# Do We Know What LLMs Don't Know? A Study of Consistency in Knowledge Probing

**Raoyuan Zhao, Abdullatif Köksal, Ali Modarressi, Michael A. Hedderich and Hinrich Schütze**

LMU Munich and Munich Center for Machine Learning (MCML)
{rzhao,akoksal,amodaresi,hedderich}@cis.lmu.de

## Abstract

The reliability of large language models (LLMs) is greatly compromised by their tendency to hallucinate, underscoring the need for precise identification of knowledge gaps within LLMs. Various methods for probing such gaps exist, ranging from calibration-based to prompting-based methods. To evaluate these probing methods, in this paper, we propose a new process based on using input variations and quantitative metrics. Through this, we expose two dimensions of inconsistency in knowledge gap probing. (1) **Intra-method inconsistency:** Minimal non-semantic perturbations in prompts lead to considerable variance in detected knowledge gaps within the same probing method; e.g., the simple variation of shuffling answer options can cause an agreement as low as 40%. (2) **Cross-method inconsistency:** Probing methods contradict each other on whether a model knows the answer. Methods are highly inconsistent – with decision consistency across methods being as low as 7% – even though the model, dataset, and prompt are all the same. These findings challenge existing probing methods and highlight the urgent need for perturbation-robust probing frameworks of knowledge gaps.

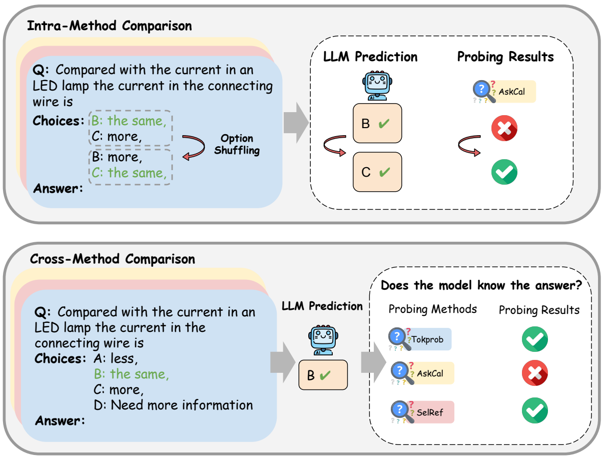


Figure 1: Examples of the two major dimensions of inconsistency in knowledge gap probing that we focus on in this paper. In the **intra-method comparison** (top), the same probing method gives contradictory assessments (certain vs not certain) for the same LLM and the same question with options shuffled, revealing internal inconsistency. In the **cross-method comparison** (bottom), different probing methods applied to the same LLM yield conflicting judgments – two probing methods maintain that the LLM is certain about the answer, while the third does not. These results illustrate that knowledge gap detection in LLMs can be unreliable and sensitive to method choice and prompt perturbation.

## 1 Introduction

While large language models (LLMs) are increasingly applied across diverse NLP tasks, understanding the limits of their knowledge remains a core challenge – particularly in mitigating hallucinations (Wang et al., 2023c), where models produce fluent yet factually incorrect outputs (Ji et al., 2023a; Maynez et al., 2020; Tam et al., 2023; Ji et al., 2023b). This has led to increasing interest in identifying *knowledge gaps* – a situation where the model lacks the necessary knowledge to answer a question, meaning it either does not know or is uncertain about the correct answer.

To address this, a growing body of work proposes probing methods that aim to act as detection tools for LLMs' knowledge gaps (Wang et al., 2023c), based on various signals such as prompting (Feng et al., 2023), self-consistency (Mündler et al., 2024; Feng et al., 2024b), token probabilities (Guo et al., 2017; Jiang et al., 2021) and calibrated hidden representations (Slobodkin et al., 2023; Azaria and Mitchell, 2023).

These knowledge probing methods are increasingly used to infer whether a model "knows" the answer to a question. However, an underexplored issue is the consistency and reliability of the probing methods themselves: Based on their predic-



*Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 23254–23280
November 4-9, 2025 

tions, how reliable are these probing methods and do we actually know what LLMs don't know?

To answer this question, we present a systematic study of consistency within and between probing methods. In practical applications, prompt variations such as typos or slight changes in word order are common. While they can influence LLM outputs (Salinas and Morstatter, 2024; Pezeshkpour and Hruschka, 2024; Hedderich et al., 2025), the underlying knowledge gaps should remain unchanged, and probing methods should be robust to such slight perturbations.

To evaluate whether *probing methods are reliable*, we conduct a systematic evaluation of consistency within and across popular probing methods. Specifically, we propose a framework with two comparison dimensions (as illustrated in Figure 1):

(i) **Intra-method consistency** – whether a method yields stable predictions under surface-level prompt perturbations (e.g., typos, answer reordering);

(ii) **Cross-method consistency** – whether different probing methods agree when applied to the same model and input.

We design four distinct prompting variants to systematically evaluate different probing methods, LLMs and datasets, and we propose new consistency metrics for the two diagnostic axes (intra-method and cross-method) to quantify the consistency. Note that we do not evaluate whether LLMs are consistent, but whether the probing methods that evaluate LLM behavior are consistent and whether we can thus trust their assessment.

Our work reveals a paradox: These tools themselves suffer from alarming inconsistencies, casting doubt on the validity of their predictions. We identify four main findings:

- Minimal prompt perturbations, such as introducing typos, reduce the consistency metric with the original prompt down to around 39%, revealing hypersensitivity to surface-level variations.
- Even when moving from zero-shot to few-shot prompting to guide the model, we still observe inconsistencies in the detected knowledge gaps – reaching down to 4%.
- The scaling rule (that bigger models are less inconsistent) does not always hold for the consistency of the probing methods. We observe that probing consistency of some methods on a 70B model is even lower than on the 1B or 3B models.
- All the probing methods exhibit large inconsistencies, both within individual methods and across different methods. The lowest observed cross-method consistency has a decision consistency of just 7% and intra-method consistency reaches a minimum of just 2%. We publicly release our evaluation code.[1]

## 2 Related Work

### 2.1 Knowledge Probing Methods

Knowledge probing methods have been proposed to extract stable signals that quantify model certainty and diagnose potential knowledge gaps (Petroni et al., 2019; Youssef et al., 2023). Several works have focused on identifying internal signals that reflect a model's certainty about the given answer, including token probabilities, response consistency, self-reported confidence scores (Kadavath et al., 2022).

To make use of these signals, researchers have developed several strategies which Feng et al. (2024b) broadly categorize into four categories: calibration-based methods align model confidence with empirical accuracy to set abstention thresholds (Sun et al., 2022; Kuhn et al., 2023); training-based methods fine-tune models or probe internal representations to estimate answer veracity (Cobbe et al., 2021); prompting-based approaches guide models to assess uncertainty through reflective or information-seeking prompts (Si et al., 2023; Wang et al., 2023a); and self-consistency methods generate multiple reasoning paths to assess stability and reliability in outputs (Feng et al., 2024b; Wang et al., 2023b; Miao et al., 2023).

More recently, efforts have moved beyond single-model paradigms, leveraging multi-LLM collaboration (Feng et al., 2024a,b) and interpretability techniques such as analyzing activation patterns (Arditi et al., 2024; Wang et al., 2024) and tracing neuron-level circuits (Yao et al., 2024). These methods aim to identify reliable indicators of model certainty and genuine knowledge, which are critical for downstream tasks such as hallucination detection (Manakul et al., 2023; Chen et al., 2023), refusal strategies (Cao, 2024; Xu et al., 2024; Zhang et al., 2024) and honesty evaluation (Chern et al., 2024; Li et al., 2024; Yang et al., 2023).

[1]https://github.com/raoyuanzhao/Probing_Uncertainty

We select probing methods from each of Feng et al. (2024b)'s categories for our evaluation:

**Calibration.** These methods define a confidence score obtained from the model and optimize a threshold to minimize misclassifications between correct and incorrect examples. Token Probability (**TokProb**) (Feng et al., 2024b) measures the model's confidence based on the output probability of the response. Ask for Calibration (**AskCal**) (Tian et al., 2023) prompts the model to output a confidence score for the response.

**Training.** In these methods, an additional classifier is trained to predict whether the model "knows" or "doesn't know" the answer to a given question. Embedding Training (**Embedding**) (Slobodkin et al., 2023; Azaria and Mitchell, 2023) involves training the classifier on the hidden states of the LLM in conjunction with the model's prediction.

**Prompting.** These methods utilize post-response prompting, where the model reassesses its previous response. Self-Reflect (**SelfRef**) (Kadavath et al., 2022) first predicts the answer for a given question and then reassesses whether this response is correct or not. If the model deems its response incorrect, it is assumed there is a knowledge gap. Similarly, in More Information (**MoreInfo**) (Liu et al., 2023), the model is asked if it requires more information. If the model responds affirmatively, it is assumed that it does not know the answer to the question.

**Consistency.** "None-of-the-Above" (**NOTA**) is an approach where an additional "NOTA" option is appended (Feng et al., 2024b). If the model selects this option, it indicates a knowledge gap.

### 2.2 Prompt Sensitivity in Language Models

Prompting has become a central interface for interacting with LLMs (Brown et al., 2020), yet accumulating evidence shows that model outputs are often highly sensitive to minor variations in prompts (Stureborg et al., 2024; Pezeshkpour and Hruschka, 2024; Errica et al., 2025; Salinas and Morstatter, 2024). This sensitivity undermines the reliability of language models in both evaluation and real-world deployment.

Sclar et al. (2024) systematically explore this issue, demonstrating that LLM performance can vary by over 70% across semantically equivalent prompts. Zhuo et al. (2024) introduce PromptSensiScore, a decoding-confidence-based measure of prompt sensitivity across tasks and datasets. Their findings reveal that larger models tend to be more robust, but even high-capacity models exhibit notable instability in complex reasoning settings.

Chatterjee et al. (2024) propose the POSIX index to evaluate the change in model log-likelihoods under paraphrastic rewrites of prompts. Their analysis highlights that instruction tuning and parameter scaling alone are insufficient to mitigate prompt sensitivity; however, few-shot prompting offers some robustness gains.

These studies highlight how fragile LLM behavior can be under minor prompt changes. This observation raises concerns not just for general prompting, but also for structured probing methods that aim to detect knowledge gaps. While prior work on knowledge gap detection has proposed various probing methods to evaluate LLM knowledge, these approaches typically assume fixed prompts and do not account for sensitivity to prompt perturbations. In our work, we examine whether prompt-dependent probing methods (as well as those based on other principles) exhibit similar inconsistencies. We also evaluate if current probes are up to this challenge and whether model scaling and few-shot prompts can help probing stability.

## 3 Consistency Evaluation Methods

Since knowledge probing methods aim to extract what a model knows (or does not know), their decisions should be consistent. First, applying the same method to semantically equivalent prompts (e.g., adding a minor typo) should yield consistent results, which we call **intra-method consistency**. Second, different methods applied to the same model should produce aligned results, avoiding contradictions. We call this **cross-method consistency**.

For the intra-method consistency, we design semantically equivalent prompts. Our zero-shot variants simulate real-world noise: inserting spaces, shuffling options and minor typos. Our one-shot variants help the model better understand the answer format. They are simple questions that do not introduce new knowledge and are assumed to have no effect on the model's knowledge gaps. See Appendix A.3 for details on these variants.

Our consistency comparisons are always between two setups that either involve the same method with original vs variant prompts (intra-method) or the same prompt applied across different methods (cross-method).

| Method | Variant | $\text{IoU}_{\text{cons}}$ | $\text{IoU}_{\text{acc}}$ | $\text{IoU}_{\text{rej}}$ | DecCons | Agr. | $\text{IoU}_{\text{cons}}$ | $\text{IoU}_{\text{acc}}$ | $\text{IoU}_{\text{rej}}$ | DecCons | Agr. | $\text{IoU}_{\text{cons}}$ | $\text{IoU}_{\text{acc}}$ | $\text{IoU}_{\text{rej}}$ | DecCons | Agr. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **Mistral-7B** | | | | | **LLaMa-3.1-8B** | | | | | **Olmo-2-7B** | | | | |
| TOKPROB | Space | .74 | .87 | .64 | .89 | **.99** | .64 | .94 | .49 | .94 | **.94** | **.69** | **.77** | **.63** | **.85** | **.96** |
| | Options | .40 | .72 | .28 | .75 | .66 | .59 | .93 | .44 | .93 | .74 | .56 | .67 | .49 | .75 | .10 |
| | Typo | .67 | .83 | .55 | .86 | .97 | .62 | .93 | .46 | .94 | .91 | **.69** | .76 | **.63** | .83 | .95 |
| | One-shot | **.97** | **.99** | **.95** | **.99** | .68 | **.69** | **.96** | **.68** | **.97** | .67 | .62 | .66 | .58 | .77 | .96 |
| ASKCAL | Space | **.76** | **.77** | **.76** | **.87** | **.94** | **.52** | **.79** | **.42** | **.81** | **.93** | **.64** | .69 | **.60** | **.79** | **.87** |
| | Options | .61 | .61 | .61 | .76 | .73 | .31 | .72 | .20 | .74 | .76 | .62 | **.70** | .55 | .78 | .13 |
| | Typo | **.76** | .75 | **.76** | .86 | .93 | .51 | .79 | .42 | .81 | .91 | .62 | .68 | .58 | .78 | .85 |
| | One-shot | .41 | .41 | .47 | .63 | .80 | .33 | .54 | .27 | .64 | .69 | .45 | .48 | .43 | .63 | .86 |
| EMBEDDING | Space | .58 | .49 | **.76** | .80 | **.95** | .50 | .70 | .40 | .75 | **.94** | **.61** | **.54** | .70 | **.78** | .85 |
| | Options | **.60** | **.49** | **.76** | **.81** | .69 | **.66** | **.84** | **.55** | **.86** | .76 | .36 | .34 | .49 | .61 | .13 |
| | Typo | .58 | .48 | .75 | .80 | .92 | .56 | .71 | .46 | .77 | .91 | **.61** | **.54** | **.71** | **.78** | **.88** |
| | One-shot | .33 | .37 | .38 | .56 | .69 | .39 | .49 | .32 | .59 | .70 | .44 | .43 | .48 | .63 | .85 |
| NOTA | Space | **.40** | .92 | **.25** | **.93** | **.90** | .36 | .90 | **.23** | .91 | **.92** | **.32** | **.91** | **.20** | **.91** | **.81** |
| | Options | .39 | **.93** | .25 | **.93** | .57 | **.39** | **.91** | .25 | **.91** | .70 | .27 | **.91** | .16 | **.91** | .22 |
| | Typo | .39 | .92 | .25 | .92 | .88 | .36 | .90 | .23 | .90 | .90 | .29 | **.91** | .17 | **.91** | .80 |
| | One-shot | .26 | .92 | .16 | .92 | .63 | .22 | .86 | .12 | .86 | .70 | .23 | .87 | .13 | .87 | .76 |
| MOREINFO | Space | **.74** | **.91** | **.62** | **.92** | **.88** | **.86** | **.98** | **.77** | **.98** | **.92** | .45 | **.77** | .32 | **.79** | **.77** |
| | Options | .62 | .88 | .47 | .89 | .55 | .79 | .97 | .67 | .97 | .72 | .41 | .74 | .29 | .76 | .26 |
| | Typo | .72 | .91 | .60 | .92 | .85 | .80 | .97 | .68 | .97 | .89 | **.46** | .76 | **.33** | **.79** | .76 |
| | One-shot | .04 | .79 | .02 | .79 | .64 | .09 | .93 | .05 | .93 | .71 | .36 | .64 | .25 | .68 | .65 |
| SELFREF | Space | **.67** | **.67** | **.67** | **.80** | **.92** | **.66** | **.66** | **.66** | **.79** | **.96** | **.49** | **.37** | **.72** | **.75** | **.87** |
| | Options | .46 | .46 | .46 | .63 | .53 | .52 | .53 | .52 | .68 | .82 | .36 | .24 | .69 | .71 | .18 |
| | Typo | .67 | .67 | .67 | .80 | .91 | .62 | .62 | .62 | .76 | .95 | .47 | .35 | .71 | **.75** | .84 |
| | One-shot | .49 | .51 | .48 | .66 | .77 | .40 | .35 | .49 | .59 | .71 | .31 | .21 | .62 | .65 | .75 |

Table 1: Intra-method consistency evaluation using six knowledge probing methods in MMLU. Best results in **bold** and the worst underlined. **$\text{IoU}_{\text{cons}}$** columns are highlighted with light yellow background as this is our main metric for consistency. We introduce four different variants, each evaluated over independent runs with different random seeds and one-shot prompt examples, and the reported values represent their mean. The variance is generally close to zero: see Appendix C for more detailed data.

As illustrated in Figure 1, there are two types of pairs:

- **Intra-Method Comparison Pair**, where the same probing method is applied to different prompt variants. (e.g., Case 1: ASKCAL Method + Original Prompt; Case 2: ASKCAL Method + Prompt with Shuffled Options)
- **Cross-Method Comparison Pair**, where different probing methods are applied to the same prompt. (e.g., Case 1: ASKCAL Method + Original Prompt; Case 2: TOKPROB Method + Original Prompt)

For any two cases 1 and 2 in the same pair, we define $R_1$ and $R_2$ as the sets of questions where the probe identifies a knowledge gap, i.e., the model should abstain from answering (**rejection**). Similarly, $A_1$ and $A_2$ represent the sets of questions where the probe claims that the models know the answer (**acceptance**).

Based on this notation, we propose four metrics:

**Acceptance/Rejection Consistency Intersection over Union ($\text{IoU}_{\text{acc}}$/$\text{IoU}_{\text{rej}}$)** is defined as the ratio of the intersection (the number of common accepted/rejected questions) to the union (total distinct accepted/rejected questions):

$$\text{IoU}_{\text{acc}} = \frac{|A_1 \cap A_2|}{|A_1 \cup A_2|}, \quad \text{IoU}_{\text{rej}} = \frac{|R_1 \cap R_2|}{|R_1 \cup R_2|}$$

Higher values indicate greater consistency in acceptance/rejection decisions.

**Harmonic Consistency IoU ($\text{IoU}_{\text{cons}}$)** We use the harmonic mean of the previous two metrics to achieve a balanced measure between the rejection and acceptance metrics. We use $\text{IoU}_{\text{cons}}$ as our main metric for intra-method evaluation.

**Decision Consistency (DecCons)** quantifies the proportion of questions consistently accepted or rejected across setups:

$$\text{DecCons} = \frac{|(A_1 \cap A_2) \cup (R_1 \cap R_2)|}{|A_1 \cup A_2 \cup R_1 \cup R_2|}$$

It is more lenient than $\text{IoU}_{\text{cons}}$, which approaches zero in cross-method setups in our experiments as it fails under extreme accept-all or reject-all behaviors, whereas DecCons counts both consistent acceptances and rejections as agreement. Thus, we use this metric as the primary indicator for cross-method analysis.

