Do LLMs Adhere to Label Definitions? Examining Their Receptivity to External Label Definitions

Seyedali Mohammadi[1], Bhaskara Hanuma Vedula[2], Hemank Lamba[3], Edward Raff[1,4], Ponnurangam Kumaraguru[2], Francis Ferraro[1], Manas Gaur[1]

[1]UMBC, [2]IIIT Hyderabad, [3]Dataminr, Inc., [4]CrowdStrike

{m294,edraff1,ferraro,manas}@umbc.edu,vedula.hanuma@research.iiit.ac.in, pk.guru@iiit.ac.in

hlamba@dataminr.com

Abstract

Do LLMs genuinely incorporate external definitions, or do they primarily rely on their parametric knowledge? To address these questions, we conduct controlled experiments across multiple explanation benchmark datasets (general and domain-specific) and label definition conditions, including expert-curated, LLM-generated, perturbed, and swapped definitions. Our results reveal that while explicit label definitions can enhance accuracy and explainability, their integration into an LLM's task-solving processes is neither guaranteed nor consistent, suggesting reliance on internalized representations in many cases. Models often default to their internal representations, particularly in general tasks, whereas domain-specific tasks benefit more from explicit definitions. These findings underscore the need for a deeper understanding of how LLMs process external knowledge alongside their pre-existing capabilities.

1 Introduction

Label Definitions are considered as grounded statements that provide context on *what an AI model* needs to do upon receiving a query (Mu et al., 2024; Deng et al., 2025). These definitions are considered as clues to disambiguate unclear labels, helping models perform their tasks more effectively (Peskine et al., 2023; Kumar et al., 2023; Xie et al., 2023). However, whether models truly process and incorporate these definitions into their decision-making remains unclear.

To highlight this, consider the example in Figure 1, where we test *LLaMA-3*'s behavior under different permutations of label definitions on an instance of the e-SNLI dataset. When permutations 1 and 3 maintain the standard definition of *"Neutral"* (cases where the premise neither entails nor contradicts the hypothesis) the model successfully classifies the relationship. However, when the *"Neutral"* label is assigned an incorrect definition (permutation 2), *LLaMA-3* misclassifies the input.

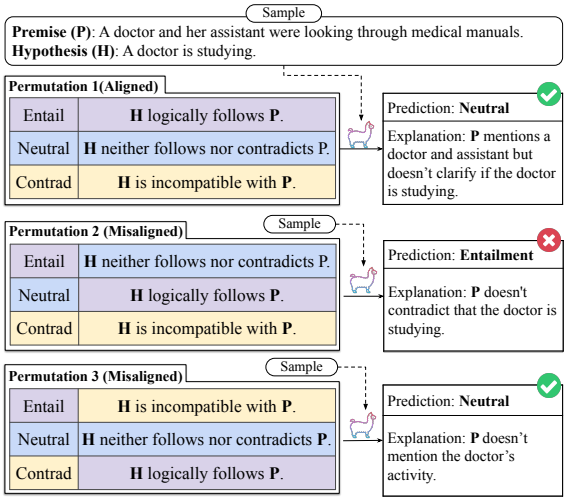

Figure 1: The receptivity of LLMs to external definitions is revealed through a permutation analysis of the e-SNLI dataset. By testing all six possible orderings of entailment, neutral, and contradiction definitions, we discovered the model's receptivity varies significantly. In the illustrated example, the ground-truth label is "neutral." Interestingly, in each permutation, the model consistently predicts the label that is mapped to the neutral definition, regardless of its original label name.

This example reveals a fundamental challenge in how LLMs process conflicting information sources. When external definitions clash with a model's training data, we see unpredictable behavior; sometimes the model follows the external guidance, other times it relies on internal knowledge (Burns et al., 2022; Duan et al., 2024). To understand this systematically, we need to examine two key aspects: how models handle *Knowledge Conflict* (when external definitions contradict training) and *Definition Integration* (how definitions are presented in prompts) (Azaria and Mitchell, 2023; Gaur et al., 2022). The *LLaMA-3* case demonstrates both issues—the model's internal understanding of "neutral" conflicts with the manipulated definition, yet it still attempts to integrate the external guidance into its reasoning. This suggests that explanation-based classification may be more vulnerable to definitional inconsistencies than previously assumed.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 32380–32393

To better understand the role of definitions in *explanation-based classification* tasks using LLMs, we investigate the following questions: ***Q1:*** Do LLMs rely more on their internalized, potentially opaque representations, raising questions about the interplay between external guidance and inherent knowledge? ***Q2:*** Do they consistently adhere to external definitions? In this work, we restrict our study to classification tasks, as they provide clear ground truth labels that make it easier to assess whether models follow external definitions or revert to internal knowledge. We chose this focus deliberately: classification enables objective evaluation on widely used datasets (e.g., e-SNLI), establishes strong baselines for comparability, and reflects many real-world applications such as content moderation, diagnosis, or categorization. At the same time, our analysis of generated explanations already shows similar domain- and definition-sensitivity patterns, suggesting that the core findings may extend beyond classification to generation and reasoning tasks, which we leave for future work.

To answer these questions systematically, we conduct experiments across four diverse datasets: e-SNLI for general natural language inference (Camburu et al., 2018), WELLXPLAIN for mental health (Garg, 2024), HATEXPLAIN for hate speech detection (Mathew et al., 2021), and WICE for fact-checking (Kamoi et al., 2023). This multi-domain approach allows us to test whether our findings generalize beyond any single task or domain. Our main contributions are as follows: 1) We develop evaluation strategies to assess the performance and explainability of LLMs when provided with label definitions, which serve as a lightweight and interpretable form of external knowledge injection. 2) We design two probing strategies, definition permutation and definitional perturbation, to assess the sensitivity of LLMs to definition quality and alignment. 3) We conduct extensive experiments across four diverse benchmarks, e-SNLI, WELLXPLAIN, HATEXPLAIN, and WICE, to assess model receptivity to label definitions.

Our analysis yields four key insights: First, we show that models heavily rely on provided label definitions, with performance dropping significantly when definitions are swapped or corrupted. Second, we find that domain-specific tasks require more precise definitions than general tasks like natural language inference. Third, we demonstrate that LLM-generated definitions often outperform expert-written ones, particularly when they are tailored to specific inputs through simple retrieval methods, such as K-Nearest Neighbors (K-NN) (Sheth et al., 2021). Finally, we reveal systematic differences in how models integrate definitions depending on the task domain and definition source. Code for reproducing our experiments is available at `https://github.com/mohammadi-ali/Definition-Receptivity-LLMs`.

2 Problem Formulation

We evaluate the proclivity of an LLM $\mathcal{M}$ to adhere to a grounded definition[1] G and generative definition[2] R through **Knowledge Conflict**, i.e., conflicting knowledge provided to the model and its internal knowledge, and through **Definition Integration**, i.e., how to provide definitions to the model.

