original passage.

Now your turn

Entity:

{entity}

Original passage:

< {passage} >

Rewritten passage (end with >):

<

D.1.4 Appending completion sentence.

Prompt: Completion Sentence Generation

Instruction:

You are given a Wikipedia passage that contains an entity. Then you are given the entity's name. Produce one additional sentence that naturally appends to the passage, reaffirming the entity's identity. You may use one of the following formats (or a similarly natural variant):

- "The *[description]* is called *[entity]*."
- "The *[description]* is named *[entity]*."
- "Locals called it *[entity]*."
- "They refer to it as *[entity]*."

IMPORTANT:

1. You must use the **exact entity name as provided**—no alterations, changes in capitalization, or partial usage.
2. Your output should only be **that single appended sentence**.
3. The sentence should be generic and do a good job thoroughly introducing the entity. It should naturally lead up to naming the entity, so that if the entity were removed, a model would be likely to complete the sentence with that entity.

One-Shot Example

Passage:

< Frankenstein is a gothic novel by Mary Shelley that was first published in 1818. The story follows a young scientist who creates a sapient creature through an unorthodox experiment, and it is often hailed as the first true work of science fiction. >

Entity:

< Mary Shelley >

Output:

< The author of the novel Frankenstein is named Mary Shelley. >

Notice how:

- The entity is introduced **exactly as given**.
- The sentence flows **naturally from the passage**.

Now it is your turn.

Entity:

< {entity} >

Original Passage:

< {passage} >

Output (ending with >):

<

D.2 SQuAD and WebQuestions

One-shot prompts for SQuAD and WebQ.

SQuAD

Context: Frank Herbert was an American science-fiction author best known for his novel Dune. Example Q: Who wrote Dune? A: Frank Herbert

Context: ...

Q: ... A: ...

WebQuestions

Example Q: Who wrote “Dune”? A: Frank Herbert

Q: ... A: ...