**Agreement (Agr.)** is the proportion of commonly accepted questions for which the model (not the probe) provides the same answer in both setups. This metric evaluates the stability of the model's answers.

$$\text{Agr.} = \frac{\sum_{x \in A_1 \cap A_2} \mathbb{1}(\text{Answer}_1(x) = \text{Answer}_2(x))}{|A_1 \cap A_2|}$$

## 4 Experimental Setup

We select a range of instruction-tuned models to apply different knowledge probing methods, including Mistral-7B (Jiang et al., 2023), LLaMA-3.2-1B-Instruct, LLaMA-3.2-3B-Instruct, LLaMA-3.1-8B-Instruct, LLaMA-3.1-70B (Dubey et al., 2024), and OLMo-2-7B-Instruct (OLMo et al., 2024). Our selection includes models from different developers and covers a range of sizes to explore how model capacity may influence the robustness and stability of knowledge probing methods. In particular, we include four models from the LLaMA-3 family – 1B, 3B, 8B and 70B – to systematically examine whether increasing model size leads to more consistent probing behavior under prompt variations.

For our probing datasets, we adopt MMLU, a benchmark designed to test knowledge and reasoning across diverse academic topics (Hendrycks et al., 2021), and Hellaswag, a commonsense inference dataset focused on everyday scenarios (Zellers et al., 2019). For each dataset, we randomly sample 1,000 examples to construct a development set and another 1,000 examples as the test set. The development set is used for threshold calibration and probing method tuning where applicable, while the test set is held out for evaluation. See Appendix A.1 for full details.

## 5 Results and Analysis

### 5.1 Intra-method Consistency

In this section, we investigate the intra-method consistency results of various probing methods on MMLU and HellaSwag. Since the results show similar patterns of extreme inconsistency, we focus on MMLU in the main text. See Table 8 in Appendix for results on HellaSwag.

**Impact of Zero-Shot Variants** All zero-shot variants (Space, Shuffle Options, Typo) affect consistency, with $\text{IoU}_{\text{cons}}$ values ranging from as low as 0.27 to a maximum of 0.86. Inserting spaces has the least overall impact on consistency. Shuffling options has the greatest impact even though shuffling does not change the semantic meaning of a question in any way. This sensitivity to shuffling in the probes is similar to the sensitivity of LLMs (Pezeshkpour and Hruschka, 2024). EMBEDDING for Mistral and Llama 8B maintains the highest consistency under option variations among methods. Its comparatively good performance may stem from its reliance on semantic patterns in hidden representations. Nevertheless, its intra-method consistency is still poor: $\text{IoU}_{\text{cons}}$ is only 0.6.

**Impact of One-Shot Variant** The impact of one-shot prompting is even greater than that of the three zero-shot variants, with $\text{IoU}_{\text{cons}}$ ranging from 0.04 to 0.97.

The impact is particularly evident for MOREINFO. In the MMLU dataset, the Mistral and Llama 8B models have $\text{IoU}_{\text{cons}}$ scores of 0.04 and 0.09 (compared to 0.74 and 0.86 for the space variant). With one-shot prompting, the abstain rate drops (see Table 13), likely because MOREINFO follows the simple one-shot pattern, where the need for more information is indicated as "No." This might encourage the model to respond similarly, even for uncertain questions. Although the pattern is less pronounced, the probing methods also exhibit reduced consistency under the one-shot variant on Olmo, compared to other variants. This suggests that the model's response consistency is more sensitive to changes in input structure than to minor formatting perturbations.

Future work could investigate whether more complex or varied one-shot examples alter this effect, shedding light on whether the phenomenon is inherent to few-shot prompting or tied to example design.

**Inconsistency is Consistent** The variance of $\text{IoU}_{\text{cons}}$ after introducing three variants with three random seeds and four one-shot examples is close to 0; see Appendix C for details. This suggests that the inconsistency of the probing methods is not due to randomness in the selection of one-shot examples or in the locations where perturbations are introduced.

**Source of Inconsistency** Table 1 shows that some methods achieve high $\text{IoU}_{\text{acc}}$ or $\text{IoU}_{\text{rej}}$ along with strong DecCons, yet exhibit low overall $\text{IoU}_{\text{cons}}$. This is primarily due to extreme rejection rates. Additionally, for TOKPROB, which achieves 0.97 in $\text{IoU}_{\text{cons}}$, Agr. is only 0.68, indicating that even though detected knowledge gaps are consistent, there is great variability in the model's answer to the same question. This may be the source of some inconsistency, as methods involving threshold finding or training rely on surface-level response matching in the training set to infer knowledge gaps. However, unstable predictions undermine the reliability of these methods.

### 5.2 Cross-method Consistency

Due to highly divergent rejection rates by the probes, cross-method consistency is much lower

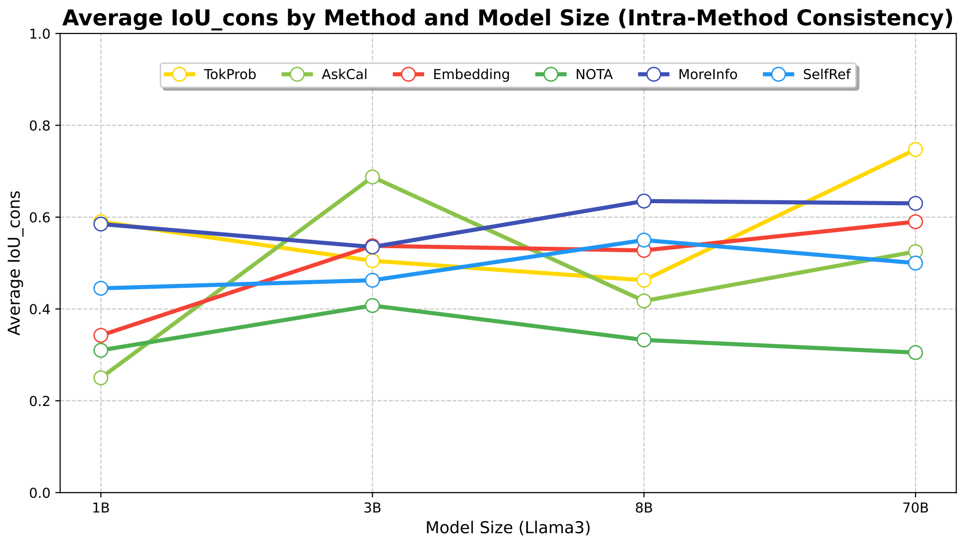


Figure 2: Average $\text{IoU}_{\text{cons}}$ across different model sizes (LLaMA3) for intra-method consistency of each method. The scaling trend does not consistently hold across all probing methods, see Table 6 in Appendix for more details.

than intra-method consistency: $\text{IoU}_{\text{cons}}$ values for cross-method combinations are near zero (full results in Appendix B). This disparity motivates our adoption of the DecCons metric (visualized through heatmaps in Figure 3) for the evaluation of cross-method consistency.

**(In)Consistency is Model- and Dataset-Specific** As can be seen in Figures 3a and 3b, for the same model on the same dataset with different variants, DecCons is similar. However, the metric differs across different models and datasets (Figure 3c). For example, with Mistral on dataset MMLU, NOTA and ASKCAL achieve a DecCons of 0.54, whereas in Mistral+HellaSwag, the same methods drop to 0.07. This stark contrast further highlights the instability of these methods across different datasets, suggesting the reliability of these probing methods depends on the dataset and model.

**Methods Using Similar Signals Exhibit Higher Consistency** EMBEDDING is less consistent with other methods (in Mistral+Hellaswag, DecCons with MOREINFO is 0.07). This may be because EMBEDDING utilizes deeper-level model outputs (signals) than other methods, specifically leveraging the model's hidden states. NOTA and MOREINFO share the highest consistency across all setups, with DecCons between 0.62 and 0.89. This may be due to the underlying similar principles the methods share, suggesting they utilize a correlated signal.

| Method | Variant | Abstain F1 |
|---|---|---|
| TokProb | Original | .47 |
| | Zero-shot | .47 |
| | One-shot | .41 |
| AskCal | Original | .65 |
| | Zero-shot | .64 |
| | One-shot | .56 |
| Embedding | Original | .64 |
| | Zero-shot | **.68** |
| | One-shot | .45 |
| MoreInfo | Original | .24 |
| | Zero-shot | **.25** |
| | One-shot | .02 |
| NOTA | Original | .16 |
| | Zero-shot | .14 |
| | One-shot | .09 |
| Reflect | Original | .50 |
| | Zero-shot | .50 |
| | One-shot | .48 |

Table 2: Evaluation of probing methods on the Mistral + MMLU setting, using a metric proposed by (Feng et al., 2024b). Zero-shot variants (space, shuffled option, typo) do not substantially reduce Abstain F1 and sometimes even **improve** it, which suggests that current metrics may not reliably reflect probing method robustness. Full results across all model-dataset combinations are provided in the Appendix D.

### 5.3 Scaling Rules for Probing Consistency

LLMs become less sensitive and robust to input variations as their scale increases (Zhuo et al., 2024). If sampling-based probing methods were robust tools for detecting knowledge gaps, their intra-method and cross-method consistency, when

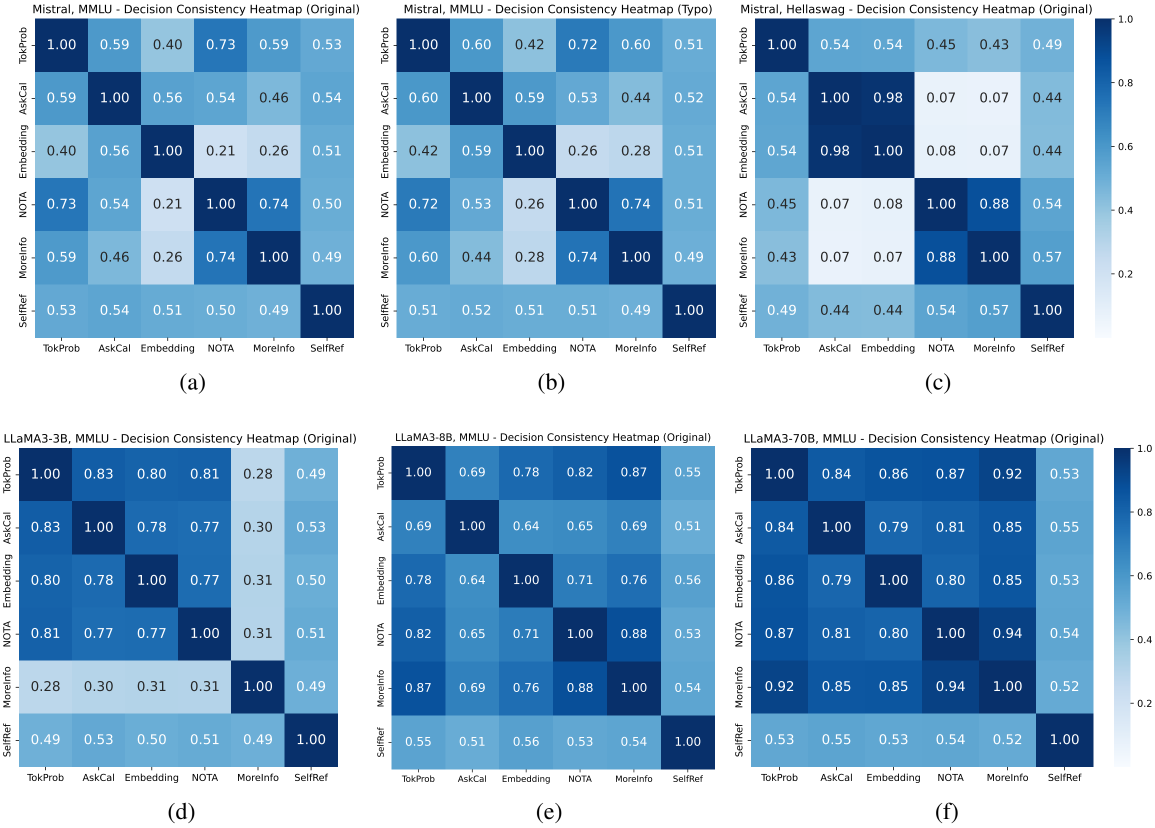


Figure 3: Heatmaps of cross-method consistency evaluation results (DecCons) under the original prompt across different datasets and model sizes. Subfigures (a)–(c) show results for Mistral models on MMLU and HellaSwag under different perturbation types. Subfigures (d)–(f) present results for LLaMA-3 models (3B, 8B, 70B) on MMLU under original prompt. See Appendix B for $\text{IoU}_{\text{cons}}$, LLaMA results and variant prompts.

applied to increasingly larger models, should also improve with scale. However, Figure 2 shows that this is not always the case. While some methods, such as TOKPROB, show a slight upward trend in consistency as model size increases, others remain flat or even decline. For example, NOTA reaches its peak consistency at 3B model and performs worse on the 70B model. Methods like EMBEDDING and ASKCAL also display inconsistent trends across different scales.

Moreover, the scaling rule does not consistently hold for cross-method consistency either. In Figure 3e and 3f, SELFREF exhibits uniformly low agreement with other methods across all model sizes, with DecCons values remaining around 0.5. On the 70B model, the agreement between SELFREF and both TOKPROB and MOREINFO is similar to, or even lower than, that on the 8B model. This indicates that increasing model size does not necessarily lead to greater convergence across different probing methods. The observed inconsistency should be attributed to the knowledge probing methods themselves, rather than to the underlying models.

### 5.4 Variant Influence on Probing Performance Metrics

Existing work commonly evaluates knowledge probing methods using metrics such as Abstain F1, which captures how well a method identifies knowledge gaps (Feng et al., 2024b). Abstain F1 is defined as the harmonic mean of precision and recall over abstention decisions, where precision reflects the proportion of predicted knowledge gaps that are correct, and recall reflects the proportion of true knowledge gaps that are successfully identified (Feng et al., 2024b; Whitehead et al., 2022).

But are these metrics sufficient to evaluate the consistency of probing methods under prompt perturbations?

To investigate this, we compare the performance of several probing methods using Abstain F1 across both original prompts and their perturbed variants. As before, these variants include common zero-shot modifications such as inserting extra spaces, shuffling multiple-choice options, and adding typos. As shown in Table 2, the Abstain F1 scores remain largely stable. For example, ASKCAL achieves 0.65 on the original prompt and 0.64 on a zero-shot variant. Similarly, REFLECT remains virtually unchanged with scores of 0.50, 0.50, and 0.48 across variants.

At first glance, these results suggest that current probing methods are robust to minor prompt changes. However, this interpretation overlooks a key discrepancy: while overall Abstain F1 scores appear stable, the actual rejection decisions vary considerably across prompts. For instance, REFLECT's Abstain F1 changes only slightly, but the $\text{IoU}_{\text{cons}}$ in shuffling option variants is just 46% (see Table 1), indicating that many of the specific questions being rejected differ.

The inconsistency becomes more striking under the one-shot setting. Although one-shot prompting is often considered to stabilize LLM outputs (Chatterjee et al., 2024), calibration-based methods like ASKCAL actually suffer a noticeable performance drop – from 0.65 to 0.56 in Abstain F1. This suggests that the instability is not due to the model itself, but rather the probe's failure to reliably capture the model's underlying uncertainty.

These findings reveal a key limitation of current evaluation practices. Metrics such as Abstain F1 emphasize aggregate correctness while failing to assess the consistency of rejection behavior across prompts. This indicates that the underlying knowledge gaps exposed by the probe differ across prompts, even when surface-level performance appears stable. Such discrepancies are invisible to established metrics, which suggests that these are not a good measure of probing reliability and highlights the need for using the metrics we propose in this work.

### 5.5 Threshold Influence Consistency

In Table 3, we observe that the probing methods exhibited poor intra-method consistency in its ASKCAL method on the HellaSwag dataset (with only 0.05 $\text{IoU}_{\text{cons}}$ in Options variant). This inconsistency could be attributed to the threshold selection process in calibration-based probing methods. These methods typically involve two steps: First, they use a validation set to compare a knowledge-probing signal (such as token probability) to actual accuracy (i.e., whether the model knows the answer or not). Then, they select the best threshold to determine which values indicate that the model does not know the answer (below the threshold) and which indicate that it does (above the threshold).

| Variant | $\text{IoU}_{\text{cons}}$ | $\text{IoU}_{\text{acc}}$ | $\text{IoU}_{\text{rej}}$ | DecCons | Agr. |
|---|---|---|---|---|---|
| **ASKCAL (w/o threshold correction)** | | | | | |
| Space | .24 | .17 | .87 | .87 | .89 |
| Options | .05 | .03 | .79 | .79 | .39 |
| Typo | .13 | .08 | .87 | .87 | 1.0 |
| One-shot | .09 | .05 | .93 | .93 | .77 |
| **ASKCAL (with threshold correction)** | | | | | |
| Space | .53 (+.29) | .45 (+.28) | .66 (-.20) | .73 (-.14) | .85 (-.04) |
| Options | .48 (+.43) | .35 (+.32) | .77 (-.03) | .79 (-.00) | .43 (+.04) |
| Typo | .41 (+.27) | .37 (+.29) | .47 (-.40) | .58 (-.29) | .82 (-.18) |
| One-shot | .28 (+.19) | .17 (+.12) | .79 (-.15) | .80 (-.17) | .65 (-.12) |

Table 3: Intra-method consistency analysis of the ASKCAL method on the HellaSwag dataset, with and without threshold correction. Without correction, the model's threshold values were highly unstable, leading to near-zero $\text{IoU}_{\text{cons}}$ scores across variants. Applying a fixed-threshold safeguard (set to 0.5) greatly improved consistency ($\text{IoU}_{\text{cons}}$), demonstrating that the correction mitigates the sensitivity to poorly calibrated thresholds.

However, during our experiments, we observed that existing threshold selection algorithms can yield suboptimal values. For instance, some thresholds were as high as 0.98 (leading the model to reject nearly all questions) while others were as low as 0.01 (effectively accepting everything). To address this issue, we introduced a threshold correction rule as a safeguard: when an unreasonable threshold is detected, we override it and set the threshold to 0.5.

After applying this correction, we observed a notable improvement in the intra-method consistency of the ASKCAL variants. As shown in Table 3, the $\text{IoU}_{\text{cons}}$ scores increased across all variants, demonstrating that the threshold correction notably mitigated the instability caused by poor threshold calibration.

## 6 Conclusion

In this study, we explore the consistency of four types of knowledge probing methods based on different principles. Our results reveal a high level of inconsistency, both intra-method and cross-method.

This variability suggests that a more robust approach is needed to reliably detect knowledge gaps across different models and datasets. Current refusal mechanisms often rely heavily on the output of the probing methods to decide whether a model "knows" an answer and should refuse to answer uncertain questions. However, if these probing signals are themselves unstable or inconsistent across variants and architectures, then the rejection behavior becomes inherently unreliable. This undermines the interpretability and trustworthiness of abstention-based frameworks.

We recommend that future work on knowledge probing explicitly consider the consistency of probing methods and routinely report consistency metrics such as those proposed in this paper. Improving the reliability of these methods is essential for building systems that can reliably assess the knowledge captured by language models.