2.1 Knowledge Conflict

Conflicts arise when a model's pre-trained knowledge conflicts with newly provided definitions, potentially affecting classification performance and explanation quality (Xie et al., 2023). To assess the model's reliance on external definitions versus its internalized knowledge, we introduce two scenarios.

Our first scenario examines *varying degrees of definition accuracy* using three categories: *incorrect*, *slightly incorrect*, and *correct* definitions. *Incorrect* definitions completely misalign labels with contradictory explanations—for instance, defining "contradiction" in e-SNLI as a scenario where the hypothesis must always be true if the premise is true. *Slightly incorrect* definitions introduce subtle inaccuracies that mislead without being entirely wrong—defining "entailment" as a relationship where the hypothesis is *possibly* true given the premise, rather than *necessarily* true. *Correct* definitions use established logical relationships that align precisely with the intended classification task.
Note. We distinguish these categories from *disputed definitions*, such as differing expert interpretations of terms like "fluency" or "readability." These are not incorrect but represent legitimate divergences in perspective. While crucial in human evaluation settings, they introduce ambiguity that makes it difficult to isolate whether model performance changes arise from definition sensitivity or

[1] Grounded refers to definitions established by experts.
[2] Generative are those produced by LLM, e.g., *GPT-4*.

from inherent conceptual disagreement. For clarity, our study therefore focuses on clear right–wrong contrasts.

The second scenario, *permutation-based*, involves swapping label definitions. Given a set of p label-definition pairs (l_i, d_i) as $G = \{(l_1, d_1), (l_2, d_2), \ldots, (l_p, d_p)\}$. The set $\mathcal{P}$ contains all $p!$ possible permutations of G which is denoted as the following:

$$\mathcal{P} = \{\pi(G) | \pi \in S_p\}$$

where S_p represents all possible ways to rearrange p elements, containing $p!$ permutations. In this set, one unique permutation (π_c) represents the correct alignment, where each label matches its true definition, while the remaining $(p! - 1)$ permutations (π_m) represent misaligned label-definition pairs.

Performance and Explanation Evaluation: To assess how definition manipulation affects model performance, we use the Matthews Correlation Coefficient (MCC), a robust metric for binary classification that accounts for all four confusion matrix categories (Matthews, 1975). MCC ranges from -1 to +1, where +1 indicates perfect classification, 0 represents random performance, and -1 indicates complete disagreement between predictions and ground truth . Unlike accuracy or F1-score, MCC remains reliable even with imbalanced datasets.

For our permutation-based experiments, we compare performance when label-definition pairs maintain their correct associations, referred to as aligned mapping (MCC_a), against the average performance when these pairs are misaligned:

$$\overline{\text{MCC}}_m = \frac{1}{(p! - 1)} \sum_{\substack{i=1 \\ i \neq c}}^{p!} \text{MCC}_{\pi_i}$$

where π_c represents the correct label-definition mapping and π_i denotes the i-th permutation.

For definition accuracy experiments, we measure performance across three levels: $\text{MCC}_{\text{Inc.}}$ (incorrect definitions), $\text{MCC}_{\text{SInc.}}$ (slightly incorrect), and $\text{MCC}_{\text{Cor.}}$ (correct definitions). For most datasets (e-SNLI, HATEXPLAIN, WELLXPLAIN), models generate textual explanations that we evaluate using MCC. However, WICE is a fact-checking dataset where explanations consist of supporting sentence indices rather than generated text. For this task, we report F1 scores calculated as:

$$\text{F1} = \begin{cases} 1.0 & \text{if } |I_{\text{gold}}| = |I_{\text{pred}}| = 0 \\ 0.0 & \text{if } |I_{\text{gold}}| = 0 \oplus |I_{\text{pred}}| = 0 \\ \frac{2 \cdot \text{P} \cdot \text{R}}{\text{P}+\text{R}} & \text{otherwise} \end{cases}$$

Where I_{gold} and I_{pred} represent gold and predicted supporting sentence indices, and P and R are precision and recall, respectively.

2.2 Definition Integration

While our Knowledge Conflict experiments examine *what happens* when definitions contradict model knowledge, Definition Integration investigates *how* the presentation and source of definitions affects model behavior.

We systematically evaluate four definition integration strategies using a test set $\mathcal{S}_{test} = \{(X_i, Y_i, E_i)\}_{i=1}^{n}$, where X_i represents the input, Y_i the true label, and E_i the ground truth explanation. The four conditions are: *(i) vanilla* (zero-shot), where models rely purely on internal knowledge without explicit definitions (Figure 2(a)); *(ii) fixed definition*, where expert-written definitions are incorporated directly into prompts (Figure 2(b)); *(iii) adjusted definition*, where definitions are dynamically generated for each input sample using the same LLM performing the classification task (Figure 2(c)); and *(iv) definition + few-shot*, which combines fixed definitions with exemplar cases. These conditions allow us to assess the trade-offs between relying on internal model knowledge versus external guidance, and between generic versus context-specific definitions. For most tasks, we evaluate explanation quality using ROUGE scores, while WICE uses F1 scores for sentence-level explanation matching.

For the adjusted definition condition, we employ a k-NN approach to generate context-specific definitions that are tailored to each input. Given an input X, we retrieve the k most similar training examples for each label l using cosine similarity in embedding space, denoted as $\mathcal{N}_k(X, l)$. An LLM $\mathcal{M}$ then generates customized definitions based on these retrieved examples. The adjusted definitions are formalized as:

$$D_{adjusted} = \{\mathcal{M}(\mathcal{N}_k(X, l))\}_{l=1}^{L}$$

We experiment with $k \in \{1, 2, 5, 10\}$ and extend to 15 or 20 when necessary to identify optimal retrieval sizes. This approach allows us to

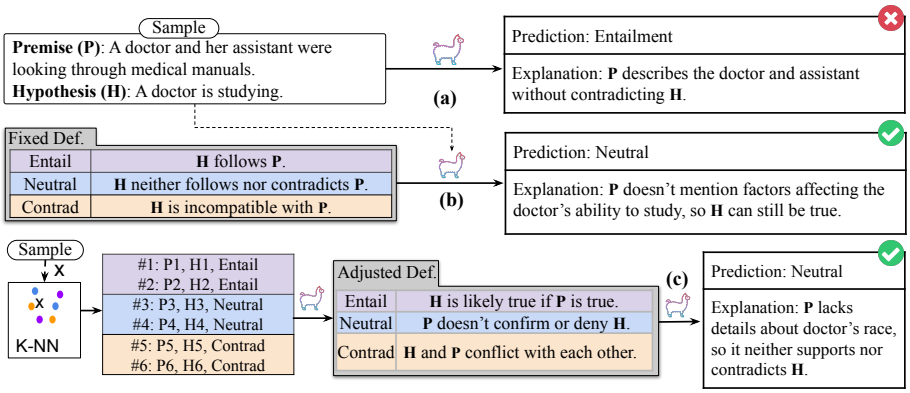

Figure 2: An illustration of how *LLaMA-3*'s natural language inference performance improves with label definition integration. (a) baseline performance where *LLaMA-3* incorrectly labels a premise-hypothesis pair as "Entailment"; (b) improved accuracy when grounded label definitions are provided, leading to correct "Neutral" classification; and (c) successful explanation and classification when presented with both definitions and few-shot examples. The same LLM (*LLaMA-3*) is used to generate label definitions. The few-shot samples used to generate such definitions are in Table 4 (appendix A). Note that the adjusted definitions shown in the figure are abbreviated; the complete versions are provided in appendix F.

test whether input-specific definitions, which potentially capture nuanced contextual information, improve both classification accuracy and explanation quality compared to static, expert-written definitions.