## Limitations

While this study provides insights into the inconsistency of knowledge probing methods, the following limitations should be acknowledged:

**Limited to Multiple-Choice Question Datasets** In order to simplify the probing and evaluation to better compare it with previous work, we focused only on multiple-choice datasets. But additional insights might be obtained from open-ended text generation tasks.

**Scope of Probing Methods** Although we evaluate six existing knowledge probing methods and show inconsistency for all of them, the list of tested probes is not exhaustive. Expanding the scope of methods may provide an even more nuanced understanding of knowledge gap detection.

**Lack of Reasoning-Oriented Probing** Our study primarily evaluates probing methods that operate on direct model outputs, such as token probabilities or calibration-based responses. These methods are not naturally compatible with multi-step reasoning processes like chain-of-thought prompting. As a result, we do not assess whether explicit reasoning could improve consistency. Incorporating reasoning-oriented probes may require adapting or redesigning the probing framework, which we leave for future work.

## Acknowledgments

This research was supported by the Deutsche Forschungsgemeinschaft DFG (grant SCHU 2246/14-1). We thank the members of MaiNLP for their valuable feedback on this project, especially Yupei Du, Beiduo Chen, Robert Litschko, Silvia Casola, Yang Janet Liu, and Andreas Säuberli. Figure 1 includes icons designed by contributors to Flaticon, which we gratefully acknowledge.

## References

Andy Arditi, Oscar Balcells Obeso, Aaquib Syed, Daniel Paleka, Nina Rimsky, Wes Gurnee, and Neel Nanda. 2024. Refusal in language models is mediated by a single direction. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

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

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Lang Cao. 2024. Learn to refuse: Making large language models more controllable and reliable through knowledge scope limitation and refusal mechanism. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3628–3646, Miami, Florida, USA. Association for Computational Linguistics.

Anwoy Chatterjee, H S V N S Kowndinya Renduchintala, Sumit Bhatia, and Tanmoy Chakraborty. 2024. POSIX: A prompt sensitivity index for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14550–14565, Miami, Florida, USA. Association for Computational Linguistics.

Yuyan Chen, Qiang Fu, Yichen Yuan, Zhihao Wen, Ge Fan, Dayiheng Liu, Dongmei Zhang, Zhixu Li, and Yanghua Xiao. 2023. Hallucination detection: Robustly discerning reliable answers in large language models. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 245–255.

Steffi Chern, Zhulin Hu, Yuqing Yang, Ethan Chern, Yuan Guo, Jiahe Jin, Binjie Wang, and Pengfei Liu. 2024. Behonest: Benchmarking honesty in large language models. *arXiv preprint arXiv:2406.13261*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Federico Errica, Davide Sanvito, Giuseppe Siracusano, and Roberto Bifulco. 2025. What did I do wrong? quantifying LLMs' sensitivity and consistency to prompt engineering. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1543–1558, Albuquerque, New Mexico. Association for Computational Linguistics.

Shangbin Feng, Weijia Shi, Yuyang Bai, Vidhisha Balachandran, Tianxing He, and Yulia Tsvetkov. 2023. Knowledge card: Filling llms' knowledge gaps with plug-in specialized language models. In *The Twelfth International Conference on Learning Representations*.

Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Orevaoghene Ahia, Shuyue Stella Li, Vidhisha Balachandran, Sunayana Sitaram, and Yulia Tsvetkov. 2024a. Teaching LLMs to abstain across languages via multilingual feedback. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4125–4150, Miami, Florida, USA. Association for Computational Linguistics.

Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. 2024b. Don't hallucinate, abstain: Identifying LLM knowledge gaps via multi-LLM collaboration. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14664–14690, Bangkok, Thailand. Association for Computational Linguistics.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

Michael A. Hedderich, Anyi Wang, Raoyuan Zhao, Florian Eichin, Jonas Fischer, and Barbara Plank. 2025. What's the difference? supporting users in identifying the effects of prompt and model changes through token patterns. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20093–20123, Vienna, Austria. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023a. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023b. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38.

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*.

Haoming Jiang et al. 2021. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. 2022. Language models (mostly) know what they know. *Preprint*, arXiv:2207.05221.

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*.

Siheng Li, Cheng Yang, Taiqiang Wu, Chufan Shi, Yuji Zhang, Xinyu Zhu, Zesen Cheng, Deng Cai, Mo Yu, Lemao Liu, Jie Zhou, Yujiu Yang, Ngai Wong, Xixin Wu, and Wai Lam. 2024. A survey on the honesty of large language models. *Trans. Mach. Learn. Res.*, 2025.

Jing Liu et al. 2023. Knowledge card: Filling llms' knowledge gaps with plug-in specialized language models. In *Proceedings of the 2023 Annual Meeting of the Association for Computational Linguistics (ACL)*.

Potsawee Manakul, Adian Liusie, and Mark 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*, pages 9004–9017, Singapore. Association for Computational Linguistics.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Ning Miao, Yee Whye Teh, and Tom Rainforth. 2023. Selfcheck: Using llms to zero-shot check their own step-by-step reasoning. *arXiv preprint arXiv:2308.00436*.

Niels Mündler, Jingxuan He, Slobodan Jenko, and Martin Vechev. 2024. Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation. In *The Twelfth International Conference on Learning Representations*.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Michal Guerquin, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James V. Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2024. 2 olmo 2 furious.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Pouya Pezeshkpour and Estevam Hruschka. 2024. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2006–2017, Mexico City, Mexico. Association for Computational Linguistics.

Abel Salinas and Fred Morstatter. 2024. The butterfly effect of altering prompts: How small changes and jailbreaks affect large language model performance. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4629–4651, Bangkok, Thailand. Association for Computational Linguistics.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *The Twelfth International Conference on Learning Representations*.

Chenglei Si, Weijia Shi, Chen Zhao, Luke Zettlemoyer, and Jordan Boyd-Graber. 2023. Getting MoRE out of mixture of language model reasoning experts. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8234–8249, Singapore. Association for Computational Linguistics.

Aviv Slobodkin, Omer Goldman, Avi Caciularu, Ido Dagan, and Shauli Ravfogel. 2023. The curious case of hallucinatory (un)answerability: Finding truths in the hidden states of over-confident large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3607–3625, Singapore. Association for Computational Linguistics.

Rickard Stureborg, Dimitris Alikaniotis, and Yoshi Suhara. 2024. Large language models are inconsistent and biased evaluators. *arXiv preprint arXiv:2405.01724*.

Meiqi Sun, Wilson Yan, Pieter Abbeel, and Igor Mordatch. 2022. Quantifying uncertainty in foundation models via ensembles. In *NeurIPS 2022 Workshop on Robustness in Sequence Modeling*.

Derek Tam, Anisha Mascarenhas, Shiyue Zhang, Sarah Kwan, Mohit Bansal, and Colin Raffel. 2023. Evaluating the factual consistency of large language models through news summarization. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5220–5255, Toronto, Canada. Association for Computational Linguistics.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442, Singapore. Association for Computational Linguistics.

Boshi Wang, Xiang Yue, and Huan Sun. 2023a. Can chatgpt defend its belief in truth? evaluating llm reasoning via debate. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11865–11881.

Xinpeng Wang, Chengzhi Hu, Paul Röttger, and Barbara Plank. 2024. Surgical, cheap, and flexible: Mitigating false refusal in language models via single vector ablation. *arXiv preprint arXiv:2410.03415*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023b. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Zezhong Wang, Luyao Ye, Hongru Wang, Wai-Chung Kwan, David Ho, and Kam-Fai Wong. 2023c. ReadPrompt: A readable prompting method for reliable knowledge probing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7468–7479, Singapore. Association for Computational Linguistics.

Spencer Whitehead, Suzanne Petryk, Vedaad Shakib, Joseph Gonzalez, Trevor Darrell, Anna Rohrbach, and Marcus Rohrbach. 2022. Reliable visual question answering: Abstain rather than answer incorrectly. In *European Conference on Computer Vision*, pages 148–166. Springer.

Hongshen Xu, Zichen Zhu, Situo Zhang, Da Ma, Shuai Fan, Lu Chen, and Kai Yu. 2024. Rejection improves reliability: Training llms to refuse unknown questions using rl from knowledge feedback. *arXiv preprint arXiv:2403.18349*.

Yuqing Yang, Ethan Chern, Xipeng Qiu, Graham Neubig, and Pengfei Liu. 2023. Alignment for honesty. *ArXiv*, abs/2312.07000.

Yunzhi Yao, Ningyu Zhang, Zekun Xi, Mengru Wang, Ziwen Xu, Shumin Deng, and Huajun Chen. 2024. Knowledge circuits in pretrained transformers. *CoRR*, abs/2405.17969.

Paul Youssef, Osman Koraş, Meijie Li, Jörg Schlötterer, 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*, pages 15588–15605, Singapore. Association for Computational Linguistics.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Hanning Zhang, Shizhe Diao, Yong Lin, Yi Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. 2024. R-tuning: Instructing large language models to say 'I don't know'. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7113–7139, Mexico City, Mexico. Association for Computational Linguistics.

Jingming Zhuo, Songyang Zhang, Xinyu Fang, Haodong Duan, Dahua Lin, and Kai Chen. 2024. ProSA: Assessing and understanding the prompt sensitivity of LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1950–1976, Miami, Florida, USA. Association for Computational Linguistics.

## A Experimental Setups

The experiments for the six probing methods were run on one H200 140G. Temperatures for both LLaMa3 and Mistral settings were 0.1 with top_k = 0.9, top_k = 50. We checked the licenses of all the models and datasets used, as well as the code, which are publicly available resources.

### A.1 Data

We randomly sampled 1,000 data points from the validation set and 1,000 data points from the test set separately, then applied both zero-shot and one-shot prompting techniques to comprehensively evaluate the consistency of these methods.

### A.2 Zero-shot Variants

We used three different random seeds (4, 44, 99) to introduce variations into the original prompt (i.e., multiple-choice questions). For the shuffling options variant, we ensured that the correct answer's option order was always changed. For the typo variant, a randomly selected non-numeric word in the question had a letter added, deleted, or swapped. In the blank space insertion variant, we ensured that numeric values remained unchanged to minimize semantic disruption.

### A.3 One-shot Variants

Table 4 presents the one-shot prompts that have been used in our experiments. We selected well-known facts such as 2+2=4 to avoid introducing new information to the model, focusing instead on providing the model with the prompt structure.

Since the previously mentioned knowledge probing methods, such as AskCal, MoreInfo, and SelfRef, involve having the model first provide an answer to the question and then immediately follow up with the probability of correctness or whether the more information is needed, specific prompt design is required when applying the one-shot prompt. This is essential to ensure the model can handle these follow-up questions effectively and consistently. Table 5 outlines the prompt designs for these methods, showing how the questions are structured to guide the model through answering and then evaluating its response.

## B Cross-method Results

The huge difference in rejection rates results in poor $\mathrm{IOU}_{\mathrm{cons}}$ values for cross-method consistency, and the rejection rates for each method with different variants can be seen in Table 11,12,13,14.

Figures 6 and 8 present heatmaps of cross-method consistency using $\mathrm{IoU}_{\mathrm{cons}}$ as the metric, comparing original and variant-introduced conditions. Figures 4 and Figure 5 display complete heatmaps based on DecCons. The heatmaps demonstrate similar patterns when using the same dataset and model, but exhibit substantial variations when either factor is altered. This further highlights the inherent instability of these probing methods.

## C Intra-method Results

We have provided an additional metric here for reference:

**Common Accept Accuracy** calculates the average accuracy on questions that were commonly accepted, which can reflect the accuracy of the

| Index | Prompt Examples |
|---|---|
| | **MMLU** |
| 0 | Question: Who sings 'Here Comes the Sun'?<br>Choices: A: Led Zeppelin, B: Queen, C: Pink Floyd, D: The Beatles<br>Answer: D |
| 1 | Question: What is 2+2?<br>Choices: A: 3, B: 4, C: 5, D: 6<br>Answer: B |
| 2 | Question: What is the capital of France?<br>Choices: A: Berlin, B: Madrid, C: Paris, D: Rome<br>Answer: C |
| 3 | Question: What is the chemical symbol for water?<br>Choices: A: H2O, B: CO2, C: NaCl, D: O2<br>Answer: A |
| | **HellaSwag** |
| 0 | Question: When the lights went out during the storm, they<br>Choices: A: started watching a movie. B: lit some candles. C: opened the refrigerator. D: went swimming in the river<br>Answer: B |
| 1 | Question: After the baby started crying, the mother<br>Choices: A: picked up the baby to comfort it. B: paint the ceiling with a toothbrush. C: whispered to the toaster. D: opened an umbrella indoors<br>Answer: A |
| 2 | Question: As the sun set over the horizon, the sky turned<br>Choices: A: white. B: completely green. C: a mix of orange and pink. D: into a checkerboard pattern<br>Answer: C |
| 3 | Question: When the doorbell rang, I went to the door and<br>Choices: A: closed the windows. B: started cooking dinner. C: went to bed. D: opened it to see who was there<br>Answer: D |

Table 4: Information about the one-shot-prompt examples for MMLU and HellaSwag. All questions are quite simple.

problems that the model is certain about and the true capability of the model.

$$\cap\text{Accuracy} = \frac{\sum_{i \in A_1 \cap A_2} \frac{(\text{Correct}_1(i) + \text{Correct}_2(i))}{2}}{|A_1 \cap A_2|}$$

Tables 7 and 8 present the intra-method results on the MMLU and Hellaswag datasets. While different variants negatively impact consistency, their variance remains small, further indicating that the inconsistency is not due to randomness.

## D Probing Performance Comparison

In Table 11,12,13,14, we report the abstain performance for both zero-shot and one-shot variant prompting. The metrics applied are the same as those defined and used by Feng et al. (2024b). The results from the zero-shot setup are similar to those reported by Feng et al. (2024b), and the numbers further indicate that the general abstaining mechanism is minimally impacted by the inclusion of variants.

| Method | 1-shot Prompt(Input) |
|---|---|
| **TokProb** | The question is:<br>What is the chemical symbol for water?<br>Choices: A:H2O, B:CO2, C:NACl D:O2<br>Choose one answer from the above choices. Guess: A<br>The question is:<br>{question}<br>{choices}<br>Choose one answer from the above choices. Guess: {response} |
| **AskCal** | The question is:<br>What is the chemical symbol for water?<br>Choices: A: H2O, B: CO2, C: NaCl, D: O2<br>Choose one answer from the above choices. Guess: A<br>Provide the probability that your guess is correct.<br>Give ONLY the probability, no other words or explanation.For example:<br>Probability: <the probability between 0.0 and 1.0 that your guess is correct, without any extra commentary whatsoever; just the probability!><br>Probability: 1.0<br>The question is:<br>{question}<br>{choices}<br>Choose one answer from the above choices. Guess: {response1}<br>Provide the probability that your guess is correct.<br>Give ONLY the probability, no other words or explanation.For example:<br>Probability: <the probability between 0.0 and 1.0 that your guess is correct, without any extra commentary whatsoever; just the probability!><br>Probability: {response2} |
| **SelfRef** | The question is:<br>What is the chemical symbol for water?<br>Choices: A: H2O, B: CO2, C: NaCl, D: O2<br>Choose one answer from the above choices. Guess: A<br>The question is:<br>{question}<br>{choices}<br>Choose one answer from the above choices. Guess: {response1}<br>The above answer is:<br>A. True<br>B. False<br>The answer is {response2} |
| **NOTA** | The question is:<br>What is the chemical symbol for water?<br>Choices: A:H2O, B:CO2, C:NACl D:O2 E: None of the above<br>Choose one answer from the above choices. Guess: A<br>The question is:<br>{question}<br>{choices}<br>Choose one answer from the above choices. Guess: {response} |
| **MoreInfo** | The question is:<br>What is the chemical symbol for water?<br>Choices: A: H2O, B: CO2, C: NaCl, D: O2<br>Choose one answer from the above choices. Guess: A<br>Do you need more information to answer this question? (Yes or No)No<br>The question is:<br>{question}<br>{choices}<br>Choose one answer from the above choices. Guess: {response1}<br>Do you need more information to answer this question? (Yes or No){response2} |

Table 5: Example of one-shot prompt inputs across different methods. This table illustrates the design of input prompts for various methods, including TokProb, AskCal, SelfRef, NOTA, and MoreInfo. Each method presents the same base question, but with tailored instructions to reflect the specific goal of each method, such as asking for a guess, a probability estimate, or additional information.

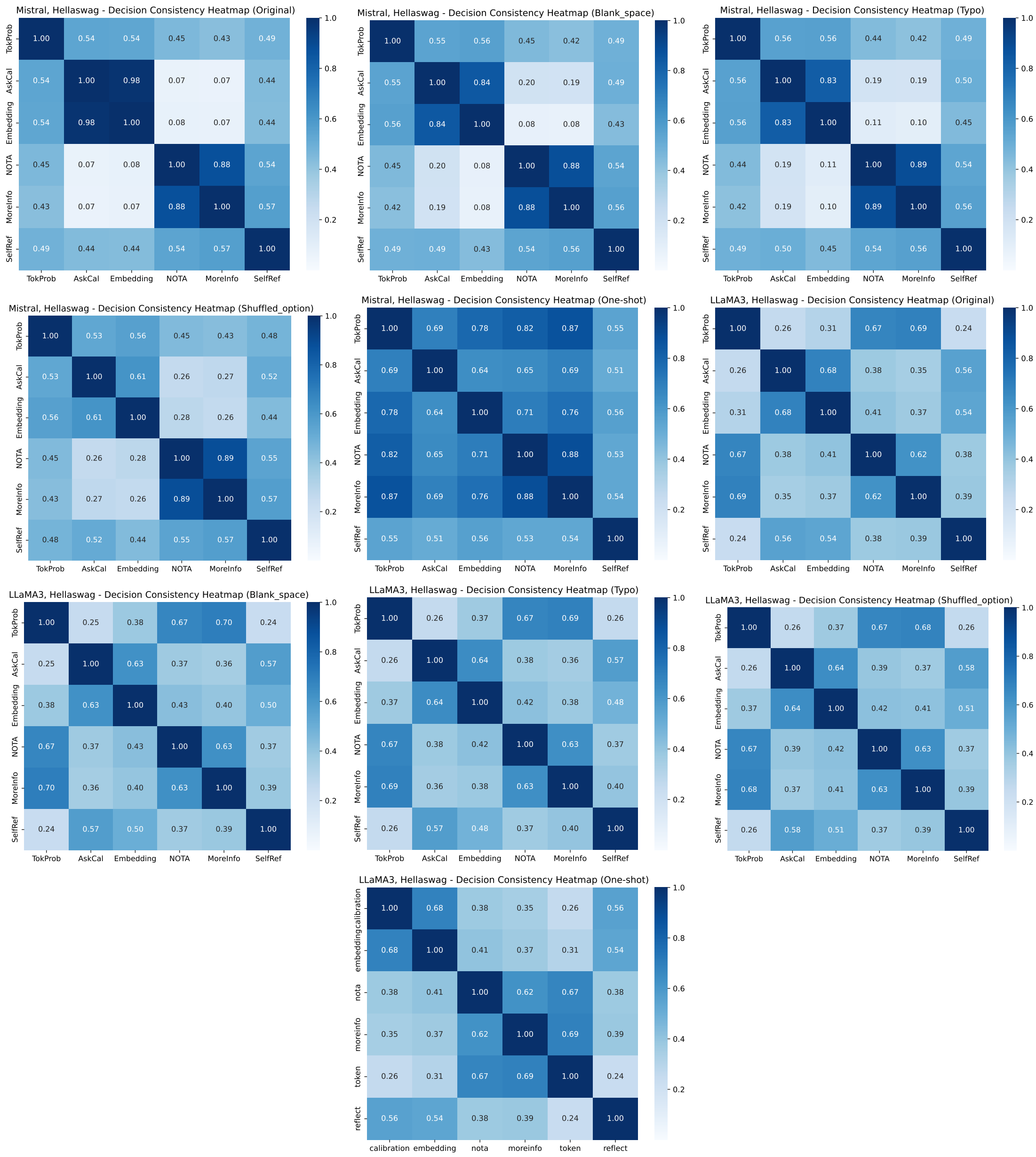

Figure 4: Heatmap of cross-method consistency evaluation results for Hellaswag. The values represent the average consistency across three different random seeds setups.