3 Setup and Methodology

Models: To investigate definition receptivity across diverse model architectures, we selected four state-of-the-art LLMs with deliberately varied characteristics: *GPT-4*, *LLaMA-3*, *Phi-3*, and *Mistral-7B*.

This selection spans from compact 3.8B parameter models to trillion-parameter architectures, enabling us to determine whether investments in larger models yield proportional improvements in definition adherence—a key consideration for AI deployment in high-stakes domains where consistent interpretation of instructions is essential. Since definitions are concise, their inclusion adds minimal overhead in terms of token length. Nonetheless, this setup is primarily designed for short to medium-length inputs. For tasks requiring longer contexts (e.g., document-level inputs), the approach may not be suitable for smaller models like *Phi-3*, which has a 4k token limit in our case, thereby restricting the ability to integrate external definitions without truncation.

Each model represents distinct design philosophies with practical implications for deploying LLMs in definition-sensitive contexts: *GPT-4* serves as our high-performance benchmark, representing proprietary models that organizations might access through APIs (Achiam et al., 2023); *LLaMA-3* exemplifies open-source alternatives increasingly adopted in industry settings requiring customization and privacy (Meta, 2024); *Phi-3* addresses the growing need for efficient, deployable models in resource-constrained environments (Bilenko, 2024); and *Mistral-7B* represents models optimized for the performance-efficiency trade-off crucial for commercial applications (Jiang et al., 2023). This diversity allows us to provide actionable insights on which architectural approaches best ensure definitional consistency across different deployment scenarios—whether in cloud infrastructure, edge devices, or hybrid settings. Detailed specifications for each model are provided in Table 5 (appendix A), with a representative prompt template shown in Table 6 (appendix A). For *GPT-4*, we balanced representation and cost by sampling 500 instances per label across datasets[3]. For all other models, we used complete test sets.

Datasets: Our experiments utilize four diverse datasets representing both general and domain-specific classification tasks. The e-SNLI dataset examines natural language inference with three labels: "entailment" (hypothesis necessarily follows from premise), "neutral" (hypothesis might be true given the premise), and "contradiction" (hypothesis cannot be true given the premise). The WELLXPLAIN dataset focuses on mental health categorization, classifying posts into four aspects: "physical," "intellectual and vocational," "social," and "spiritual and emotional" well-being. HAT-

[3]This approach ensures balanced class representation while managing computational resources, a practical constraint in production settings as well.

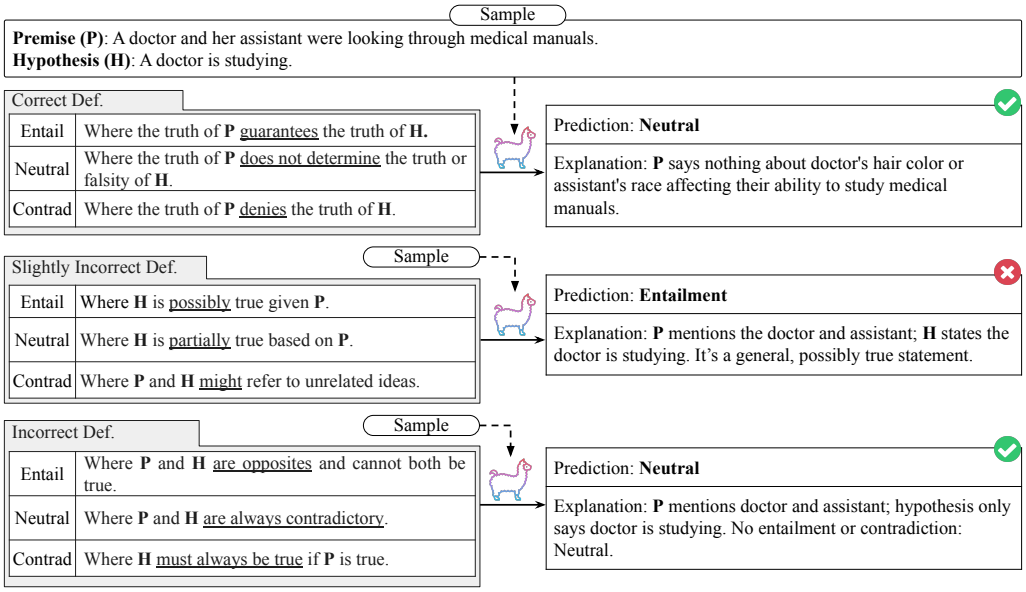

Figure 3: Label Definition Accuracy (Incorrect Vs. Slightly Incorrect Vs. Correct Definitions): As anticipated, when the model was provided with correct or slightly incorrect definitions, it produced the expected outputs: correct outputs for correct definitions and incorrect outputs for slightly incorrect definitions. However, when given incorrect definitions, the model generated correct outputs unexpectedly.

EXPLAIN provides text classification across three categories: "hatespeech," "normal," or "offensive" content, representing typical content moderation scenarios. The WICE dataset differs from the others as a fact-checking benchmark where models must identify which sentences from provided evidence support a given claim, requiring sentence-level explanation rather than just classification labels.

Statistical details of these datasets are presented in Table 7 (appendix A), with the standard label definitions we use for fixed definition experiments provided in appendices B to E. This diverse collection allows us to assess model behavior across varying tasks, from general logical reasoning to highly specialized domains where definition interpretation is particularly consequential.

4 Experiments, Results, and Analysis

4.1 LLM Receptivity: Results and Analysis

Building on our experimental datasets, we now examine how different models integrate and respond to label definitions across our designed scenarios. Our analysis focuses on how models balance external definitions with their internalized knowledge.