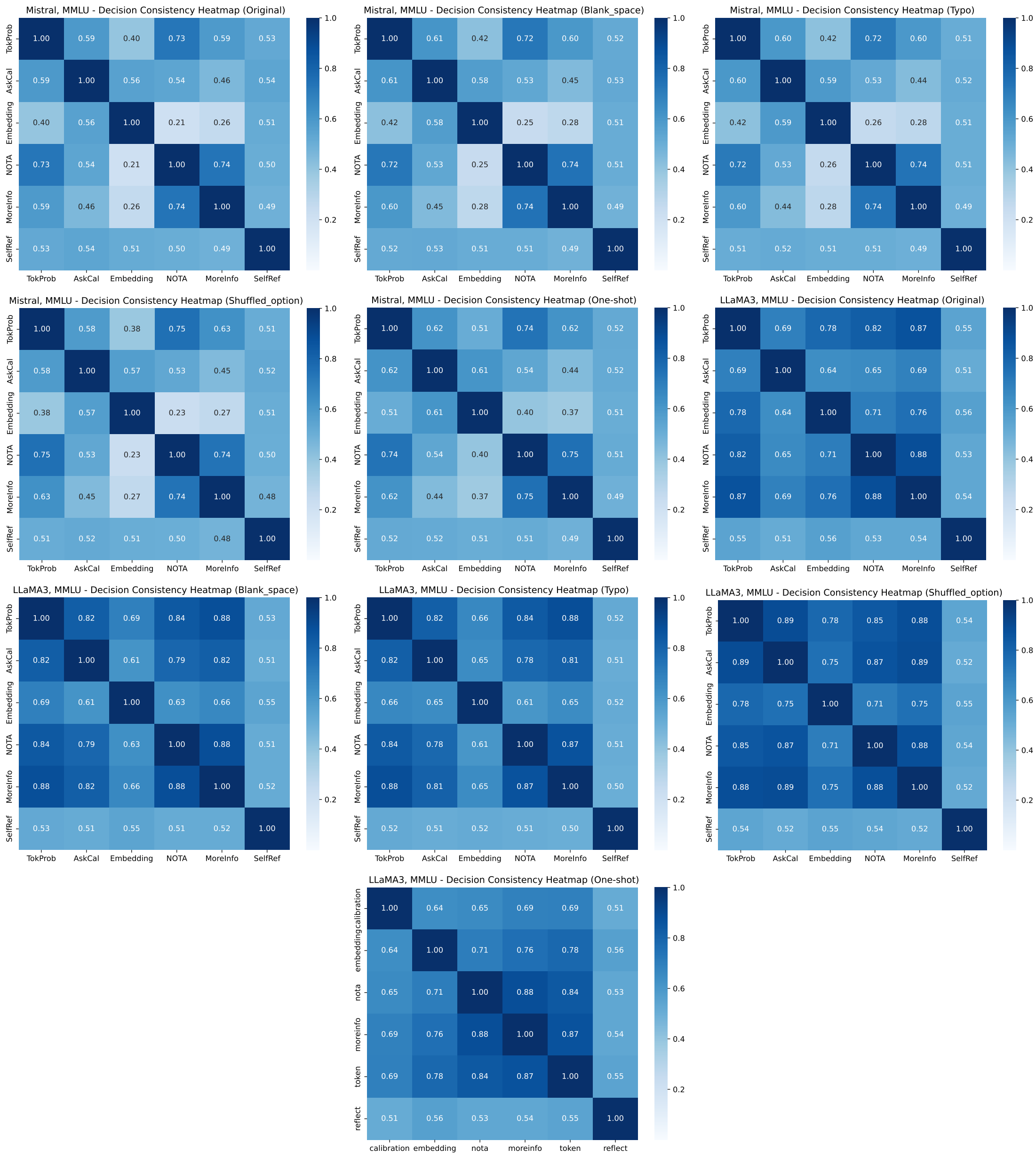


Figure 5: Heatmap of cross-method consistency evaluation results for MMLU. The values represent the average consistency across three different random seeds setups or different one-shot examples.


Figure 6: Heatmap of cross-method consistency evaluation results. The values represent the average consistency across three different random seeds setups.

Figure 7: Heatmap of cross-method consistency evaluation results. The values represent the average consistency across three different random seeds setups.

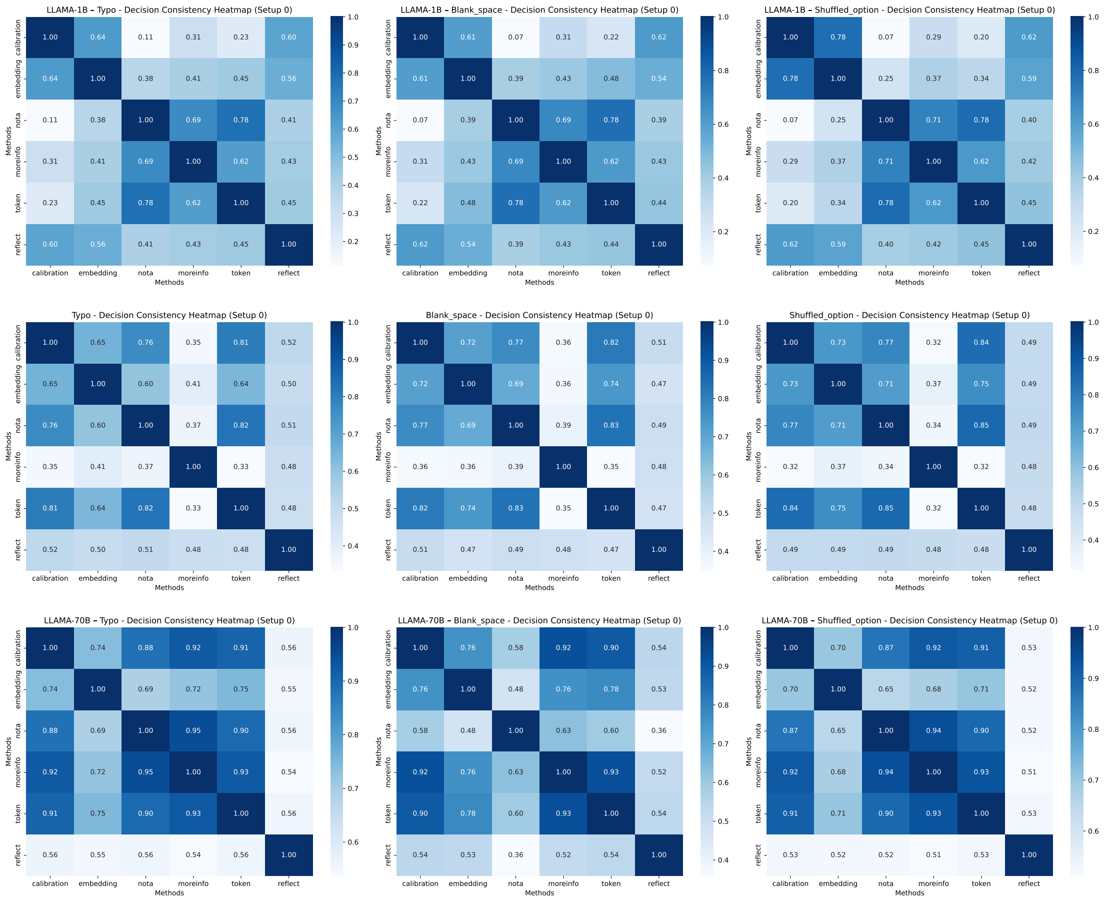


Figure 8: Decision consistency heatmaps for three LLaMA models across three prompt variants.

| Method | Variant | $\text{IoU}_{\text{cons}}$ | $\text{IoU}_{\text{acc}}$ | $\text{IoU}_{\text{rej}}$ | DecCons | Agr. | $\text{IoU}_{\text{cons}}$ | $\text{IoU}_{\text{acc}}$ | $\text{IoU}_{\text{rej}}$ | DecCons | Agr. | $\text{IoU}_{\text{cons}}$ | $\text{IoU}_{\text{acc}}$ | $\text{IoU}_{\text{rej}}$ | DecCons | Agr. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | LLaMa-3.2-1B | | | | | LLaMa-3.2-3B | | | | | LLaMa-3.1-70B | | | | |
| TOKPROB | Space | .60 | .84 | .47 | .86 | .90 | .52 | .90 | .37 | .91 | .90 | .78 | .97 | .65 | .97 | .91 |
| | Options | .41 | .78 | .28 | .80 | .28 | .44 | .89 | .29 | .89 | .18 | .70 | .96 | .55 | .96 | .19 |
| | Typo | .55 | .83 | .41 | .85 | .87 | .48 | .89 | .33 | .89 | .87 | .71 | .96 | .57 | .96 | .89 |
| | One-shot | .80 | .93 | .71 | .94 | .50 | .58 | .92 | .42 | .92 | .70 | .80 | .97 | .68 | .97 | .80 |
| ASKCAL | Space | .32 | .19 | .97 | .97 | .75 | .84 | .95 | .75 | .95 | .88 | .58 | .90 | .43 | .91 | .91 |
| | Options | .42 | .28 | .98 | .98 | .36 | .71 | .91 | .59 | .92 | .18 | .52 | .89 | .37 | .90 | .20 |
| | Typo | .21 | .12 | .94 | .94 | .70 | .79 | .93 | .68 | .94 | .85 | .55 | .90 | .40 | .90 | .89 |
| | One-shot | .05 | .03 | .99 | .99 | .00 | .41 | .81 | .28 | .82 | .70 | .45 | .87 | .31 | .88 | .80 |
| EMBEDDING | Space | .37 | .26 | .62 | .67 | .92 | .64 | .83 | .52 | .86 | .90 | .61 | .84 | .48 | .86 | .91 |
| | Options | .41 | .28 | .75 | .77 | .17 | .54 | .80 | .41 | .83 | .18 | .52 | .76 | .40 | .79 | .17 |
| | Typo | .38 | .27 | .65 | .69 | .86 | .54 | .72 | .43 | .77 | .92 | .84 | .54 | .79 | .41 | .82 |
| | One-shot | .21 | .14 | .80 | .81 | .48 | .43 | .64 | .32 | .69 | .82 | .39 | .85 | .25 | .86 | .80 |
| NOTA | Space | .32 | .94 | .19 | .94 | .83 | .47 | .91 | .32 | .91 | .92 | .23 | .62 | .14 | .62 | .61 |
| | Options | .42 | .94 | .27 | .94 | .30 | .44 | .91 | .29 | .91 | .80 | .40 | .93 | .26 | .93 | .63 |
| | Typo | .35 | .93 | .21 | .93 | .80 | .45 | .90 | .30 | .91 | .91 | .33 | .93 | .20 | .93 | .89 |
| | One-shot | .15 | .84 | .08 | .84 | .51 | .27 | .86 | .16 | .86 | .79 | .26 | .93 | .15 | .93 | .79 |
| MOREINFO | Space | .79 | .84 | .74 | .93 | .73 | .69 | .61 | .80 | .85 | .89 | .80 | 1.00 | .67 | 1.00 | .90 |
| | Options | .76 | .86 | .68 | .89 | .31 | .71 | .62 | .82 | .86 | .83 | .64 | 1.00 | .48 | 1.00 | .24 |
| | Typo | .77 | .83 | .73 | .93 | .79 | .70 | .61 | .81 | .86 | .89 | .65 | .99 | .49 | .99 | .88 |
| | One-shot | .08 | .72 | .04 | .72 | .44 | .04 | .26 | .02 | .28 | .80 | .43 | .99 | .28 | .99 | .76 |
| SELFREF | Space | .49 | .41 | .59 | .68 | .89 | .56 | .54 | .58 | .71 | .95 | .47 | .48 | .46 | .55 | .63 |
| | Options | .44 | .37 | .56 | .65 | .32 | .48 | .46 | .49 | .64 | .80 | .46 | .49 | .44 | .63 | .20 |
| | Typo | .49 | .42 | .59 | .68 | .86 | .54 | .53 | .56 | .70 | .92 | .65 | .67 | .62 | .79 | .93 |
| | One-shot | .36 | .30 | .44 | .55 | .56 | .27 | .22 | .38 | .43 | .68 | .42 | .40 | .46 | .59 | .87 |

Table 6: Intra-method consistency evaluation using six knowledge probing methods in MMLU with Llama model in different size. We introduce four different variants, each evaluated over independent runs with different random seeds or one-shot prompt examples, and the reported values represent their mean. The variance is generally is close to zero.

| Method | Variant | $IoU_{cons}$ | $IoU_{acc}$ | $IoU_{rej}$ | ∩Accuracy | Agr. |
|---|---|---|---|---|---|---|
| **Mistral-7B** | | | | | | |
| **TokProb** | Space | 0.736 ± 0.000 | 0.866 ± 0.000 | 0.640 ± 0.000 | 0.993 ± 0.000 | 0.989 ± 0.000 |
| | Options | 0.398 ± 0.001 | 0.721 ± 0.000 | 0.275 ± 0.001 | 0.756 ± 0.000 | 0.663 ± 0.000 |
| | Typo | 0.665 ± 0.000 | 0.833 ± 0.000 | 0.553 ± 0.000 | 0.980 ± 0.000 | 0.972 ± 0.000 |
| | One-shot | 0.969 ± 0.000 | 0.985 ± 0.000 | 0.952 ± 0.000 | 0.790 ± 0.002 | 0.678 ± 0.005 |
| **AskCal** | Space | 0.763 ± 0.000 | 0.765 ± 0.000 | 0.761 ± 0.000 | 0.957 ± 0.000 | 0.937 ± 0.000 |
| | Options | 0.613 ± 0.000 | 0.614 ± 0.000 | 0.612 ± 0.000 | 0.795 ± 0.000 | 0.727 ± 0.000 |
| | Typo | 0.756 ± 0.000 | 0.753 ± 0.000 | 0.758 ± 0.000 | 0.945 ± 0.000 | 0.927 ± 0.000 |
| | One-shot | 0.414 ± 0.003 | 0.413 ± 0.005 | 0.469 ± 0.018 | 0.861 ± 0.002 | 0.801 ± 0.003 |
| **Embedding** | Space | 0.584 ± 0.016 | 0.488 ± 0.023 | 0.758 ± 0.002 | 0.964 ± 0.000 | 0.945 ± 0.001 |
| | Options | 0.599 ± 0.000 | 0.495 ± 0.000 | 0.760 ± 0.000 | 0.741 ± 0.001 | 0.693 ± 0.001 |
| | Typo | 0.583 ± 0.008 | 0.481 ± 0.012 | 0.752 ± 0.001 | 0.943 ± 0.001 | 0.921 ± 0.001 |
| | One-shot | 0.332 ± 0.007 | 0.366 ± 0.006 | 0.380 ± 0.034 | 0.789 ± 0.001 | 0.691 ± 0.003 |
| **NOTA** | Space | 0.395 ± 0.001 | 0.924 ± 0.000 | 0.251 ± 0.001 | 0.941 ± 0.000 | 0.898 ± 0.000 |
| | Options | 0.393 ± 0.001 | 0.925 ± 0.000 | 0.249 ± 0.000 | 0.731 ± 0.000 | 0.571 ± 0.000 |
| | Typo | 0.390 ± 0.000 | 0.921 ± 0.000 | 0.248 ± 0.000 | 0.930 ± 0.000 | 0.878 ± 0.000 |
| | One-shot | 0.265 ± 0.002 | 0.919 ± 0.000 | 0.156 ± 0.001 | 0.778 ± 0.001 | 0.630 ± 0.002 |
| **MoreInfo** | Space | 0.740 ± 0.000 | 0.914 ± 0.000 | 0.622 ± 0.000 | 0.934 ± 0.000 | 0.884 ± 0.000 |
| | Options | 0.615 ± 0.001 | 0.879 ± 0.000 | 0.474 ± 0.001 | 0.720 ± 0.000 | 0.546 ± 0.000 |
| | Typo | 0.720 ± 0.000 | 0.906 ± 0.000 | 0.598 ± 0.000 | 0.905 ± 0.000 | 0.853 ± 0.000 |
| | One-shot | 0.037 ± 0.000 | 0.794 ± 0.000 | 0.019 ± 0.000 | 0.781 ± 0.001 | 0.640 ± 0.001 |
| **SelfRef** | Space | 0.673 ± 0.000 | 0.673 ± 0.000 | 0.672 ± 0.000 | 0.957 ± 0.000 | 0.924 ± 0.000 |
| | Options | 0.458 ± 0.000 | 0.462 ± 0.000 | 0.455 ± 0.000 | 0.708 ± 0.000 | 0.528 ± 0.000 |
| | Typo | 0.668 ± 0.000 | 0.668 ± 0.000 | 0.668 ± 0.000 | 0.943 ± 0.000 | 0.907 ± 0.000 |
| | One-shot | 0.494 ± 0.000 | 0.508 ± 0.000 | 0.482 ± 0.001 | 0.843 ± 0.000 | 0.773 ± 0.002 |
| **LLaMa-3.1-8B** | | | | | | |
| **TokProb** | Space | 0.643 ± 0.001 | 0.937 ± 0.000 | 0.491 ± 0.001 | 0.952 ± 0.000 | 0.936 ± 0.000 |
| | Options | 0.593 ± 0.001 | 0.930 ± 0.000 | 0.435 ± 0.001 | 0.827 ± 0.000 | 0.743 ± 0.000 |
| | Typo | 0.615 ± 0.001 | 0.933 ± 0.000 | 0.460 ± 0.001 | 0.934 ± 0.000 | 0.912 ± 0.000 |
| | One-shot | 0.693 ± 0.000 | 0.931 ± 0.000 | 0.552 ± 0.000 | 0.736 ± 0.004 | 0.666 ± 0.006 |
| **AskCal** | Space | 0.515 ± 0.055 | 0.789 ± 0.006 | 0.419 ± 0.072 | 0.946 ± 0.000 | 0.926 ± 0.000 |
| | Options | 0.312 ± 0.000 | 0.724 ± 0.000 | 0.199 ± 0.000 | 0.849 ± 0.000 | 0.757 ± 0.000 |
| | Typo | 0.514 ± 0.053 | 0.786 ± 0.006 | 0.418 ± 0.070 | 0.931 ± 0.000 | 0.905 ± 0.000 |
| | One-shot | 0.325 ± 0.005 | 0.544 ± 0.027 | 0.274 ± 0.013 | 0.742 ± 0.006 | 0.688 ± 0.007 |
| **Embedding** | Space | 0.500 ± 0.010 | 0.695 ± 0.033 | 0.401 ± 0.006 | 0.956 ± 0.000 | 0.940 ± 0.000 |
| | Options | 0.663 ± 0.000 | 0.835 ± 0.000 | 0.550 ± 0.000 | 0.832 ± 0.000 | 0.762 ± 0.000 |
| | Typo | 0.561 ± 0.007 | 0.714 ± 0.016 | 0.462 ± 0.004 | 0.935 ± 0.000 | 0.907 ± 0.000 |
| | One-shot | 0.385 ± 0.007 | 0.487 ± 0.019 | 0.322 ± 0.003 | 0.766 ± 0.007 | 0.702 ± 0.010 |
| **NOTA** | Space | 0.364 ± 0.000 | 0.904 ± 0.000 | 0.228 ± 0.000 | 0.942 ± 0.000 | 0.921 ± 0.000 |
| | Options | 0.387 ± 0.000 | 0.910 ± 0.000 | 0.246 ± 0.000 | 0.827 ± 0.000 | 0.698 ± 0.000 |
| | Typo | 0.361 ± 0.000 | 0.898 ± 0.000 | 0.226 ± 0.000 | 0.929 ± 0.000 | 0.896 ± 0.000 |
| | One-shot | 0.215 ± 0.002 | 0.862 ± 0.000 | 0.123 ± 0.001 | 0.763 ± 0.001 | 0.698 ± 0.001 |
| **MoreInfo** | Space | 0.863 ± 0.001 | 0.980 ± 0.000 | 0.772 ± 0.001 | 0.944 ± 0.000 | 0.916 ± 0.000 |
| | Options | 0.789 ± 0.000 | 0.969 ± 0.000 | 0.666 ± 0.000 | 0.826 ± 0.000 | 0.715 ± 0.000 |
| | Typo | 0.796 ± 0.000 | 0.968 ± 0.000 | 0.676 ± 0.000 | 0.922 ± 0.000 | 0.889 ± 0.000 |
| | One-shot | 0.088 ± 0.000 | 0.928 ± 0.000 | 0.046 ± 0.000 | 0.789 ± 0.000 | 0.713 ± 0.001 |
| **SelfRef** | Space | 0.663 ± 0.001 | 0.663 ± 0.001 | 0.662 ± 0.001 | 0.971 ± 0.000 | 0.962 ± 0.000 |
| | Options | 0.523 ± 0.000 | 0.532 ± 0.000 | 0.515 ± 0.000 | 0.880 ± 0.000 | 0.817 ± 0.000 |
| | Typo | 0.617 ± 0.000 | 0.615 ± 0.000 | 0.620 ± 0.000 | 0.960 ± 0.000 | 0.948 ± 0.000 |
| | One-shot | 0.404 ± 0.002 | 0.349 ± 0.003 | 0.485 ± 0.000 | 0.762 ± 0.007 | 0.709 ± 0.009 |

Table 7: Intra-method consistency evaluation using six knowledge probing methods in MMLU. Results represent the mean and standard deviation across six comparisons derived from three different variants generated with three different random seeds and four distinct one-shot prompting setups.

| Method | Variant | $\text{IoU}_{\text{cons}}$ | $\text{IoU}_{\text{acc}}$ | $\text{IoU}_{\text{rej}}$ | ∩Accuracy | Agr. |
|---|---|---|---|---|---|---|
| | | | **Mistral-7B** | | | |
| **TokProb** | Space | 0.781 ± 0.000 | 0.762 ± 0.000 | 0.801 ± 0.000 | 0.979 ± 0.000 | 0.963 ± 0.000 |
| | Options | 0.474 ± 0.000 | 0.439 ± 0.000 | 0.514 ± 0.000 | 0.615 ± 0.000 | 0.450 ± 0.000 |
| | Typo | 0.740 ± 0.000 | 0.717 ± 0.000 | 0.765 ± 0.000 | 0.979 ± 0.000 | 0.970 ± 0.000 |
| | One-shot | 0.904 ± 0.000 | 0.896 ± 0.000 | 0.913 ± 0.000 | 0.676 ± 0.005 | 0.488 ± 0.015 |
| **AskCal** | Space | 0.243 ± 0.069 | 0.168 ± 0.038 | 0.865 ± 0.009 | 0.889 ± 0.006 | 0.889 ± 0.006 |
| | Options | 0.049 ± 0.000 | 0.026 ± 0.000 | 0.793 ± 0.000 | 0.389 ± 0.025 | 0.389 ± 0.025 |
| | Typo | 0.134 ± 0.011 | 0.076 ± 0.004 | 0.870 ± 0.008 | 1.000 ± 0.000 | 1.000 ± 0.000 |
| | One-shot | 0.090 ± 0.025 | 0.048 ± 0.000 | 0.931 ± 0.000 | 0.771 ± 0.019 | 0.771 ± 0.019 |
| **Embedding** | Space | 0.239 ± 0.005 | 0.138 ± 0.002 | 0.975 ± 0.000 | 0.806 ± 0.020 | 0.806 ± 0.020 |
| | Options | 0.099 ± 0.003 | 0.054 ± 0.001 | 0.745 ± 0.030 | 0.835 ± 0.026 | 0.658 ± 0.080 |
| | Typo | 0.157 ± 0.000 | 0.086 ± 0.000 | 0.943 ± 0.001 | 0.620 ± 0.041 | 0.583 ± 0.032 |
| | One-shot | 0.070 ± 0.004 | 0.038 ± 0.001 | 0.709 ± 0.055 | 0.754 ± 0.032 | 0.403 ± 0.017 |
| **NOTA** | Space | 0.159 ± 0.001 | 0.883 ± 0.000 | 0.088 ± 0.001 | 0.908 ± 0.000 | 0.830 ± 0.000 |
| | Options | 0.149 ± 0.000 | 0.888 ± 0.000 | 0.082 ± 0.000 | 0.592 ± 0.000 | 0.329 ± 0.000 |
| | Typo | 0.145 ± 0.001 | 0.885 ± 0.000 | 0.079 ± 0.000 | 0.896 ± 0.000 | 0.807 ± 0.000 |
| | One-shot | 0.120 ± 0.002 | 0.900 ± 0.000 | 0.065 ± 0.001 | 0.631 ± 0.003 | 0.375 ± 0.010 |
| **MoreInfo** | Space | 0.711 ± 0.002 | 0.964 ± 0.000 | 0.565 ± 0.003 | 0.898 ± 0.000 | 0.818 ± 0.000 |
| | Options | 0.500 ± 0.001 | 0.934 ± 0.000 | 0.341 ± 0.000 | 0.594 ± 0.000 | 0.328 ± 0.000 |
| | Typo | 0.678 ± 0.001 | 0.960 ± 0.000 | 0.525 ± 0.002 | 0.895 ± 0.000 | 0.802 ± 0.000 |
| | One-shot | 0.126 ± 0.003 | 0.930 ± 0.000 | 0.068 ± 0.001 | 0.639 ± 0.002 | 0.415 ± 0.008 |
| **SelfRef** | Space | 0.660 ± 0.000 | 0.691 ± 0.000 | 0.631 ± 0.000 | 0.960 ± 0.000 | 0.933 ± 0.000 |
| | Options | 0.462 ± 0.001 | 0.510 ± 0.001 | 0.422 ± 0.000 | 0.574 ± 0.000 | 0.307 ± 0.000 |
| | Typo | 0.641 ± 0.000 | 0.672 ± 0.000 | 0.613 ± 0.000 | 0.953 ± 0.000 | 0.927 ± 0.000 |
| | One-shot | 0.463 ± 0.001 | 0.495 ± 0.006 | 0.445 ± 0.001 | 0.691 ± 0.008 | 0.509 ± 0.035 |
| | | | **LLaMa-3.1-8B** | | | |
| **TokProb** | Space | 0.526 ± 0.000 | 0.911 ± 0.000 | 0.370 ± 0.000 | 0.980 ± 0.000 | 0.973 ± 0.000 |
| | Options | 0.202 ± 0.001 | 0.841 ± 0.000 | 0.116 ± 0.001 | 0.577 ± 0.000 | 0.370 ± 0.000 |
| | Typo | 0.495 ± 0.000 | 0.898 ± 0.000 | 0.342 ± 0.000 | 0.975 ± 0.000 | 0.966 ± 0.000 |
| | One-shot | 0.799 ± 0.000 | 0.963 ± 0.000 | 0.683 ± 0.000 | 0.574 ± 0.001 | 0.347 ± 0.002 |
| **AskCal** | Space | 0.840 ± 0.000 | 0.761 ± 0.000 | 0.937 ± 0.000 | 0.956 ± 0.000 | 0.948 ± 0.000 |
| | Options | 0.660 ± 0.000 | 0.535 ± 0.000 | 0.862 ± 0.000 | 0.703 ± 0.002 | 0.525 ± 0.000 |
| | Typo | 0.833 ± 0.000 | 0.752 ± 0.000 | 0.934 ± 0.000 | 0.934 ± 0.000 | 0.924 ± 0.000 |
| | One-shot | 0.192 ± 0.037 | 0.126 ± 0.016 | 0.810 ± 0.000 | 0.284 ± 0.081 | 0.235 ± 0.057 |
| **Embedding** | Space | 0.602 ± 0.002 | 0.510 ± 0.000 | 0.736 ± 0.006 | 0.960 ± 0.000 | 0.942 ± 0.000 |
| | Options | 0.606 ± 0.003 | 0.510 ± 0.003 | 0.752 ± 0.006 | 0.612 ± 0.000 | 0.391 ± 0.000 |
| | Typo | 0.595 ± 0.000 | 0.497 ± 0.000 | 0.743 ± 0.001 | 0.939 ± 0.000 | 0.920 ± 0.000 |
| | One-shot | 0.220 ± 0.005 | 0.144 ± 0.004 | 0.604 ± 0.008 | 0.559 ± 0.021 | 0.347 ± 0.028 |
| **NOTA** | Space | 0.433 ± 0.001 | 0.658 ± 0.000 | 0.323 ± 0.001 | 0.962 ± 0.000 | 0.944 ± 0.000 |
| | Options | 0.413 ± 0.000 | 0.638 ± 0.000 | 0.306 ± 0.000 | 0.601 ± 0.001 | 0.416 ± 0.000 |
| | Typo | 0.428 ± 0.000 | 0.655 ± 0.000 | 0.318 ± 0.000 | 0.944 ± 0.000 | 0.930 ± 0.000 |
| | One-shot | 0.225 ± 0.000 | 0.613 ± 0.000 | 0.138 ± 0.000 | 0.594 ± 0.001 | 0.388 ± 0.002 |
| **MoreInfo** | Space | 0.904 ± 0.000 | 0.943 ± 0.000 | 0.868 ± 0.000 | 0.954 ± 0.000 | 0.944 ± 0.000 |
| | Options | 0.756 ± 0.000 | 0.850 ± 0.000 | 0.681 ± 0.000 | 0.567 ± 0.001 | 0.384 ± 0.000 |
| | Typo | 0.871 ± 0.000 | 0.921 ± 0.000 | 0.826 ± 0.000 | 0.949 ± 0.000 | 0.938 ± 0.000 |
| | One-shot | 0.225 ± 0.000 | 0.613 ± 0.000 | 0.138 ± 0.000 | 0.594 ± 0.001 | 0.388 ± 0.002 |
| **SelfRef** | Space | 0.714 ± 0.000 | 0.610 ± 0.001 | 0.861 ± 0.000 | 0.978 ± 0.000 | 0.976 ± 0.000 |
| | Options | 0.425 ± 0.000 | 0.304 ± 0.000 | 0.708 ± 0.000 | 0.783 ± 0.000 | 0.721 ± 0.001 |
| | Typo | 0.696 ± 0.000 | 0.592 ± 0.000 | 0.845 ± 0.000 | 0.983 ± 0.000 | 0.977 ± 0.000 |
| | One-shot | 0.326 ± 0.001 | 0.219 ± 0.001 | 0.644 ± 0.001 | 0.492 ± 0.012 | 0.346 ± 0.009 |

Table 8: Intra-method consistency evaluation using six knowledge probing methods in Hellaswag. Results represent the mean and standard deviation across six comparisons derived from three different variants generated with three different random seeds and four distinct one-shot prompting setups.

| Method | Variant | $\mathrm{IoU}_{\mathrm{cons}}$ | $\mathrm{IoU}_{\mathrm{acc}}$ | $\mathrm{IoU}_{\mathrm{rej}}$ | ∩Accuracy | Agr. |
|---|---|---|---|---|---|---|
| | | | **LLaMa-3.2-1B** | | | |
| **TokProb** | Space | 0.736 ± 0.000 | 0.817 ± 0.000 | 0.670 ± 0.000 | 0.950 ± 0.000 | 0.915 ± 0.000 |
| | Options | 0.229 ± 0.012 | 0.590 ± 0.001 | 0.156 ± 0.010 | 0.555 ± 0.002 | 0.402 ± 0.013 |
| | Typo | 0.718 ± 0.000 | 0.806 ± 0.000 | 0.647 ± 0.000 | 0.948 ± 0.000 | 0.915 ± 0.000 |
| | One-shot | 0.817 ± 0.000 | 0.876 ± 0.000 | 0.765 ± 0.000 | 0.620 ± 0.010 | 0.320 ± 0.035 |
| **AskCal** | Space | 0.000 ± 0.000 | 0.000 ± 0.000 | 0.999 ± 0.000 | 0.000 ± 0.000 | 0.000 ± 0.000 |
| | Options | 0.000 ± 0.000 | 0.000 ± 0.000 | 0.455 ± 0.149 | 0.000 ± 0.000 | 0.000 ± 0.000 |
| | Typo | 0.000 ± 0.000 | 0.000 ± 0.000 | 0.998 ± 0.000 | 0.000 ± 0.000 | 0.000 ± 0.000 |
| | One-shot | 0.000 ± 0.000 | 0.000 ± 0.000 | 0.997 ± 0.000 | 0.000 ± 0.000 | 0.000 ± 0.000 |
| **Embedding** | Space | 0.347 ± 0.016 | 0.747 ± 0.027 | 0.230 ± 0.010 | 0.939 ± 0.000 | 0.916 ± 0.000 |
| | Options | 0.215 ± 0.021 | 0.487 ± 0.114 | 0.162 ± 0.004 | 0.607 ± 0.013 | 0.254 ± 0.001 |
| | Typo | 0.056 ± 0.002 | 0.120 ± 0.014 | 0.073 ± 0.000 | 0.388 ± 0.075 | 0.312 ± 0.049 |
| | One-shot | 0.003 ± 0.000 | 0.002 ± 0.000 | 0.084 ± 0.000 | 0.350 ± 0.168 | 0.350 ± 0.168 |
| **NOTA** | Space | 0.132 ± 0.000 | 0.895 ± 0.000 | 0.071 ± 0.000 | 0.896 ± 0.000 | 0.814 ± 0.000 |
| | Options | 0.059 ± 0.001 | 0.893 ± 0.000 | 0.031 ± 0.000 | 0.531 ± 0.002 | 0.361 ± 0.007 |
| | Typo | 0.114 ± 0.001 | 0.892 ± 0.000 | 0.061 ± 0.000 | 0.887 ± 0.000 | 0.792 ± 0.000 |
| | One-shot | 0.103 ± 0.001 | 0.800 ± 0.000 | 0.055 ± 0.000 | 0.625 ± 0.008 | 0.313 ± 0.033 |
| **MoreInfo** | Space | 0.851 ± 0.000 | 0.858 ± 0.000 | 0.843 ± 0.000 | 0.862 ± 0.000 | 0.749 ± 0.000 |
| | Options | 0.692 ± 0.027 | 0.704 ± 0.025 | 0.681 ± 0.029 | 0.519 ± 0.001 | 0.512 ± 0.019 |
| | Typo | 0.853 ± 0.000 | 0.860 ± 0.000 | 0.845 ± 0.000 | 0.857 ± 0.000 | 0.749 ± 0.000 |
| | One-shot | 0.151 ± 0.022 | 0.517 ± 0.000 | 0.110 ± 0.016 | 0.615 ± 0.018 | 0.265 ± 0.061 |
| **SelfRef** | Space | 0.407 ± 0.000 | 0.290 ± 0.000 | 0.683 ± 0.000 | 0.874 ± 0.000 | 0.805 ± 0.000 |
| | Options | 0.272 ± 0.001 | 0.181 ± 0.000 | 0.548 ± 0.005 | 0.506 ± 0.001 | 0.414 ± 0.032 |
| | Typo | 0.419 ± 0.000 | 0.300 ± 0.000 | 0.695 ± 0.000 | 0.864 ± 0.001 | 0.796 ± 0.000 |
| | One-shot | 0.225 ± 0.001 | 0.139 ± 0.000 | 0.606 ± 0.001 | 0.565 ± 0.017 | 0.336 ± 0.027 |
| | | | **LLaMa-3.2-1B** | | | |
| **TokProb** | Space | 0.489 ± 0.007 | 0.897 ± 0.000 | 0.342 ± 0.006 | 0.971 ± 0.000 | 0.957 ± 0.000 |
| | Options | 0.279 ± 0.000 | 0.841 ± 0.000 | 0.167 ± 0.000 | 0.718 ± 0.000 | 0.247 ± 0.000 |
| | Typo | 0.501 ± 0.001 | 0.889 ± 0.000 | 0.350 ± 0.001 | 0.957 ± 0.000 | 0.943 ± 0.000 |
| | One-shot | 0.669 ± 0.000 | 0.923 ± 0.000 | 0.524 ± 0.000 | 0.642 ± 0.001 | 0.519 ± 0.004 |
| **AskCal** | Space | 0.794 ± 0.000 | 0.920 ± 0.000 | 0.699 ± 0.000 | 0.947 ± 0.000 | 0.930 ± 0.000 |
| | Options | 0.520 ± 0.000 | 0.814 ± 0.000 | 0.382 ± 0.000 | 0.701 ± 0.000 | 0.260 ± 0.000 |
| | Typo | 0.779 ± 0.000 | 0.915 ± 0.000 | 0.679 ± 0.000 | 0.937 ± 0.000 | 0.915 ± 0.000 |
| | One-shot | 0.209 ± 0.007 | 0.338 ± 0.072 | 0.202 ± 0.000 | 0.692 ± 0.000 | 0.605 ± 0.001 |
| **Embedding** | Space | 0.434 ± 0.002 | 0.424 ± 0.002 | 0.447 ± 0.004 | 0.945 ± 0.000 | 0.927 ± 0.000 |
| | Options | 0.513 ± 0.009 | 0.483 ± 0.006 | 0.565 ± 0.020 | 0.698 ± 0.000 | 0.232 ± 0.000 |
| | Typo | 0.446 ± 0.000 | 0.398 ± 0.002 | 0.519 ± 0.002 | 0.925 ± 0.000 | 0.904 ± 0.000 |
| | One-shot | 0.191 ± 0.012 | 0.122 ± 0.006 | 0.625 ± 0.000 | 0.548 ± 0.059 | 0.447 ± 0.035 |
| **NOTA** | Space | 0.274 ± 0.002 | 0.930 ± 0.000 | 0.161 ± 0.001 | 0.945 ± 0.000 | 0.921 ± 0.000 |
| | Options | 0.234 ± 0.004 | 0.930 ± 0.000 | 0.136 ± 0.002 | 0.701 ± 0.000 | 0.259 ± 0.000 |
| | Typo | 0.227 ± 0.002 | 0.928 ± 0.000 | 0.130 ± 0.001 | 0.930 ± 0.000 | 0.905 ± 0.000 |
| | One-shot | 0.081 ± 0.000 | 0.824 ± 0.000 | 0.043 ± 0.000 | 0.666 ± 0.001 | 0.526 ± 0.003 |
| **MoreInfo** | Space | 0.820 ± 0.001 | 0.810 ± 0.001 | 0.831 ± 0.000 | 0.945 ± 0.000 | 0.930 ± 0.000 |
| | Options | 0.688 ± 0.000 | 0.672 ± 0.000 | 0.704 ± 0.000 | 0.682 ± 0.000 | 0.225 ± 0.000 |
| | Typo | 0.807 ± 0.000 | 0.794 ± 0.000 | 0.820 ± 0.000 | 0.924 ± 0.000 | 0.906 ± 0.000 |
| | One-shot | 0.026 ± 0.000 | 0.473 ± 0.000 | 0.013 ± 0.000 | 0.664 ± 0.002 | 0.552 ± 0.002 |
| **SelfRef** | Space | 0.638 ± 0.000 | 0.583 ± 0.000 | 0.703 ± 0.000 | 0.960 ± 0.000 | 0.947 ± 0.000 |
| | Options | 0.422 ± 0.000 | 0.357 ± 0.000 | 0.515 ± 0.000 | 0.743 ± 0.000 | 0.228 ± 0.000 |
| | Typo | 0.629 ± 0.000 | 0.575 ± 0.000 | 0.694 ± 0.000 | 0.958 ± 0.000 | 0.933 ± 0.000 |
| | One-shot | 0.041 ± 0.000 | 0.022 ± 0.000 | 0.278 ± 0.001 | 0.608 ± 0.048 | 0.422 ± 0.070 |