In the *definition permutation* scenario, Table 1 shows clear performance differences between aligned and misaligned label-definition mappings. Across all datasets, models achieve significantly higher MCC scores with correctly permuted definitions (*aligned*), with ***Mistral-7B* on HATEXPLAIN being the striking exception** where alignment produced minimal improvement—a surprising deviation from the otherwise consistent pattern. These findings align with the distributional hypothesis of language learning (Firth, 1957; Saxena et al., 2024), whereby LLMs learn word meanings through their contextual associations. When faced with conflicting cues—explicit definitions versus internalized patterns—models generally prioritize the explicit definitions provided in the prompt context, suggesting that instruction-following capabilities can override pre-trained associations.

Model	Permutation	e-SNLI	WELLXPLAIN	HATEXPLAIN	WICE
GPT-4	MCC_a	**0.83**	**0.35**	**0.42**	**0.32**
	$\overline{MCC}_m$	0.74	0.17	0.27	0.16
LLaMA-3	MCC_a	**0.42**	**0.20**	**0.21**	**0.13**
	$\overline{MCC}_m$	0.34	0.15	0.17	0.09
Phi-3	MCC_a	**0.24**	**0.32**	**0.20**	**0.05**
	$\overline{MCC}_m$	0.23	0.11	0.06	0.04
Mistral-7B	MCC_a	**0.51**	**0.10**	0.05	**0.05**
	$\overline{MCC}_m$	0.48	0.07	0.05	0.03

Table 1: MCC for aligned vs. misaligned definitions across e-SNLI, WELLXPLAIN, and HATEXPLAIN. Aligned definitions improve performance, while misaligned ones reduce MCC. Subscripts *a* and *m* denote aligned and misaligned label–definition mappings, respectively.

The *label definition accuracy* scenario, summarized in Table 2, reveals varying sensitivities to definition quality across models and tasks. For **e-SNLI**, *LLaMA-3* demonstrates remarkably strong responsiveness, with MCC values increasing more than

	e-SNLI			WELLXPLAIN			HATEXPLAIN			WICE		
Model	$MCC_{Inc.}$	$MCC_{SInc.}$	$MCC_{Cor.}$	$MCC_{Inc.}$	$MCC_{SInc.}$	$MCC_{Cor.}$	$MCC_{Inc.}$	$MCC_{SInc.}$	$MCC_{Cor.}$	$MCC_{Inc.}$	$MCC_{SInc.}$	$MCC_{Cor.}$
GPT-4	Meta-Response†	0.83	0.83	0.31	0.31	**0.38**	0.31	0.36	**0.44**	0.13	0.31	**0.32**
LLaMA-3	0.08	0.32	**0.36**	0.16	0.20	**0.21**	0.10	0.13	**0.15**	0.09	0.13	0.13
Phi-3	0.22	0.21	**0.23**	0.29	0.35	**0.38**	0.11	0.08	**0.14**	0.04	0.01	0.05
Mistral-7B	0.20	**0.48**	0.40	0.18	**0.20**	**0.20**	0.05	0.07	**0.09**	0.03	0.03	0.05

Table 2: Impact of Definition Quality: Incorrect, Slightly Incorrect, and Correct Definitions across the general domain dataset (e-SNLI) and domain-specific datasets (WELLXPLAIN and HATEXPLAIN). Results highlight distinct patterns in model performance, with *LLaMA-3* showing steady improvement from Incorrect to Correct, while *Mistral-7B* exhibits strong sensitivity to definition accuracy, particularly in domain-specific settings. Subscripts Inc., SInc., and Cor. denote Incorrect, Slightly Incorrect, and Correct, respectively. "Meta-Response†" reflects the model declining to answer due to conflicting definitions.

threefold when moving from *incorrect* to *slightly incorrect* definitions, and nearly fourfold when comparing *incorrect* to *correct* definitions. *Mistral-7B* shows substantial but less dramatic improvements (140% and 100% respectively), while ***Phi-3* exhibits a surprisingly minimal sensitivity** with near-negligible changes (-4.55% and +4.55%)—an unexpected resilience to definitional variations. Most unexpectedly, ***GPT-4* surprised us entirely** by detecting definitional inconsistencies and declining to provide predictions—a sophisticated metacognitive response none of the other models exhibited. This pattern reveals a key distinction between linguistic competence (internalized knowledge of language structures) and performance (how that knowledge is applied in specific contexts), with larger models demonstrating stronger metacognitive capabilities to detect contradictions between provided definitions and their pre-training.

Notably, *GPT-4* exhibits a unique *meta-response* behavior, choosing abstention when faced with conflicting label definitions. This refusal, absent in other models, suggests that alignment strategies such as Reinforcement Learning from Human Feedback (RLHF) encourage diagnostic caution rather than committing to potentially inconsistent outputs. We emphasize this as an important extension of the competence–performance distinction and highlight it as a promising direction for mechanistic follow-up work on safety-aligned LLMs.

For the **WELLXPLAIN** mental health categorization task, all models show moderate improvements with better definitions, though the sensitivity varies. *GPT-4* and *LLaMA-3* show similar improvement patterns (22.58% and 25% respectively) when moving from incorrect to more accurate definitions. *Phi-3* demonstrates greater definitional sensitivity, with performance improving by approximately one-fifth when using *slightly incorrect* definitions and by nearly one-third with *correct* definitions. **Contrary to expectations, *Mistral-7B* shows the least variation** (11.11%), suggesting it relies more heavily on internal knowledge for this domain—a surprising finding given its stronger responsiveness in the e-SNLI task. This aligns with theories of knowledge acquisition, suggesting that when LLMs lack extensive pre-training exposure to specialized domains, they become more dependent on explicit contextual definitions, similar to how humans learn novel concepts by relying on explicit instruction when they lack prior experience.

In the **HATEXPLAIN** content moderation task, definition quality impacts performance substantially across all models. *GPT-4*'s accuracy improves steadily as definitions become more precise (16.13% with *slightly incorrect* and 41.94% with *correct* definitions). *LLaMA-3* follows a similar pattern with more pronounced gains (30% and 50%). **Interestingly, both *Phi-3* and *Mistral-7B* show unexpectedly large improvements** when given correct definitions (75% and 80% respectively)—a stark contrast to their behavior on other tasks. This suggests a "knowledge integration threshold" where models shift from relying on internal representations to trusting external definitions when the latter are sufficiently precise and consistent with their training.

For the **WICE** fact-checking task, model performance under varying definition accuracies reveals minimal responsiveness across all models. Unlike other datasets, the MCC scores remain nearly flat, with negligible gains when moving from incorrect to correct definitions. *LLaMA-3* and *Mistral-7B* show marginal improvement (from 0.09 to 0.13 and from 0.03 to 0.05, respectively), while *Phi-3* even declines slightly from slightly incorrect to correct. These results suggest that **WICE, as an evidence-retrieval task, is less sensitive to semantic label definitions** and more reliant on factual pattern matching—highlighting a limitation in current LLMs' ability to incorporate abstract definitional guidance into decision-making for fact-based

inference tasks.