Table 9: Intra-method consistency evaluation using six knowledge probing methods in LLaMa-3.2-1B and 3 B with Hellaswag. Results represent the mean and standard deviation across six comparisons derived from three different variants generated with three different random seeds and four distinct one-shot prompting setups.

| Method | Variant | $IoU_{cons}$ | $IoU_{acc}$ | $IoU_{rej}$ | ∩Accuracy | Agr. |
|---|---|---|---|---|---|---|
| | | | **LLaMa-3.1-70B** | | | |
| **TokProb** | Space | 0.206 ± 0.000 | 0.972 ± 0.000 | 0.116 ± 0.000 | 0.972 ± 0.000 | 0.967 ± 0.000 |
| | Options | 0.095 ± 0.001 | 0.959 ± 0.000 | 0.050 ± 0.000 | 0.959 ± 0.000 | 0.165 ± 0.000 |
| | Typo | 0.160 ± 0.004 | 0.941 ± 0.000 | 0.089 ± 0.001 | 0.942 ± 0.000 | 0.964 ± 0.000 |
| | One-shot | 0.222 ± 0.008 | 0.965 ± 0.000 | 0.128 ± 0.004 | 0.966 ± 0.000 | 0.847 ± 0.000 |
| **AskCal** | Space | 0.000 ± 0.000 | 1.000 ± 0.000 | 0.000 ± 0.000 | 1.000 ± 0.000 | 0.957 ± 0.000 |
| | Options | 0.000 ± 0.000 | 0.999 ± 0.000 | 0.000 ± 0.000 | 0.999 ± 0.000 | 0.175 ± 0.000 |
| | Typo | 0.000 ± 0.000 | 1.000 ± 0.000 | 0.000 ± 0.000 | 1.000 ± 0.000 | 0.948 ± 0.000 |
| | One-shot | 0.000 ± 0.000 | 0.964 ± 0.002 | 0.000 ± 0.000 | 0.964 ± 0.002 | 0.834 ± 0.000 |
| **Embedding** | Space | 0.361 ± 0.002 | 0.827 ± 0.007 | 0.231 ± 0.001 | 0.836 ± 0.006 | 0.957 ± 0.000 |
| | Options | 0.374 ± 0.013 | 0.930 ± 0.000 | 0.241 ± 0.009 | 0.931 ± 0.000 | 0.160 ± 0.000 |
| | Typo | 0.404 ± 0.002 | 0.861 ± 0.004 | 0.264 ± 0.001 | 0.868 ± 0.003 | 0.944 ± 0.000 |
| | One-shot | 0.122 ± 0.015 | 0.930 ± 0.000 | 0.070 ± 0.005 | 0.931 ± 0.000 | 0.835 ± 0.000 |
| **NOTA** | Space | 0.220 ± 0.001 | 0.912 ± 0.000 | 0.125 ± 0.000 | 0.913 ± 0.000 | 0.955 ± 0.000 |
| | Options | 0.252 ± 0.003 | 0.914 ± 0.000 | 0.147 ± 0.001 | 0.916 ± 0.000 | 0.161 ± 0.000 |
| | Typo | 0.227 ± 0.003 | 0.913 ± 0.000 | 0.131 ± 0.001 | 0.914 ± 0.000 | 0.948 ± 0.000 |
| | One-shot | 0.103 ± 0.001 | 0.939 ± 0.000 | 0.055 ± 0.000 | 0.940 ± 0.000 | 0.842 ± 0.000 |
| **MoreInfo** | Space | 0.000 ± 0.000 | 0.000 ± 0.000 | 0.000 ± 0.000 | 0.000 ± 0.000 | 0.000 ± 0.000 |
| | Options | 0.000 ± 0.000 | 0.000 ± 0.000 | 0.000 ± 0.000 | 0.000 ± 0.000 | 0.000 ± 0.000 |
| | Typo | 0.000 ± 0.000 | 0.000 ± 0.000 | 0.000 ± 0.000 | 0.000 ± 0.000 | 0.000 ± 0.000 |
| | One-shot | 0.000 ± 0.000 | 0.000 ± 0.000 | 0.000 ± 0.000 | 0.000 ± 0.000 | 0.000 ± 0.000 |
| **SelfRef** | Space | 0.619 ± 0.000 | 0.558 ± 0.000 | 0.694 ± 0.000 | 0.764 ± 0.000 | 0.991 ± 0.000 |
| | Options | 0.372 ± 0.000 | 0.286 ± 0.000 | 0.532 ± 0.000 | 0.598 ± 0.000 | 0.119 ± 0.000 |
| | Typo | 0.591 ± 0.000 | 0.530 ± 0.000 | 0.666 ± 0.000 | 0.737 ± 0.000 | 0.975 ± 0.000 |
| | One-shot | 0.301 ± 0.004 | 0.211 ± 0.002 | 0.530 ± 0.005 | 0.549 ± 0.009 | 0.934 ± 0.000 |
| | | | **Olmo-2-7B** | | | |
| **TokProb** | Space | 0.693 ± 0.001 | 0.708 ± 0.001 | 0.681 ± 0.002 | 0.820 ± 0.000 | 0.890 ± 0.001 |
| | Options | 0.448 ± 0.000 | 0.495 ± 0.001 | 0.415 ± 0.002 | 0.630 ± 0.000 | 0.785 ± 0.001 |
| | Typo | 0.695 ± 0.000 | 0.692 ± 0.000 | 0.698 ± 0.000 | 0.820 ± 0.000 | 0.889 ± 0.000 |
| | One-shot | 0.720 ± 0.001 | 0.708 ± 0.002 | 0.733 ± 0.001 | 0.838 ± 0.001 | 0.840 ± 0.000 |
| **AskCal** | Space | 0.497 ± 0.001 | 0.563 ± 0.007 | 0.449 ± 0.000 | 0.680 ± 0.002 | 0.765 ± 0.001 |
| | Options | 0.459 ± 0.001 | 0.623 ± 0.000 | 0.366 ± 0.002 | 0.691 ± 0.000 | 0.655 ± 0.000 |
| | Typo | 0.520 ± 0.003 | 0.681 ± 0.000 | 0.424 ± 0.004 | 0.742 ± 0.000 | 0.752 ± 0.000 |
| | One-shot | 0.439 ± 0.000 | 0.487 ± 0.001 | 0.400 ± 0.000 | 0.618 ± 0.001 | 0.725 ± 0.000 |
| **Embedding** | Space | 0.000 ± 0.000 | 0.000 ± 0.000 | 0.000 ± 0.000 | 0.000 ± 0.000 | 0.000 ± 0.000 |
| | Options | 0.000 ± 0.000 | 0.000 ± 0.000 | 0.000 ± 0.000 | 0.000 ± 0.000 | 0.000 ± 0.000 |
| | Typo | 0.000 ± 0.000 | 0.000 ± 0.000 | 0.000 ± 0.000 | 0.000 ± 0.000 | 0.000 ± 0.000 |
| | One-shot | 0.435 ± 0.000 | 0.507 ± 0.001 | 0.382 ± 0.000 | 0.622 ± 0.000 | 0.738 ± 0.000 |
| **NOTA** | Space | 0.341 ± 0.000 | 0.779 ± 0.000 | 0.218 ± 0.000 | 0.792 ± 0.000 | 0.748 ± 0.000 |
| | Options | 0.342 ± 0.001 | 0.782 ± 0.000 | 0.219 ± 0.001 | 0.794 ± 0.000 | 0.638 ± 0.000 |
| | Typo | 0.345 ± 0.001 | 0.792 ± 0.000 | 0.221 ± 0.001 | 0.803 ± 0.000 | 0.756 ± 0.000 |
| | One-shot | 0.160 ± 0.000 | 0.804 ± 0.000 | 0.089 ± 0.000 | 0.808 ± 0.000 | 0.701 ± 0.000 |
| **MoreInfo** | Space | 0.289 ± 0.001 | 0.832 ± 0.000 | 0.175 ± 0.001 | 0.838 ± 0.000 | 0.741 ± 0.000 |
| | Options | 0.221 ± 0.000 | 0.807 ± 0.000 | 0.128 ± 0.000 | 0.812 ± 0.000 | 0.630 ± 0.000 |
| | Typo | 0.302 ± 0.000 | 0.818 ± 0.000 | 0.185 ± 0.000 | 0.825 ± 0.000 | 0.736 ± 0.000 |
| | One-shot | 0.066 ± 0.001 | 0.864 ± 0.000 | 0.035 ± 0.000 | 0.864 ± 0.000 | 0.667 ± 0.000 |
| **SelfRef** | Space | 0.334 ± 0.001 | 0.210 ± 0.000 | 0.811 ± 0.000 | 0.818 ± 0.000 | 0.889 ± 0.000 |
| | Options | 0.228 ± 0.000 | 0.133 ± 0.000 | 0.794 ± 0.000 | 0.799 ± 0.000 | 0.561 ± 0.000 |
| | Typo | 0.333 ± 0.000 | 0.210 ± 0.000 | 0.810 ± 0.000 | 0.819 ± 0.000 | 0.870 ± 0.000 |
| | One-shot | 0.216 ± 0.000 | 0.126 ± 0.000 | 0.769 ± 0.002 | 0.775 ± 0.002 | 0.771 ± 0.001 |

Table 10: Intra-method consistency evaluation using six knowledge probing methods in LLaMa-3.1-70B and olmo-2-7B with Hellaswag. Results represent the mean and standard deviation across six comparisons derived from three different variants generated with three different random seeds and four distinct one-shot prompting setups.

| Method | Source | Reliable Acc. | Effective Acc. | Abstain Acc. | Abstain Prec. | Abstain Rec. | Abstain Rate | Abstain F1 |
|---|---|---|---|---|---|---|---|---|
| **AskCal** | Original | 0.500 | 0.000 | 0.631 | 0.632 | 0.995 | 0.994 | 0.773 |
| | Blank Space | 0.521 | 0.008 | 0.626 | 0.650 | 0.854 | 0.812 | 0.738 |
| | Blank Space 1 | 0.450 | -0.022 | 0.600 | 0.643 | 0.804 | 0.778 | 0.714 |
| | Blank Space 2 | 0.714 | 0.003 | 0.632 | 0.631 | 0.997 | 0.993 | 0.773 |
| | Shuffled Option | 0.455 | -0.020 | 0.618 | 0.665 | 0.809 | 0.776 | 0.730 |
| | Shuffled Option 1 | 0.443 | -0.024 | 0.624 | 0.672 | 0.819 | 0.790 | 0.739 |
| | Shuffled Option 2 | 0.467 | -0.013 | 0.622 | 0.660 | 0.835 | 0.803 | 0.737 |
| | Typo | 0.471 | -0.011 | 0.621 | 0.656 | 0.842 | 0.811 | 0.737 |
| | Typo 1 | 1.000 | 0.001 | 0.647 | 0.647 | 1.000 | 0.999 | 0.785 |
| | Typo 2 | 0.469 | -0.013 | 0.624 | 0.665 | 0.827 | 0.793 | 0.737 |
| | One-shot 1 | 0.714 | 0.030 | 0.598 | 0.589 | 0.965 | 0.930 | 0.732 |
| | One-shot 2 | 0.756 | 0.044 | 0.571 | 0.554 | 0.960 | 0.914 | 0.702 |
| | One-shot 3 | 0.653 | 0.023 | 0.558 | 0.550 | 0.951 | 0.925 | 0.697 |
| | One-shot 4 | 0.711 | 0.019 | 0.694 | 0.693 | 0.981 | 0.955 | 0.812 |
| **Embedding** | Original | 0.308 | -0.005 | 0.620 | 0.624 | 0.986 | 0.987 | 0.764 |
| | Blank Space | 0.333 | -0.004 | 0.614 | 0.617 | 0.987 | 0.988 | 0.760 |
| | Blank Space 1 | 0.462 | -0.001 | 0.632 | 0.634 | 0.989 | 0.987 | 0.773 |
| | Blank Space 2 | 0.333 | -0.012 | 0.623 | 0.634 | 0.962 | 0.964 | 0.764 |
| | Shuffled Option | 0.366 | -0.124 | 0.528 | 0.668 | 0.549 | 0.536 | 0.603 |
| | Shuffled Option 1 | 0.388 | -0.061 | 0.577 | 0.648 | 0.738 | 0.727 | 0.690 |
| | Shuffled Option 2 | 0.516 | 0.001 | 0.641 | 0.645 | 0.977 | 0.969 | 0.777 |
| | Typo | 0.361 | -0.027 | 0.615 | 0.642 | 0.903 | 0.903 | 0.751 |
| | Typo 1 | 0.273 | -0.010 | 0.618 | 0.626 | 0.975 | 0.978 | 0.762 |
| | Typo 2 | 0.467 | -0.003 | 0.628 | 0.636 | 0.962 | 0.955 | 0.765 |
| | One-shot 1 | 0.417 | -0.059 | 0.511 | 0.563 | 0.635 | 0.643 | 0.597 |
| | One-shot 2 | 0.551 | 0.013 | 0.542 | 0.541 | 0.892 | 0.873 | 0.673 |
| | One-shot 3 | 0.479 | -0.027 | 0.509 | 0.562 | 0.381 | 0.363 | 0.454 |
| | One-shot 4 | 0.381 | -0.005 | 0.669 | 0.675 | 0.981 | 0.979 | 0.800 |
| **NOTA** | Original | 0.377 | -0.230 | 0.400 | 0.721 | 0.078 | 0.068 | 0.140 |
| | Blank Space | 0.377 | -0.228 | 0.406 | 0.770 | 0.090 | 0.074 | 0.161 |
| | Blank Space 1 | 0.376 | -0.233 | 0.392 | 0.649 | 0.059 | 0.057 | 0.109 |
| | Blank Space 2 | 0.364 | -0.253 | 0.388 | 0.704 | 0.078 | 0.071 | 0.140 |
| | Shuffled Option | 0.356 | -0.270 | 0.376 | 0.683 | 0.063 | 0.060 | 0.116 |
| | Shuffled Option 1 | 0.367 | -0.250 | 0.392 | 0.774 | 0.075 | 0.062 | 0.136 |
| | Shuffled Option 2 | 0.363 | -0.257 | 0.386 | 0.730 | 0.072 | 0.063 | 0.130 |
| | Typo | 0.370 | -0.242 | 0.394 | 0.727 | 0.075 | 0.066 | 0.137 |
| | Typo 1 | 0.365 | -0.251 | 0.391 | 0.746 | 0.078 | 0.067 | 0.141 |
| | Typo 2 | 0.364 | -0.256 | 0.382 | 0.661 | 0.064 | 0.062 | 0.117 |
| | One-shot 1 | 0.432 | -0.133 | 0.437 | 0.621 | 0.032 | 0.029 | 0.060 |
| | One-shot 2 | 0.469 | -0.060 | 0.475 | 0.727 | 0.030 | 0.022 | 0.057 |
| | One-shot 3 | 0.461 | -0.075 | 0.475 | 0.755 | 0.067 | 0.049 | 0.124 |
| | One-shot 4 | 0.353 | -0.276 | 0.376 | 0.719 | 0.071 | 0.064 | 0.128 |
| **MoreInfo** | Original | 0.369 | -0.246 | 0.385 | 0.625 | 0.063 | 0.064 | 0.115 |
| | Blank Space | 0.373 | -0.240 | 0.385 | 0.589 | 0.053 | 0.056 | 0.097 |
| | Blank Space 1 | 0.369 | -0.246 | 0.387 | 0.667 | 0.063 | 0.060 | 0.115 |
| | Blank Space 2 | 0.354 | -0.275 | 0.370 | 0.623 | 0.059 | 0.061 | 0.108 |
| | Shuffled Option | 0.334 | -0.309 | 0.355 | 0.638 | 0.066 | 0.069 | 0.120 |
| | Shuffled Option 1 | 0.357 | -0.265 | 0.379 | 0.662 | 0.073 | 0.071 | 0.131 |
| | Shuffled Option 2 | 0.360 | -0.264 | 0.374 | 0.603 | 0.055 | 0.058 | 0.101 |
| | Typo | 0.368 | -0.249 | 0.384 | 0.655 | 0.057 | 0.055 | 0.105 |
| | Typo 1 | 0.372 | -0.242 | 0.386 | 0.621 | 0.057 | 0.058 | 0.105 |
| | Typo 2 | 0.357 | -0.268 | 0.375 | 0.645 | 0.062 | 0.062 | 0.113 |
| | One-shot 1 | 0.496 | -0.007 | 0.497 | 0.520 | 0.026 | 0.025 | 0.049 |
| | One-shot 2 | 0.498 | -0.004 | 0.499 | 0.600 | 0.012 | 0.010 | 0.023 |
| | One-shot 3 | 0.526 | 0.051 | 0.528 | 0.727 | 0.017 | 0.011 | 0.033 |
| | One-shot 4 | 0.386 | -0.224 | 0.389 | 0.550 | 0.018 | 0.020 | 0.035 |
| **Reflect** | Original | 0.392 | -0.120 | 0.517 | 0.674 | 0.468 | 0.442 | 0.552 |
| | Blank Space | 0.374 | -0.140 | 0.497 | 0.651 | 0.454 | 0.444 | 0.535 |
| | Blank Space 1 | 0.390 | -0.121 | 0.502 | 0.639 | 0.461 | 0.449 | 0.535 |
| | Blank Space 2 | 0.363 | -0.153 | 0.493 | 0.657 | 0.450 | 0.443 | 0.534 |
| | Shuffled Option | 0.348 | -0.174 | 0.488 | 0.675 | 0.437 | 0.428 | 0.530 |
| | Shuffled Option 1 | 0.341 | -0.179 | 0.463 | 0.620 | 0.422 | 0.437 | 0.502 |
| | Shuffled Option 2 | 0.403 | -0.105 | 0.526 | 0.672 | 0.487 | 0.457 | 0.564 |
| | Typo | 0.365 | -0.150 | 0.494 | 0.655 | 0.452 | 0.444 | 0.535 |
| | Typo 1 | 0.364 | -0.150 | 0.502 | 0.672 | 0.462 | 0.448 | 0.547 |
| | Typo 2 | 0.397 | -0.111 | 0.516 | 0.656 | 0.480 | 0.459 | 0.554 |
| | One-shot 1 | 0.613 | 0.066 | 0.637 | 0.647 | 0.802 | 0.708 | 0.716 |
| | One-shot 2 | 0.500 | 0.000 | 0.522 | 0.556 | 0.421 | 0.396 | 0.479 |
| | One-shot 3 | 0.491 | -0.010 | 0.525 | 0.568 | 0.469 | 0.442 | 0.514 |
| | One-shot 4 | 0.366 | -0.158 | 0.503 | 0.700 | 0.434 | 0.410 | 0.536 |
| **TokenProb** | Original | 0.453 | -0.043 | 0.579 | 0.686 | 0.596 | 0.541 | 0.638 |
| | Blank Space | 0.458 | -0.037 | 0.589 | 0.694 | 0.615 | 0.555 | 0.652 |
| | Blank Space 1 | 0.444 | -0.049 | 0.584 | 0.693 | 0.616 | 0.563 | 0.652 |
| | Blank Space 2 | 0.465 | -0.030 | 0.610 | 0.721 | 0.638 | 0.566 | 0.677 |
| | Shuffled Option | 0.440 | -0.051 | 0.606 | 0.729 | 0.638 | 0.575 | 0.680 |
| | Shuffled Option 1 | 0.437 | -0.057 | 0.584 | 0.704 | 0.605 | 0.551 | 0.651 |
| | Shuffled Option 2 | 0.436 | -0.059 | 0.578 | 0.699 | 0.592 | 0.539 | 0.641 |
| | Typo | 0.459 | -0.035 | 0.611 | 0.724 | 0.642 | 0.573 | 0.681 |
| | Typo 1 | 0.465 | -0.030 | 0.611 | 0.721 | 0.641 | 0.570 | 0.679 |
| | Typo 2 | 0.454 | -0.041 | 0.584 | 0.689 | 0.610 | 0.553 | 0.647 |
| | One-shot 1 | 0.458 | -0.037 | 0.541 | 0.607 | 0.585 | 0.557 | 0.596 |
| | One-shot 2 | 0.523 | 0.021 | 0.552 | 0.576 | 0.595 | 0.549 | 0.585 |
| | One-shot 3 | 0.527 | 0.024 | 0.562 | 0.589 | 0.613 | 0.560 | 0.601 |
| | One-shot 4 | 0.378 | -0.109 | 0.563 | 0.712 | 0.588 | 0.555 | 0.644 |