These findings collectively reveal a dynamic interplay between pre-trained knowledge and contextual definitions that varies by model architecture, task domain, and definition quality. **Contrary to the conventional wisdom that larger models are universally more capable**, smaller models generally show higher definition-sensitivity, particularly in specialized domains, suggesting they have weaker internal representations and greater reliance on explicit guidance. **The most surprising finding is that model responses to definitions are not consistent across tasks**—the same model may be highly sensitive to definitions in one domain while showing minimal responsiveness in another. These results have important implications for prompt engineering in specialized applications, suggesting that careful definition crafting is crucial for smaller models and domain-specific tasks, while larger models benefit from their ability to incorporate definitions while maintaining critical evaluation of their validity.

To further understand how models incorporate external definitions into their decision-making, we next examine their receptivity across various definition integration strategies.

4.2 Definitions Integration: Results and Analysis

Building on our previous analysis of how models respond to definition conflicts, we now examine how different integration strategies affect performance. Table 3 compares four conditions: (i) *vanilla*, (ii) *fixed definition*, (iii) *adjusted definition*, and (iv) *fixed definition + few-shot*.

For **e-SNLI**, **models perform best with the definition-free *vanilla* setting—a counterintuitive finding** that challenges conventional wisdom about explicit guidance. While *GPT-4* experiences modest degradation with definitions (6.98% and 3.49% decreases), *LLaMA-3* and *Mistral-7B* show more substantial performance drops (23.40% to 43.86%). This pattern aligns with the bayesian inference framework of in-context learning proposed by Xie et al. (2021), where models already possess robust internal representations of common linguistic tasks through pretraining and may experience interference when explicit definitions contradict these learned representations (Min et al., 2022). **Most surprisingly, adding few-shot examples alongside definitions further reduces performance across all models**, with *GPT-4* dropping by 18.60%—suggesting that exemplars can actually create confusion when combined with abstract definitions for general language tasks.

In contrast, for domain-specific tasks **WELLXPLAIN** and **HATEXPLAIN**, definitions generally enhance performance with some notable exceptions. *GPT-4* benefits modestly (34.15% in WELLXPLAIN, 9.09% in HATEXPLAIN), while **Mistral demonstrates an extraordinary definition dependence with a 110% improvement in WELLXPLAIN and a remarkable tenfold increase in HATEXPLAIN**. This dramatic improvement can be understood through the perspective of knowledge-base integration, where models with limited exposure to specialized domains during pretraining rely heavily on explicit definitions as surrogate knowledge, a phenomenon highlighted in prior work on definition-based supervision (Wang et al., 2021; Mishra et al., 2022; Mueller et al., 2022). *LLaMA-3* exhibits task-dependent patterns, with **a peculiar reversal where *fixed definitions* severely harm performance in WELLXPLAIN** (-44.83%) despite helping considerably in HATEXPLAIN (+64.29%). This inconsistency reflects what Reynolds and McDonell (Reynolds and McDonell, 2021) describe as the brittleness of in-context learning, where slight variations in definition format can produce dramatically different outcomes depending on how well they align with a model's pretraining experiences. **Perhaps most unexpected, the smaller *Phi-3* with *fixed definition+few-shot* achieves an MCC of 0.62 for *wellxplain*, outperforming all other models and conditions**—challenging assumptions about model scale advantages through what Rafailov et al. (2023) termed the effectiveness of direct preference optimization in smaller models.

The explanation quality tells a different story from classification accuracy. *Adjusted definition* dramatically improves Rouge scores in WELLXPLAIN, with ***LLaMA-3* and *Mistral-7B* showing unprecedented explanation improvements** (9× and 4.5×, respectively). Similarly, *fixed definition+few-shot* boosts explanation quality in HATEXPLAIN by **a staggering sixfold**. **Paradoxically, these dramatic improvements in explanation quality often fail to translate to higher classification accuracy**—revealing what Kosinski (2023) described as a distinction between a model's ability to reason about concept relationships (evident in explanations) versus its ability to apply those concepts in classification decisions, similar

		e-SNLI		WELLXPLAIN		HATEXPLAIN		WICE	
Model	**Scenario**	*MCC*	*Rouge*	*MCC*	*Rouge*	*MCC*	*Rouge*	*MCC*	*F1*
GPT-4	Vanilla	**0.86**	0.21	0.41	0.06	0.33	0.05	0.28	0.71
	Fixed Definition	0.80	0.20	0.46	0.05	**0.36**	0.04	0.32	**0.73**
	Adjusted Definition	0.83	0.20	**0.55**	0.07	0.33	0.06	**0.35**	0.69
	Fixed Definition+Few-shot	0.70	**0.26**	0.46	**0.10**	0.30	**0.31**	0.32	0.73
LLaMA-3	Vanilla	**0.47**	0.18	0.29	0.04	0.14	**0.31**	0.14	0.69
	Fixed Definition	0.35	0.15	0.16	0.04	**0.23**	0.18	0.13	0.69
	Adjusted Definition	0.36	0.14	**0.44**	**0.40**	0.14	0.10	0.14	0.69
	Fixed Definition+Few-shot	0.36	**0.20**	0.38	0.38	0.15	0.04	0.11	**0.74**
Phi-3	Vanilla	**0.48**	0.19	0.33	0.06	0.13	0.01	0.03	0.68
	Fixed Definition	0.38	0.12	0.38	0.06	**0.19**	0.02	0.05	0.68
	Adjusted Definition	0.46	0.15	0.49	**0.14**	0.17	0.01	**0.12**	0.67
	Fixed Definition+Few-shot	0.31	**0.22**	**0.62**	0.13	0.16	**0.10**	0.02	**0.69**
Mistral-7B	Vanilla	**0.57**	0.17	0.20	0.07	0.01	**0.47**	0.03	0.70
	Fixed Definition	0.32	0.20	0.20	0.07	0.12	0.35	**0.05**	0.69
	Adjusted Definition	0.34	0.19	**0.42**	**0.39**	**0.14**	0.01	0.04	0.69
	Fixed Definition+Few-shot	0.49	**0.24**	0.39	0.36	0.14	0.05	0.01	0.70

Table 3: When external guidance is actually useful: definitions make general NLI harder, but they are very helpful in specialized domains, and adjusted definitions also improve performance in *wice* for most models.

to the competence-performance distinction in human cognition (Chomsky, 1965).

This dissociation highlights a deeper competence–performance gap in LLMs: while definitions enhance the models' ability to articulate conceptual relationships in explanations (competence), they do not always shift the fast, categorical processes underlying prediction (performance). Our results suggest that explanation generation and classification rely on partially distinct subsystems, with external definitions influencing deliberative reasoning more readily than decision-making boundaries. Recognizing this gap is crucial for deploying LLMs in high-stakes applications where both accurate predictions and trustworthy explanations are required.