Table 11: Comparative abstain performance between different variant setups and original setup on Mistral-7B in Hellaswag.

| Method | Source | Reliable Acc. | Effective Acc. | Abstain Acc. | Abstain Prec. | Abstain Rec. | Abstain Rate | Abstain F1 |
|---|---|---|---|---|---|---|---|---|
| **AskCal** | Original | 0.492 | -0.003 | 0.559 | 0.575 | 0.822 | 0.803 | 0.677 |
| | Blank Space | 0.505 | 0.002 | 0.568 | 0.583 | 0.834 | 0.810 | 0.686 |
| | Blank Space 1 | 0.513 | 0.005 | 0.569 | 0.582 | 0.839 | 0.813 | 0.687 |
| | Blank Space 2 | 0.497 | -0.001 | 0.564 | 0.580 | 0.832 | 0.811 | 0.683 |
| | Shuffled Option | 0.539 | 0.015 | 0.577 | 0.586 | 0.842 | 0.807 | 0.691 |
| | Shuffled Option 1 | 0.520 | 0.008 | 0.592 | 0.610 | 0.836 | 0.800 | 0.705 |
| | Shuffled Option 2 | 0.528 | 0.011 | 0.595 | 0.611 | 0.842 | 0.805 | 0.708 |
| | Typo | 0.479 | -0.008 | 0.563 | 0.583 | 0.827 | 0.810 | 0.684 |
| | Typo 1 | 0.472 | -0.011 | 0.564 | 0.586 | 0.823 | 0.807 | 0.685 |
| | Typo 2 | 0.508 | 0.003 | 0.572 | 0.587 | 0.833 | 0.807 | 0.689 |
| | One-shot 1 | 0.500 | 0.000 | 0.746 | 0.746 | 1.000 | 1.000 | 0.855 |
| | One-shot 2 | 0.500 | 0.000 | 0.743 | 0.743 | 1.000 | 1.000 | 0.853 |
| | One-shot 3 | 0.488 | -0.003 | 0.644 | 0.666 | 0.904 | 0.879 | 0.767 |
| | One-shot 4 | 0.508 | 0.001 | 0.631 | 0.639 | 0.951 | 0.937 | 0.765 |
| **Embedding** | Original | 0.419 | -0.043 | 0.511 | 0.544 | 0.722 | 0.735 | 0.621 |
| | Blank Space | 0.423 | -0.052 | 0.512 | 0.557 | 0.656 | 0.664 | 0.603 |
| | Blank Space 1 | 0.410 | -0.093 | 0.473 | 0.540 | 0.463 | 0.485 | 0.499 |
| | Blank Space 2 | 0.412 | -0.041 | 0.524 | 0.558 | 0.758 | 0.767 | 0.643 |
| | Shuffled Option | 0.481 | -0.008 | 0.551 | 0.569 | 0.807 | 0.792 | 0.668 |
| | Shuffled Option 1 | 0.451 | -0.031 | 0.555 | 0.603 | 0.705 | 0.685 | 0.650 |
| | Shuffled Option 2 | 0.455 | -0.043 | 0.538 | 0.614 | 0.552 | 0.523 | 0.582 |
| | Typo | 0.407 | -0.063 | 0.517 | 0.573 | 0.653 | 0.661 | 0.611 |
| | Typo 1 | 0.432 | -0.037 | 0.541 | 0.582 | 0.734 | 0.729 | 0.649 |
| | Typo 2 | 0.399 | -0.086 | 0.485 | 0.549 | 0.552 | 0.574 | 0.550 |
| | One-shot 1 | 0.296 | -0.066 | 0.681 | 0.755 | 0.847 | 0.838 | 0.799 |
| | One-shot 2 | 0.394 | -0.014 | 0.732 | 0.756 | 0.946 | 0.934 | 0.840 |
| | One-shot 3 | 0.392 | -0.050 | 0.595 | 0.656 | 0.781 | 0.768 | 0.713 |
| | One-shot 4 | 0.396 | -0.103 | 0.530 | 0.661 | 0.529 | 0.507 | 0.588 |
| **NOTA** | Original | 0.472 | -0.039 | 0.522 | 0.642 | 0.335 | 0.293 | 0.440 |
| | Blank Space | 0.456 | -0.064 | 0.495 | 0.596 | 0.299 | 0.280 | 0.398 |
| | Blank Space 1 | 0.441 | -0.086 | 0.487 | 0.610 | 0.290 | 0.272 | 0.393 |
| | Blank Space 2 | 0.475 | -0.035 | 0.531 | 0.664 | 0.346 | 0.295 | 0.455 |
| | Shuffled Option | 0.471 | -0.042 | 0.521 | 0.646 | 0.330 | 0.288 | 0.437 |
| | Shuffled Option 1 | 0.434 | -0.094 | 0.493 | 0.637 | 0.317 | 0.292 | 0.423 |
| | Shuffled Option 2 | 0.461 | -0.055 | 0.517 | 0.650 | 0.337 | 0.297 | 0.444 |
| | Typo | 0.456 | -0.063 | 0.506 | 0.635 | 0.309 | 0.277 | 0.416 |
| | Typo 1 | 0.451 | -0.070 | 0.503 | 0.630 | 0.321 | 0.292 | 0.425 |
| | Typo 2 | 0.452 | -0.070 | 0.505 | 0.646 | 0.308 | 0.274 | 0.417 |
| | One-shot 1 | 0.261 | -0.388 | 0.361 | 0.793 | 0.199 | 0.188 | 0.318 |
| | One-shot 2 | 0.246 | -0.406 | 0.341 | 0.720 | 0.193 | 0.200 | 0.304 |
| | One-shot 3 | 0.305 | -0.316 | 0.379 | 0.695 | 0.190 | 0.190 | 0.298 |
| | One-shot 4 | 0.354 | -0.242 | 0.413 | 0.700 | 0.182 | 0.170 | 0.288 |
| **MoreInfo** | Original | 0.475 | -0.035 | 0.523 | 0.637 | 0.337 | 0.295 | 0.441 |
| | Blank Space | 0.471 | -0.042 | 0.528 | 0.670 | 0.339 | 0.288 | 0.450 |
| | Blank Space 1 | 0.465 | -0.049 | 0.527 | 0.675 | 0.345 | 0.295 | 0.457 |
| | Blank Space 2 | 0.472 | -0.039 | 0.527 | 0.657 | 0.345 | 0.297 | 0.452 |
| | Shuffled Option | 0.491 | -0.013 | 0.554 | 0.697 | 0.377 | 0.307 | 0.490 |
| | Shuffled Option 1 | 0.442 | -0.081 | 0.508 | 0.662 | 0.336 | 0.299 | 0.446 |
| | Shuffled Option 2 | 0.461 | -0.054 | 0.526 | 0.673 | 0.355 | 0.306 | 0.465 |
| | Typo | 0.457 | -0.060 | 0.513 | 0.641 | 0.342 | 0.306 | 0.446 |
| | Typo 1 | 0.472 | -0.038 | 0.537 | 0.681 | 0.367 | 0.310 | 0.477 |
| | Typo 2 | 0.471 | -0.041 | 0.527 | 0.657 | 0.350 | 0.303 | 0.457 |
| | One-shot 1 | 0.349 | -0.298 | 0.354 | 0.714 | 0.015 | 0.014 | 0.030 |
| | One-shot 2 | 0.390 | -0.185 | 0.437 | 0.688 | 0.174 | 0.157 | 0.277 |
| | One-shot 3 | 0.414 | -0.171 | 0.414 | 0.333 | 0.002 | 0.003 | 0.003 |
| | One-shot 4 | 0.499 | -0.002 | 0.499 | 0.500 | 0.014 | 0.014 | 0.027 |
| **Reflect** | Original | 0.643 | 0.066 | 0.625 | 0.619 | 0.853 | 0.770 | 0.718 |
| | Blank Space | 0.632 | 0.060 | 0.630 | 0.630 | 0.853 | 0.772 | 0.724 |
| | Blank Space 1 | 0.662 | 0.075 | 0.639 | 0.632 | 0.862 | 0.769 | 0.729 |
| | Blank Space 2 | 0.691 | 0.083 | 0.650 | 0.639 | 0.882 | 0.783 | 0.741 |
| | Shuffled Option | 0.637 | 0.067 | 0.627 | 0.624 | 0.841 | 0.755 | 0.716 |
| | Shuffled Option 1 | 0.667 | 0.077 | 0.669 | 0.670 | 0.870 | 0.769 | 0.757 |
| | Shuffled Option 2 | 0.634 | 0.065 | 0.646 | 0.650 | 0.847 | 0.757 | 0.735 |
| | Typo | 0.655 | 0.077 | 0.655 | 0.655 | 0.851 | 0.751 | 0.740 |
| | Typo 1 | 0.667 | 0.080 | 0.652 | 0.647 | 0.860 | 0.760 | 0.739 |
| | Typo 2 | 0.632 | 0.060 | 0.631 | 0.631 | 0.853 | 0.772 | 0.725 |
| | One-shot 1 | 0.290 | -0.107 | 0.636 | 0.754 | 0.756 | 0.745 | 0.755 |
| | One-shot 2 | 0.309 | -0.084 | 0.655 | 0.753 | 0.794 | 0.780 | 0.773 |
| | One-shot 3 | 0.484 | -0.005 | 0.637 | 0.665 | 0.877 | 0.847 | 0.756 |
| | One-shot 4 | 0.471 | -0.015 | 0.602 | 0.647 | 0.781 | 0.745 | 0.708 |
| **TokenProb** | Original | 0.451 | -0.089 | 0.466 | 0.611 | 0.105 | 0.095 | 0.178 |
| | Blank Space | 0.448 | -0.096 | 0.469 | 0.702 | 0.104 | 0.084 | 0.182 |
| | Blank Space 1 | 0.436 | -0.116 | 0.457 | 0.667 | 0.105 | 0.090 | 0.181 |
| | Blank Space 2 | 0.447 | -0.096 | 0.472 | 0.708 | 0.120 | 0.096 | 0.205 |
| | Shuffled Option | 0.456 | -0.079 | 0.484 | 0.716 | 0.139 | 0.109 | 0.232 |
| | Shuffled Option 1 | 0.417 | -0.150 | 0.433 | 0.580 | 0.099 | 0.100 | 0.170 |
| | Shuffled Option 2 | 0.433 | -0.122 | 0.455 | 0.670 | 0.109 | 0.094 | 0.188 |
| | Typo | 0.423 | -0.140 | 0.440 | 0.606 | 0.098 | 0.094 | 0.169 |
| | Typo 1 | 0.437 | -0.113 | 0.464 | 0.695 | 0.127 | 0.105 | 0.214 |
| | Typo 2 | 0.441 | -0.105 | 0.465 | 0.661 | 0.126 | 0.109 | 0.212 |
| | One-shot 1 | 0.246 | -0.462 | 0.285 | 0.678 | 0.082 | 0.090 | 0.146 |
| | One-shot 2 | 0.245 | -0.463 | 0.285 | 0.681 | 0.083 | 0.091 | 0.148 |
| | One-shot 3 | 0.356 | -0.266 | 0.378 | 0.641 | 0.078 | 0.078 | 0.139 |
| | One-shot 4 | 0.368 | -0.243 | 0.388 | 0.614 | 0.081 | 0.083 | 0.143 |

Table 12: Comparative abstain performance between different variant setups and original setup on LLaMa-3.1-8B in Hellaswag.