The fact-checking task (**WICE**) reveals a different pattern. While *GPT-4* shows modest benefits from definitions (25% MCC improvement with *adjusted definitions*), **all models demonstrate remarkably consistent F1 scores for explanation alignment regardless of definition condition**—a striking contrast to the variable impacts seen in other metrics and tasks. This unusual stability aligns with findings from Lin et al. (Lin et al., 2024), suggesting that for evidence identification tasks, which require factual recall rather than conceptual understanding, models rely primarily on distributional patterns learned during pretraining rather than explicit definitional guidance.

These findings extend our understanding beyond the earlier knowledge conflict experiments by demonstrating that not only does definition sensitivity vary by model and domain, but **the specific integration strategy can dramatically affect both classification performance and explanation quality—sometimes in opposing directions**. The phenomenon mirrors what Wei et al. (Wei et al., 2022) termed "emergent abilities" in language models, where certain capabilities appear only under specific combinations of model architecture, task domain, and prompt structure—a finding with significant implications for how these models should be deployed in real-world applications.

These findings highlight the varying degrees to which LLMs integrate external definitions, with performance trends differing across general and domain-specific tasks.

5 Related Work

Label definitions have shown promise in improving zero-shot classification (Peskine et al., 2023). While the work establishes the basic utility of definitions, our study delves deeper into how LLMs process and utilize these definitions, particularly in generating explanations. Through our analysis of definition conditions, we reveal crucial insights into whether LLMs truly incorporate external definitions or default to their pre-trained knowledge.

While recent advances in reasoning methods, including chain-of-thought (CoT) prompting (Wang et al., 2024; Zhao et al., 2024; Wei et al., 2023), Progressive-Hint Prompting (Zheng et al., 2024), and PromptWizard (Agarwal et al., 2024), have improved model performance, they cannot distinguish between genuine expert knowledge incorporation and superficial alignment (Greenblatt et al., 2024). Our work uniquely reveals how LLMs semantically process external definitions, offering insights into whether models truly incorporate or merely imitate expert knowledge.

Building on various prompt engineering techniques (Wu et al., 2024; Zhang et al., 2022; Yang

et al., 2024; Fernando et al., 2023) and explanation strategies (Lampinen et al., 2022), our work provides the first systematic analysis of how LLMs interpret and integrate semantic definitions. This is particularly crucial for high-stakes applications like mental health and hate speech detection, where eliciting the model's true behavior rather than superficial compliance is essential—an aspect previous methods haven't fully explored.

While prior studies have examined knowledge conflicts in LLMs, our work is distinct in four ways. First, we introduce a dual framing that jointly considers both knowledge conflict and definition integration, whereas most work studies only conflict. Second, we provide the first systematic cross-domain comparison of definition receptivity, showing stark contrasts between general and domain-specific tasks. Third, we propose an exhaustive permutation-based probing strategy that tests all possible label-definition mappings, unlike prior binary conflict settings. Finally, we uncover an explanation–classification disconnect, revealing that explanation quality does not necessarily correlate with classification accuracy, a phenomenon not previously explored in this context.

In summary, our work is unique in four respects: (1) We jointly analyze both knowledge conflict and definition integration, whereas prior work has largely focused on conflict alone; (2) we provide the first cross-domain comparison, revealing that definition receptivity differs substantially between general and specialized domains; (3) we introduce a permutation-based probing strategy that exhaustively tests all possible label–definition mappings, beyond the binary conflict settings previously studied; and (4) we uncover a novel explanation–classification disconnect, showing that gains in explanation quality do not necessarily translate to improved classification accuracy. Together, these contributions differentiate our work from existing literature and open new directions for understanding LLM receptivity to definitions.

6 Conclusion and Future Work

To evaluate LLMs' receptivity to label definitions, we study LLM across three domains—everyday language understanding, mental health analysis, and hate speech detection. We make the following findings: Firstly, LLMs perform significantly better when handling definitions in specialized tasks than general everyday tasks. This indicates that their ability to follow instructions improves with domain-specific contexts. Based on this, we recommend crafting precise and context-specific definitions for specialized tasks such as mental health and hate speech. Clear definitions tailored to the task can greatly enhance model performance.

Secondly, the quality of definitions, whether correct or flawed, profoundly affects LLM behavior and output accuracy. Poorly constructed definitions can severely degrade model performance. Third, definitions customized for specific scenarios outperform generic, one-size-fits-all approaches. This highlights the importance of aligning definitions with the nuances of the task at hand. To maximize effectiveness, we recommend using a mix of fixed label definitions and a few shot examples that are semantically proximal to user queries in order to adapt LLMs to the specific user context.

Finally, our observations hold true across various tasks ranging from general language understanding (e.g., entailment tasks) to specialized domains like mental health analysis and hate speech detection. This consistency underscores the versatility of LLMs but also reveals varying levels of task complexity. By adopting these strategies, domain experts and practitioners can unlock the full potential of LLMs across both general-purpose and specialized applications.

Our findings suggest actionable insights for deployment. On edge devices, smaller models (e.g., *Phi-3*, *Mistral-7B*) gain the most from carefully tailored definitions in specialized domains, offering lightweight performance boosts. In cloud settings, larger models (e.g., *GPT-4*) rely more on internal knowledge but often refuse outputs when definitions conflict, making them better suited for high-stakes applications where safety is critical. Even when definitions do not improve classification, they reliably enhance explanation quality, supporting user trust, transparency, and regulatory compliance.

A promising future direction is to apply mechanistic interpretability techniques to understand better how external definitions influence model behavior. Such analysis can reveal whether specific components of the network are responsible for encoding and applying definition semantics, shedding light on the internal mechanisms that support or override external guidance during the tasks.

An additional avenue for future work is to examine LLM behavior under *disputed but valid definitions*, where experts may legitimately disagree (e.g., on "fluency" vs. "readability").

Limitations

While our study provides strong evidence of the impact of label definitions on LLM performance, certain aspects warrant further exploration. Our analysis focuses on classification tasks, and while the results generalize well within this domain, additional studies in more complex reasoning tasks could provide deeper insights. Additionally, while our experiments rely on prompting strategies, future work could explore fine-tuning methods to further enhance definition adherence. Lastly, although we analyze definition accuracy and alignment, investigating finer-grained model behaviors through interpretability techniques could offer even more precise insights into how LLMs integrate external definitions. We did not employ attention-weight analysis, since prior work has shown that attention weights do not consistently reflect true model decision-making processes (Jain and Wallace, 2019; Wiegreffe and Pinter, 2019; Mohammadi et al., 2024; Naim and Asher, 2024). Nonetheless, we consider probing internal mechanisms, including attention, as valuable future directions.