| Method | Source | Reliable Acc. | Effective Acc. | Abstain Acc. | Abstain Prec. | Abstain Rec. | Abstain Rate | Abstain F1 |
|---|---|---|---|---|---|---|---|---|
| AskCal | Original | 0.648 | 0.149 | 0.650 | 0.653 | 0.645 | 0.495 | 0.649 |
| | Blank Space | 0.621 | 0.125 | 0.634 | 0.647 | 0.617 | 0.485 | 0.632 |
| | Blank Space 1 | 0.621 | 0.119 | 0.638 | 0.654 | 0.642 | 0.509 | 0.648 |
| | Blank Space 2 | 0.621 | 0.123 | 0.639 | 0.657 | 0.628 | 0.493 | 0.642 |
| | Shuffled Option | 0.604 | 0.106 | 0.620 | 0.637 | 0.605 | 0.488 | 0.621 |
| | Shuffled Option 1 | 0.640 | 0.135 | 0.641 | 0.642 | 0.656 | 0.517 | 0.649 |
| | Shuffled Option 2 | 0.649 | 0.148 | 0.640 | 0.631 | 0.644 | 0.502 | 0.638 |
| | Typo | 0.618 | 0.112 | 0.638 | 0.656 | 0.656 | 0.526 | 0.656 |
| | Typo 1 | 0.629 | 0.126 | 0.650 | 0.670 | 0.655 | 0.512 | 0.662 |
| | Typo 2 | 0.606 | 0.105 | 0.633 | 0.659 | 0.633 | 0.507 | 0.645 |
| | One-shot 1 | 0.588 | 0.056 | 0.589 | 0.590 | 0.752 | 0.680 | 0.661 |
| | One-shot 2 | 0.657 | 0.067 | 0.574 | 0.551 | 0.856 | 0.787 | 0.671 |
| | One-shot 3 | 0.543 | 0.071 | 0.545 | 0.556 | 0.200 | 0.171 | 0.295 |
| | One-shot 4 | 0.600 | 0.062 | 0.553 | 0.532 | 0.747 | 0.690 | 0.622 |
| Embedding | Original | 0.641 | 0.104 | 0.598 | 0.573 | 0.731 | 0.630 | 0.642 |
| | Blank Space | 0.675 | 0.067 | 0.581 | 0.559 | 0.879 | 0.809 | 0.683 |
| | Blank Space 1 | 0.773 | 0.065 | 0.574 | 0.547 | 0.947 | 0.881 | 0.694 |
| | Blank Space 2 | 0.637 | 0.094 | 0.615 | 0.604 | 0.760 | 0.656 | 0.673 |
| | Shuffled Option | 0.712 | 0.087 | 0.610 | 0.584 | 0.887 | 0.795 | 0.704 |
| | Shuffled Option 1 | 0.648 | 0.062 | 0.569 | 0.548 | 0.854 | 0.790 | 0.668 |
| | Shuffled Option 2 | 0.715 | 0.086 | 0.576 | 0.541 | 0.884 | 0.800 | 0.671 |
| | Typo | 0.667 | 0.056 | 0.572 | 0.553 | 0.891 | 0.832 | 0.682 |
| | Typo 1 | 0.614 | 0.068 | 0.593 | 0.584 | 0.781 | 0.702 | 0.668 |
| | Typo 2 | 0.695 | 0.068 | 0.597 | 0.576 | 0.900 | 0.826 | 0.703 |
| | One-shot 1 | 0.511 | 0.015 | 0.548 | 0.619 | 0.394 | 0.339 | 0.482 |
| | One-shot 2 | 0.499 | -0.002 | 0.531 | 0.615 | 0.321 | 0.278 | 0.422 |
| | One-shot 3 | 0.482 | -0.032 | 0.501 | 0.663 | 0.129 | 0.104 | 0.217 |
| | One-shot 4 | 0.603 | 0.030 | 0.563 | 0.556 | 0.891 | 0.854 | 0.685 |
| NOTA | Original | 0.515 | 0.028 | 0.527 | 0.710 | 0.088 | 0.062 | 0.157 |
| | Blank Space | 0.498 | -0.004 | 0.513 | 0.788 | 0.079 | 0.052 | 0.144 |
| | Blank Space 1 | 0.499 | -0.002 | 0.512 | 0.760 | 0.074 | 0.050 | 0.135 |
| | Blank Space 2 | 0.494 | -0.012 | 0.512 | 0.765 | 0.099 | 0.068 | 0.176 |
| | Shuffled Option | 0.495 | -0.010 | 0.507 | 0.700 | 0.081 | 0.060 | 0.146 |
| | Shuffled Option 1 | 0.502 | 0.003 | 0.511 | 0.667 | 0.075 | 0.057 | 0.135 |
| | Shuffled Option 2 | 0.517 | 0.033 | 0.520 | 0.566 | 0.062 | 0.053 | 0.111 |
| | Typo | 0.477 | -0.043 | 0.480 | 0.527 | 0.055 | 0.055 | 0.100 |
| | Typo 1 | 0.490 | -0.018 | 0.503 | 0.707 | 0.079 | 0.058 | 0.142 |
| | Typo 2 | 0.479 | -0.039 | 0.490 | 0.661 | 0.074 | 0.059 | 0.133 |
| | One-shot 1 | 0.511 | 0.020 | 0.519 | 0.641 | 0.082 | 0.064 | 0.146 |
| | One-shot 2 | 0.511 | 0.022 | 0.515 | 0.625 | 0.041 | 0.032 | 0.076 |
| | One-shot 3 | 0.543 | 0.084 | 0.548 | 0.750 | 0.039 | 0.024 | 0.074 |
| | One-shot 4 | 0.517 | 0.033 | 0.518 | 0.560 | 0.029 | 0.025 | 0.055 |
| MoreInfo | Original | 0.508 | 0.014 | 0.512 | 0.536 | 0.151 | 0.140 | 0.235 |
| | Blank Space | 0.485 | -0.024 | 0.495 | 0.539 | 0.187 | 0.180 | 0.278 |
| | Blank Space 1 | 0.490 | -0.016 | 0.492 | 0.500 | 0.159 | 0.162 | 0.242 |
| | Blank Space 2 | 0.486 | -0.024 | 0.490 | 0.512 | 0.169 | 0.170 | 0.254 |
| | Shuffled Option | 0.484 | -0.027 | 0.488 | 0.510 | 0.148 | 0.149 | 0.229 |
| | Shuffled Option 1 | 0.498 | -0.003 | 0.504 | 0.538 | 0.154 | 0.145 | 0.239 |
| | Shuffled Option 2 | 0.501 | 0.002 | 0.493 | 0.451 | 0.151 | 0.164 | 0.226 |
| | Typo | 0.475 | -0.041 | 0.471 | 0.452 | 0.156 | 0.177 | 0.232 |
| | Typo 1 | 0.484 | -0.026 | 0.493 | 0.537 | 0.170 | 0.164 | 0.258 |
| | Typo 2 | 0.474 | -0.042 | 0.484 | 0.528 | 0.179 | 0.178 | 0.267 |
| | One-shot 1 | 0.523 | 0.045 | 0.523 | 0.667 | 0.004 | 0.003 | 0.008 |
| | One-shot 2 | 0.522 | 0.043 | 0.524 | 0.778 | 0.015 | 0.009 | 0.029 |
| | One-shot 3 | 0.551 | 0.100 | 0.550 | 0.500 | 0.011 | 0.010 | 0.022 |
| | One-shot 4 | 0.527 | 0.053 | 0.527 | 1.000 | 0.002 | 0.001 | 0.004 |
| Reflect | Original | 0.508 | 0.008 | 0.509 | 0.510 | 0.497 | 0.488 | 0.504 |
| | Blank Space | 0.483 | -0.017 | 0.496 | 0.509 | 0.499 | 0.503 | 0.504 |
| | Blank Space 1 | 0.497 | -0.003 | 0.508 | 0.519 | 0.517 | 0.509 | 0.518 |
| | Blank Space 2 | 0.474 | -0.026 | 0.485 | 0.496 | 0.483 | 0.498 | 0.490 |
| | Shuffled Option | 0.484 | -0.016 | 0.501 | 0.518 | 0.509 | 0.508 | 0.513 |
| | Shuffled Option 1 | 0.493 | -0.007 | 0.500 | 0.507 | 0.523 | 0.523 | 0.515 |
| | Shuffled Option 2 | 0.507 | 0.007 | 0.498 | 0.488 | 0.473 | 0.475 | 0.480 |
| | Typo | 0.482 | -0.018 | 0.495 | 0.508 | 0.499 | 0.504 | 0.503 |
| | Typo 1 | 0.482 | -0.018 | 0.500 | 0.518 | 0.512 | 0.512 | 0.515 |
| | Typo 2 | 0.461 | -0.040 | 0.487 | 0.514 | 0.480 | 0.492 | 0.497 |
| | One-shot 1 | 0.544 | 0.050 | 0.524 | 0.498 | 0.458 | 0.436 | 0.477 |
| | One-shot 2 | 0.537 | 0.033 | 0.502 | 0.474 | 0.554 | 0.549 | 0.511 |
| | One-shot 3 | 0.556 | 0.060 | 0.513 | 0.463 | 0.472 | 0.462 | 0.468 |
| | One-shot 4 | 0.512 | 0.013 | 0.498 | 0.481 | 0.454 | 0.457 | 0.467 |
| TokenProb | Original | 0.570 | 0.108 | 0.609 | 0.737 | 0.341 | 0.232 | 0.467 |
| | Blank Space | 0.566 | 0.096 | 0.600 | 0.694 | 0.369 | 0.268 | 0.482 |
| | Blank Space 1 | 0.564 | 0.097 | 0.608 | 0.741 | 0.358 | 0.247 | 0.483 |
| | Blank Space 2 | 0.567 | 0.099 | 0.607 | 0.719 | 0.372 | 0.263 | 0.490 |
| | Shuffled Option | 0.553 | 0.083 | 0.598 | 0.768 | 0.315 | 0.211 | 0.446 |
| | Shuffled Option 1 | 0.544 | 0.069 | 0.566 | 0.644 | 0.284 | 0.219 | 0.394 |
| | Shuffled Option 2 | 0.567 | 0.107 | 0.597 | 0.712 | 0.298 | 0.205 | 0.420 |
| | Typo | 0.555 | 0.083 | 0.608 | 0.776 | 0.356 | 0.241 | 0.488 |
| | Typo 1 | 0.552 | 0.078 | 0.607 | 0.774 | 0.363 | 0.248 | 0.494 |
| | Typo 2 | 0.551 | 0.075 | 0.601 | 0.743 | 0.369 | 0.261 | 0.493 |
| | One-shot 1 | 0.511 | 0.017 | 0.541 | 0.641 | 0.282 | 0.231 | 0.392 |
| | One-shot 2 | 0.534 | 0.053 | 0.562 | 0.656 | 0.293 | 0.227 | 0.405 |
| | One-shot 3 | 0.582 | 0.127 | 0.598 | 0.651 | 0.316 | 0.229 | 0.426 |
| | One-shot 4 | 0.558 | 0.090 | 0.581 | 0.659 | 0.303 | 0.226 | 0.416 |

Table 13: Comparative abstain performance between different variant setups and original setup on Mistral-7B in MMLU.

| Method | Source | Reliable Acc. | Effective Acc. | Abstain Acc. | Abstain Prec. | Abstain Rec. | Abstain Rate | Abstain F1 |
|---|---|---|---|---|---|---|---|---|
| AskCal | Original | 0.739 | 0.327 | 0.672 | 0.527 | 0.484 | 0.317 | 0.505 |
| | Blank Space | 0.671 | 0.315 | 0.675 | 0.727 | 0.156 | 0.077 | 0.256 |
| | Blank Space 1 | 0.681 | 0.335 | 0.685 | 0.727 | 0.160 | 0.077 | 0.262 |
| | Blank Space 2 | 0.736 | 0.322 | 0.666 | 0.516 | 0.477 | 0.318 | 0.495 |
| | Shuffled Option | 0.665 | 0.307 | 0.665 | 0.667 | 0.128 | 0.069 | 0.215 |
| | Shuffled Option 1 | 0.654 | 0.286 | 0.655 | 0.671 | 0.127 | 0.070 | 0.214 |
| | Shuffled Option 2 | 0.673 | 0.323 | 0.664 | 0.536 | 0.109 | 0.069 | 0.180 |
| | Typo | 0.720 | 0.294 | 0.660 | 0.539 | 0.489 | 0.332 | 0.513 |
| | Typo 1 | 0.657 | 0.290 | 0.664 | 0.744 | 0.155 | 0.078 | 0.257 |
| | Typo 2 | 0.659 | 0.291 | 0.664 | 0.723 | 0.161 | 0.083 | 0.263 |
| | One-shot 1 | 0.629 | 0.237 | 0.635 | 0.704 | 0.143 | 0.081 | 0.238 |
| | One-shot 2 | 0.515 | 0.008 | 0.583 | 0.608 | 0.774 | 0.732 | 0.681 |
| | One-shot 3 | 0.575 | 0.058 | 0.530 | 0.502 | 0.653 | 0.614 | 0.567 |
| | One-shot 4 | 0.563 | 0.116 | 0.577 | 0.750 | 0.124 | 0.076 | 0.212 |
| Embedding | Original | 0.709 | 0.320 | 0.668 | 0.534 | 0.359 | 0.234 | 0.430 |
| | Blank Space | 0.754 | 0.171 | 0.526 | 0.410 | 0.766 | 0.663 | 0.534 |
| | Blank Space 1 | 0.692 | 0.294 | 0.652 | 0.521 | 0.341 | 0.234 | 0.412 |
| | Blank Space 2 | 0.679 | 0.325 | 0.677 | 0.659 | 0.170 | 0.091 | 0.271 |
| | Shuffled Option | 0.700 | 0.310 | 0.662 | 0.531 | 0.341 | 0.226 | 0.415 |
| | Shuffled Option 1 | 0.680 | 0.290 | 0.655 | 0.551 | 0.296 | 0.196 | 0.385 |
| | Shuffled Option 2 | 0.711 | 0.296 | 0.639 | 0.470 | 0.411 | 0.300 | 0.439 |
| | Typo | 0.707 | 0.182 | 0.546 | 0.420 | 0.646 | 0.560 | 0.509 |
| | Typo 1 | 0.691 | 0.296 | 0.665 | 0.575 | 0.352 | 0.226 | 0.437 |
| | Typo 2 | 0.685 | 0.257 | 0.629 | 0.502 | 0.411 | 0.305 | 0.452 |
| | One-shot 1 | 0.695 | 0.241 | 0.640 | 0.551 | 0.526 | 0.381 | 0.538 |
| | One-shot 2 | 0.448 | -0.048 | 0.543 | 0.625 | 0.567 | 0.536 | 0.594 |
| | One-shot 3 | 0.596 | 0.083 | 0.564 | 0.540 | 0.636 | 0.567 | 0.584 |
| | One-shot 4 | 0.641 | 0.071 | 0.540 | 0.506 | 0.808 | 0.749 | 0.622 |
| NOTA | Original | 0.667 | 0.309 | 0.662 | 0.603 | 0.125 | 0.073 | 0.207 |
| | Blank Space | 0.663 | 0.304 | 0.658 | 0.586 | 0.116 | 0.070 | 0.193 |
| | Blank Space 1 | 0.672 | 0.320 | 0.664 | 0.557 | 0.113 | 0.070 | 0.188 |
| | Blank Space 2 | 0.679 | 0.333 | 0.674 | 0.606 | 0.126 | 0.071 | 0.209 |
| | Shuffled Option | 0.656 | 0.291 | 0.644 | 0.477 | 0.088 | 0.065 | 0.148 |
| | Shuffled Option 1 | 0.642 | 0.263 | 0.640 | 0.613 | 0.122 | 0.075 | 0.204 |
| | Shuffled Option 2 | 0.675 | 0.329 | 0.665 | 0.508 | 0.092 | 0.061 | 0.156 |
| | Typo | 0.655 | 0.286 | 0.646 | 0.539 | 0.114 | 0.076 | 0.188 |
| | Typo 1 | 0.651 | 0.277 | 0.648 | 0.614 | 0.137 | 0.083 | 0.225 |
| | Typo 2 | 0.647 | 0.271 | 0.639 | 0.544 | 0.117 | 0.079 | 0.192 |
| | One-shot 1 | 0.592 | 0.169 | 0.587 | 0.531 | 0.103 | 0.081 | 0.172 |
| | One-shot 2 | 0.485 | -0.026 | 0.494 | 0.565 | 0.117 | 0.108 | 0.194 |
| | One-shot 3 | 0.534 | 0.060 | 0.536 | 0.549 | 0.141 | 0.122 | 0.224 |
| | One-shot 4 | 0.538 | 0.070 | 0.543 | 0.605 | 0.097 | 0.076 | 0.168 |
| MoreInfo | Original | 0.686 | 0.346 | 0.685 | 0.676 | 0.136 | 0.070 | 0.226 |
| | Blank Space | 0.671 | 0.318 | 0.671 | 0.676 | 0.130 | 0.068 | 0.219 |
| | Blank Space 1 | 0.663 | 0.304 | 0.666 | 0.700 | 0.135 | 0.070 | 0.227 |
| | Blank Space 2 | 0.670 | 0.316 | 0.672 | 0.700 | 0.138 | 0.070 | 0.230 |
| | Shuffled Option | 0.662 | 0.301 | 0.659 | 0.620 | 0.123 | 0.071 | 0.205 |
| | Shuffled Option 1 | 0.656 | 0.288 | 0.654 | 0.635 | 0.128 | 0.074 | 0.214 |
| | Shuffled Option 2 | 0.686 | 0.346 | 0.678 | 0.574 | 0.117 | 0.068 | 0.195 |
| | Typo | 0.662 | 0.299 | 0.666 | 0.714 | 0.150 | 0.077 | 0.248 |
| | Typo 1 | 0.667 | 0.307 | 0.675 | 0.765 | 0.168 | 0.081 | 0.276 |
| | Typo 2 | 0.660 | 0.296 | 0.660 | 0.658 | 0.137 | 0.076 | 0.227 |
| | One-shot 1 | 0.577 | 0.153 | 0.576 | 0.444 | 0.009 | 0.009 | 0.019 |
| | One-shot 2 | 0.534 | 0.067 | 0.537 | 0.889 | 0.017 | 0.009 | 0.033 |
| | One-shot 3 | 0.611 | 0.221 | 0.611 | 0.571 | 0.010 | 0.007 | 0.020 |
| | One-shot 4 | 0.579 | 0.156 | 0.580 | 0.700 | 0.017 | 0.010 | 0.032 |
| Reflect | Original | 0.738 | 0.248 | 0.598 | 0.446 | 0.611 | 0.480 | 0.516 |
| | Blank Space | 0.741 | 0.241 | 0.597 | 0.453 | 0.635 | 0.499 | 0.529 |
| | Blank Space 1 | 0.729 | 0.234 | 0.586 | 0.437 | 0.608 | 0.490 | 0.508 |
| | Blank Space 2 | 0.740 | 0.240 | 0.588 | 0.436 | 0.626 | 0.500 | 0.514 |
| | Shuffled Option | 0.750 | 0.251 | 0.600 | 0.449 | 0.639 | 0.497 | 0.527 |
| | Shuffled Option 1 | 0.723 | 0.231 | 0.599 | 0.466 | 0.611 | 0.483 | 0.529 |
| | Shuffled Option 2 | 0.741 | 0.257 | 0.596 | 0.430 | 0.593 | 0.467 | 0.499 |
| | Typo | 0.727 | 0.223 | 0.580 | 0.438 | 0.625 | 0.509 | 0.515 |
| | Typo 1 | 0.727 | 0.223 | 0.590 | 0.458 | 0.635 | 0.509 | 0.532 |
| | Typo 2 | 0.723 | 0.222 | 0.594 | 0.466 | 0.629 | 0.502 | 0.535 |
| | One-shot 1 | 0.600 | 0.086 | 0.552 | 0.516 | 0.633 | 0.572 | 0.568 |
| | One-shot 2 | 0.534 | 0.019 | 0.628 | 0.665 | 0.785 | 0.719 | 0.720 |
| | One-shot 3 | 0.653 | 0.104 | 0.561 | 0.514 | 0.742 | 0.660 | 0.607 |
| | One-shot 4 | 0.585 | 0.048 | 0.527 | 0.504 | 0.756 | 0.718 | 0.605 |
| TokenProb | Original | 0.694 | 0.350 | 0.698 | 0.740 | 0.204 | 0.096 | 0.320 |
| | Blank Space | 0.678 | 0.331 | 0.691 | 0.870 | 0.167 | 0.069 | 0.280 |
| | Blank Space 1 | 0.691 | 0.356 | 0.705 | 0.897 | 0.175 | 0.068 | 0.293 |
| | Blank Space 2 | 0.683 | 0.330 | 0.685 | 0.708 | 0.192 | 0.096 | 0.302 |
| | Shuffled Option | 0.691 | 0.355 | 0.698 | 0.781 | 0.166 | 0.073 | 0.274 |
| | Shuffled Option 1 | 0.676 | 0.325 | 0.688 | 0.840 | 0.174 | 0.075 | 0.288 |
| | Shuffled Option 2 | 0.700 | 0.372 | 0.713 | 0.875 | 0.185 | 0.072 | 0.305 |
| | Typo | 0.686 | 0.345 | 0.701 | 0.890 | 0.183 | 0.073 | 0.303 |
| | Typo 1 | 0.667 | 0.309 | 0.681 | 0.853 | 0.172 | 0.075 | 0.286 |
| | Typo 2 | 0.671 | 0.315 | 0.685 | 0.848 | 0.181 | 0.079 | 0.298 |
| | One-shot 1 | 0.630 | 0.229 | 0.636 | 0.678 | 0.201 | 0.121 | 0.311 |
| | One-shot 2 | 0.441 | -0.103 | 0.479 | 0.744 | 0.160 | 0.125 | 0.263 |
| | One-shot 3 | 0.541 | 0.072 | 0.556 | 0.661 | 0.169 | 0.124 | 0.270 |
| | One-shot 4 | 0.568 | 0.119 | 0.586 | 0.715 | 0.188 | 0.123 | 0.298 |

Table 14: Comparative abstain performance between different variant setups and original setup on LLaMa-3.1-8B in MMLU.