Ethics Statement

This research investigates the extent to which Large Language Models (LLMs) adhere to explicit label definitions. Our study is conducted using publicly available datasets and does not involve human subjects or personally identifiable information. While our analysis includes scenarios where models are presented with incorrect definitions, this is done solely to evaluate their robustness and receptivity to external knowledge. We acknowledge that LLMs may exhibit biases due to their pre-trained knowledge, and we do not make normative claims about their outputs. Our findings aim to contribute to the broader understanding of model alignment and interpretability for domain experts without promoting misuse or adversarial manipulation of AI systems.

Acknowledgments

We thank Mohammad Eskandari for his valuable assistance in revising the figures presented in this paper. We also thank the anonymous reviewers for their constructive feedback and suggestions, which helped improve the quality of this work. We gratefully acknowledge support from the UMBC Cybersecurity Leadership – Exploratory Grant and the USISTEF Award. The opinions, conclusions, and recommendations expressed here are solely those of the authors and do not necessarily reflect the views of USISTEF, UMBC, or CrowdStrike.

This material is also based on research that is in part supported by the DARPA for the SciFy program under agreement number HR00112520301. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either express or implied, of DARPA or the U.S. Government.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Eshaan Agarwal, Joykirat Singh, Vivek Dani, Raghav Magazine, Tanuja Ganu, and Akshay Nambi. 2024. Promptwizard: Task-aware prompt optimization framework. *Preprint*, arXiv:2405.18369.

Amos Azaria and Tom Mitchell. 2023. The internal state of an llm knows when it's lying. *arXiv preprint arXiv:2304.13734*.

Misha Bilenko. 2024. Introducing Phi-3: Redefining what's possible with SLMs. Microsoft Azure Blog. [Accessed 16-02-2025].

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2022. Discovering latent knowledge in language models without supervision. *arXiv preprint arXiv:2212.03827*.

Oana-Maria Camburu, Tim Rockt"aschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 9539–9549. Curran Associates, Inc.

Noam Chomsky. 1965. *Aspects of the Theory of Syntax*. MIT Press, Cambridge, MA.

Shiling Deng, Serge Belongie, and Peter Ebert Christensen. 2025. Large vision-language models for knowledge-grounded data annotation of memes. *arXiv preprint arXiv:2501.13851*.

Hanyu Duan, Yi Yang, and Kar Yan Tam. 2024. Do llms know about hallucination? an empirical investigation of llm's hidden states. *arXiv preprint arXiv:2402.09733*.

Chrisantha Fernando, Dylan Banarse, Henryk Michalewski, Simon Osindero, and Tim Rocktäschel. 2023. Promptbreeder: Self-referential self-improvement via prompt evolution. *Preprint*, arXiv:2309.16797.

J. R. Firth. 1957. A synopsis of linguistic theory, 1930-1955. In *Studies in Linguistic Analysis*, pages 1–32. Blackwell, Oxford. Special volume of the Philological Society.

Muskan Garg. 2024. Wellxplain: Wellness concept extraction and classification in reddit posts for mental health analysis. *Knowledge-Based Systems*, 284:111228.

Manas Gaur, Kalpa Gunaratna, Shreyansh Bhatt, and Amit Sheth. 2022. Knowledge-infused learning: A sweet spot in neuro-symbolic ai. *IEEE Internet Computing*, 26(4):5–11.

Ryan Greenblatt, Carson Denison, Benjamin Wright, Fabien Roger, Monte MacDiarmid, Sam Marks, Johannes Treutlein, Tim Belonax, Jack Chen, David Duvenaud, et al. 2024. Alignment faking in large language models. *arXiv preprint arXiv:2412.14093*.

Sarthak Jain and Byron C Wallace. 2019. Attention is not explanation. *arXiv preprint arXiv:1902.10186*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Ryo Kamoi, Tanya Goyal, Juan Diego Rodriguez, and Greg Durrett. 2023. Wice: Real-world entailment for claims in wikipedia. *arXiv preprint arXiv:2303.01432*.

Michal Kosinski. 2023. Theory of mind may have spontaneously emerged in large language models. *Nature Human Behaviour*, 7(5):755–767.

Sachin Kumar, Chan Young Park, and Yulia Tsvetkov. 2023. Gen-z: Generative zero-shot text classification with contextualized label descriptions. *arXiv preprint arXiv:2311.07115*.

Andrew Lampinen, Ishita Dasgupta, Stephanie Chan, Kory Mathewson, Mh Tessler, Antonia Creswell, James McClelland, Jane Wang, and Felix Hill. 2022. Can language models learn from explanations in context? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 537–563, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jerry Lin, Roy Fox, Isa Dalai, Aditya Kusupati, Sewon Min, Luke Zettlemoyer, Hannaneh Hajishirzi, Jon Shlens, Sham Kakade, and Hossein Adeli. 2024. In-context learning creates task vectors. *arXiv preprint arXiv:2402.16509*.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection.

Brian W Matthews. 1975. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2):442–451.

Meta. 2024. Introducing Meta Llama 3: The most capable openly available LLM to date — ai.meta.com. https://ai.meta.com/blog/meta-llama-3/?utm_source=chatgpt.com. [Accessed 16-02-2025].

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. Cross-task generalization via natural language instructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3470–3487, Dublin, Ireland. Association for Computational Linguistics.

Seyedali Mohammadi, Edward Raff, Jinendra Malekar, Vedant Palit, Francis Ferraro, and Manas Gaur. 2024. WellDunn: On the robustness and explainability of language models and large language models in identifying wellness dimensions. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 364–388, Miami, Florida, US. Association for Computational Linguistics.

Yida Mu, Ben P. Wu, William Thorne, Ambrose Robinson, Nikolaos Aletras, Carolina Scarton, Kalina Bontcheva, and Xingyi Song. 2024. Navigating prompt complexity for zero-shot classification: A study of large language models in computational social science. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12074–12086, Torino, Italia. ELRA and ICCL.

Aaron Mueller, Jason Krone, Salvatore Romeo, Saab Mansour, Elman Mansimov, Yi Zhang, and Dan Roth. 2022. Label semantic aware pre-training for few-shot text classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8318–8334, Dublin, Ireland. Association for Computational Linguistics.

Omar Naim and Nicholas Asher. 2024. On explaining with attention matrices. *arXiv preprint arXiv:2410.18541*.

Youri Peskine, Damir Korenčić, Ivan Grubisic, Paolo Papotti, Raphael Troncy, and Paolo Rosso. 2023. Definitions matter: Guiding gpt for multi-label classification. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4054–4063.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*.

Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. *CHI Extended Abstracts*.

Yash Saxena, Deepa Tilwani, Ali Mohammadi, Edward Raff, Amit Sheth, Srinivasan Parthasarathy, and Manas Gaur. 2024. Attribution in scientific literature: New benchmark and methods. *arXiv preprint arXiv:2405.02228*.

Amit Sheth, Manas Gaur, Kaushik Roy, and Keyur Faldu. 2021. Knowledge-intensive language understanding for explainable ai. *IEEE Internet Computing*, 25(5):19–24.

Sinong Wang, Han Fang, Madian Khabsa, Hanzi Mao, and Hao Ma. 2021. Entailment as few-shot learner. *arXiv preprint arXiv:2104.14690*.

Zecheng Wang, Chunshan Li, Zhao Yang, Qingbin Liu, Yanchao Hao, Xi Chen, Dianhui Chu, and Dianbo Sui. 2024. Analyzing chain-of-thought prompting in black-box large language models via estimated V-information. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 893–903, Torino, Italia. ELRA and ICCL.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. *Transactions on Machine Learning Research*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. *arXiv preprint arXiv:1908.04626*.

Zhaoxuan Wu, Xiaoqiang Lin, Zhongxiang Dai, Wenyang Hu, Yao Shu, See-Kiong Ng, Patrick Jaillet, and Bryan Kian Hsiang Low. 2024. Prompt optimization with ease? efficient ordering-aware automated selection of exemplars. *Preprint*, arXiv:2405.16122.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2023. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. *arXiv preprint arXiv:2305.13300*.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2021. An explanation of in-context learning as implicit bayesian inference. *arXiv preprint arXiv:2111.02080*.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V. Le, Denny Zhou, and Xinyun Chen. 2024. Large language models as optimizers. *Preprint*, arXiv:2309.03409.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2022. Automatic chain of thought prompting in large language models. *Preprint*, arXiv:2210.03493.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2024. Explainability for large language models: A survey. *ACM Trans. Intell. Syst. Technol.*, 15(2).

Chuanyang Zheng, Zhengying Liu, Enze Xie, Zhenguo Li, and Yu Li. 2024. Progressive-hint prompting improves reasoning in large language models. *Preprint*, arXiv:2304.09797.

A Additional tables

Premise	Hypothesis	Label
A doctor checks on his medical equipment.	A doctor preparing for work.	Entailment
A child is given a medical check up by a doctor.	A child with doctor.	Entailment
A doctor checks on his medical equipment.	A doctor getting ready for a patient.	Neutral
A child is given a medical check up by a doctor.	A sick child with doctor.	Neutral
A male doctor looking at a female patient's hand.	The doctor is on a date.	Contradiction
A doctor in blue scrubs performing surgery.	A doctor is eating lunch.	Contradiction

Table 4: Few-shot samples used to generate adjusted definitions (see Appendix F and Figure 2c).

Model	Version	# parameters
GPT-4	gpt-4-0613	-
LLaMA-3	meta-LLaMA/Meta-LLaMA-3-8B-Instruct	8B
Phi-3	microsoft/Phi-3-mini-4k-instruct	3.82B
Mistral-7B	mistralai/Mistral-7B-Instruct-v0.1	7.24B

Table 5: Models' detail used in experimental setup.

B Definitions of labels in e-SNLI

- **Entailment (0):** if the premise entails the hypothesis.
- **Neutral (1):** if neither entailment nor contradiction hold.
- **Contradiction (2):** if the hypothesis contradicts the premise.

Template Prompt
Prompt: Read the following definitions for neutral (1), entailment (0), and contradiction (2) and come up with a label and explanation for the given promise-hypothesis pair: *Label Definitions*: {definitions (fixed or adjusted) here} *Premise*: {premise here} *Hypothesis*: {hypothesis here}
Response: *Predicted Label*: {your label here} *Explanation*: {your explanation here}

Table 6: Template Prompt Used for *Mistral-7B* in the Definition Scenario, for e-SNLI. For other scenario prompts, refer to the corresponding code.

Dataset	Total	Train	Valid	Test
e-SNLI	569,051	549,367	9,824	9,842
WELLXPLAIN	3,092	-	-	-
HATEXPLAIN	19,229	15,383	1,922	1,924
WICE(claim)	1967	1260	349	358

Table 7: Statistical summary of e-SNLI, WELLXPLAIN, HATEXPLAIN, and WICE datasets.

C Definitions of labels in WELLXPLAIN

- **Physical Aspect (PA)**: Physical wellness fosters healthy dietary practices while discouraging harmful behaviors like tobacco use, drug misuse, and excessive alcohol consumption. Achieving optimal physical wellness involves regular physical activity, sufficient sleep, vitality, enthusiasm, and beneficial eating habits. Body shaming can negatively affect physical well-being by increasing awareness of medical history and appearance issues.
- **Intellectual Aspect (IA)**: Utilizing intellectual and cultural activities, both inside and outside the classroom, and leveraging human and learning resources enhance the wellness of an individual by nurturing intellectual growth and stimulation.
- **Vocational Aspect (VA)**: The Vocational Dimension acknowledges the role of personal gratification and enrichment derived from one's occupation in shaping life satisfaction. It influences an individual's perspective on creative problem-solving, professional development, and the management of financial obligations.
- **Social Aspect (SA)**: The Social Dimension highlights the interplay between society and the natural environment, increasing individuals' awareness of their role in society and their impact on ecosystems. Social bonds enhance interpersonal traits, enabling a better understanding and appreciation of cultural influences.
- **Spiritual Aspect (SpA)**: The Spiritual Dimension involves seeking the meaning and purpose of human life, appreciating its vastness and natural forces, and achieving harmony within oneself.
- **Emotional Aspect (EA)**: The Emotional Dimension enhances self-awareness and positivity, promoting better emotional control, realistic self-appraisal, independence, and effective stress management.

D Definitions of labels in HATEXPLAIN

- **Hatespeech:** is used to describe posts that express hatred towards a specific group or individual based on attributes like race, religion, gender, or sexual orientation. It is defined in the context of online content that devalues or discriminates against minority members and can lead to increased prejudice and real-world violence against these groups.
- **Normal:** Posts labeled as normal are those that do not contain any hate speech or offensive content. These posts are considered benign and do not target any individual or group with harmful or abusive language.
- **Offensive:** The term "offensive" is used to describe language that is abusive or insulting but does not necessarily meet the criteria of hate speech. Offensive speech includes harmful or derogatory language that can hurt individuals or groups but lacks the broader discriminatory intent characteristic of hate speech.

E Definitions of labels in WICE

- **Supported:** Entire claim is backed by evidence.
- **Partially supported:** Some parts are supported, others not.
- **Not supported:** No part is supported.

F Adjusted label definitions for sample in Figure 2(c)

- **Entailment:** is the relationship between a premise and a hypothesis where the truth of the premise logically guarantees the truth of the hypothesis, often indicating a necessary condition or a step in a process.
- **Neutral:** refers to a statement or situation that lacks a clear positive or negative connotation, implication, or emotional tone, often indicating a lack of direct or explicit relationship between the premises and the hypothesis.
- **Contradiction:** A contradiction is a statement that cannot be true or valid because it involves two or more mutually exclusive or incompatible circumstances, actions, or facts.