On the Role of Model Prior in Real-World Inductive Reasoning

Zhuo Liu
University of Rochester
zhuo.liu@rochester.edu

Ding Yu
University of Rochester
dyu18@ur.rochester.edu

Hangfeng He
University of Rochester
hangfeng.he@rochester.edu

Abstract

Large Language Models (LLMs) show impressive inductive reasoning capabilities, enabling them to generate hypotheses that could generalize effectively to new instances when guided by in-context demonstrations. However, in real-world applications, LLMs' hypothesis generation is not solely determined by these demonstrations but is significantly shaped by task-specific model priors. Despite their critical influence, the distinct contributions of model priors versus demonstrations to hypothesis generation have been underexplored. This study bridges this gap by systematically evaluating three inductive reasoning strategies across five real-world tasks with three LLMs. Our empirical findings reveal that, hypothesis generation is primarily driven by the model's inherent priors; removing demonstrations results in minimal loss of hypothesis quality and downstream usage. Further analysis shows the result is consistent across various label formats with different label configurations, and prior is hard to override, even under flipped labeling. These insights advance our understanding of the dynamics of hypothesis generation in LLMs and highlight the potential for better utilizing model priors in real-world inductive reasoning tasks.

1 Introduction

Large Language Models (LLMs) have drawn significant interests due to their performance on a diverse range of reasoning tasks (Kojima et al., 2022), such as mathematical reasoning, commonsense reasoning and symbolic reasoning. *Inductive reasoning*– an important component of reasoning (Yang et al., 2022; Heit, 2000), as a way to derive abstract hypothesis from limited specific observations, is widely regarded as a core aspect of human intelligence.

Existing studies primarily assess the inductive reasoning capabilities of LLMs (Wang et al., 2023; Qiu et al., 2023; Cheng et al., 2024; Bowen et al., 2024) by evaluating their ability to generate textual hypotheses based on in-context input-output pairs and subsequently test these hypotheses on unseen examples, thereby evaluating their generalization abilities. These studies demonstrated that LLMs can propose high-quality hypotheses, establishing them as exceptional hypothesis generators (Qiu et al., 2023; Cheng et al., 2024; Li et al., 2024).

LLMs employ various approaches to generate hypotheses depending on the nature of the task. For symbolic tasks, such as mathematical function discovery (Shojaee et al., 2024), LLMs rely primarily on input-output mappings in demonstrations, often with minimal prior knowledge about the mathematical functions. In contrast, research by Qi et al. (2023) demonstrated that LLMs can formulate hypotheses solely from provided background information, leveraging the extensive and diverse knowledge gained during pre-training. In real-world applications, hypothesis generation tends to be data-driven, such as generating hypotheses for trending Twitter headline patterns (Zhou et al., 2024), where both prior knowledge and demonstrations are utilized. In these cases, the interaction between the model's task-specific priors and provided examples is mixed.

In empirical science, data-driven hypothesis generation serves as the foundational step toward scientific discovery (Majumder et al., 2024a,b). When employing LLMs for hypothesis generation, the goal is to uncover novel hypotheses that contribute fresh insights and ideas to the existing literature (Zhou et al., 2024). However, due to the combined influence of the model's prior knowledge and the provided examples, the origin of generated hypotheses often remains unclear. For certain tasks, where LLMs are pre-trained on extensive knowledge bases, a strong model prior may even overshadow the potential for generating genuinely novel insights from the provided examples. This raises a critical question: ***What is the role of model prior in real-world inductive reasoning?***

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 10572–10595
November 4-9, 2025

To address this issue, this paper presents a systematic empirical study on real-world inductive reasoning problems, focusing on classification tasks, where hypotheses are generated to capture patterns specific to the positive class. We evaluate three representative baselines: direct input-output prompting (Qiu et al., 2023), iterative refinement with ranking (Qiu et al., 2023; Shojaee et al., 2024), and HypoGeniC (Zhou et al., 2024; Liu et al., 2024), across five diverse real-world tasks covering text, image, and image-text modalities. For each baseline, we conduct experiments where LLMs generate hypotheses both with and without demonstrations. The quality of the generated hypotheses is then evaluated from three perspectives: hypothesis-based classification performance, LLM-based assessments, and human evaluation.

Our experimental results reveal that, for real-world tasks *where LLMs have been trained on substantial amounts of relevant data*, task-specific model prior plays a dominant role in hypothesis generation. Notably, removing in-context demonstrations has minimal impact on the quality of the hypotheses. This trend holds consistently across three baselines with three LLMs: GPT-4o, Qwen2-VL and Gemini-pro, strongly suggesting that, counterintuitively, LLMs depend more on task-specific prior knowledge than on in-context demonstrations for generating hypotheses. Further analysis across various label configurations and formats supports this conclusion, indicating that model prior is often so robust that it is minimally affected by the provided examples. Our code is available at `https://github.com/joeliuz6/Model-prior-in-inductive-reasoning`.

2 Related Work

Inductive Reasoning with LLMs. Primary studies on inductive reasoning mainly focus on evaluating their inductive reasoning capabilities. Qiu et al. (2023) evaluate LLMs by inducting rules from examples, demonstrated that LLMs are good hypothesis proposers. Wang et al. (2023) uses Python programs to select better hypothesis, thus improving the inductive reasoning performance. Besides these evaluations on symbolic tasks, Yang et al. (2022) propose to induce natural language rules from natural language facts while Hypotheses-to-Theories (Zhu et al., 2023) learns rules from deduction. Similarly, Honovich et al. (2022) also show LLMs are able to infer a natural task description by provided demonstrations. Recently, some works employ LLMs to generate hypothesis that can describe the difference or shift between two distributions in different modalities, such as text (Zhong et al., 2022, 2023; Singh et al., 2022), and image (Dunlap et al., 2024; Kim et al., 2024). Distinct from these studies, our work delves into understanding how LLMs perform inductive reasoning for real-world tasks, offering insights into their underlying mechanisms.

Hypothesis Generation with LLMs. Yang et al. (2023b) uses raw web corpus as observations to generate scientific hypothesis, and Pham et al. (2023) generates hypothesis to uncover latent topics in a text collection. In Qi et al. (2023), it shows LLMs are good hypothesis proposers with only background knowledge. Majumder et al. (2024a) provides initial evidence for LLMs to do data-driven discovery, where both search and verification of hypotheses may be carried out using a dataset alone. HypoGeniC (Zhou et al., 2024) also uses LLMs to generate hypothesis from real-world labeled examples. Si et al. (2024) and Baek et al. (2024) further explore the potential to generate hypothesis in research with LLMs to provide insights and ideas for the literature. Additionally, Liu et al. (2024) combines theory-based generation and data-driven generation to get better hypothesis. However, these works do not clearly distinguish whether the hypotheses originate from hidden knowledge or provided examples, a distinction that is the central focus of our work. Similar efforts have also explored this disentanglement in in-context learning (Min et al., 2022; Sia et al., 2024; Wei et al., 2023) and question answering (Du et al., 2024), while we extend this perspective to hypothesis generation.

3 Natural Language Hypothesis Generation

Let $\mathcal{Z} = \mathcal{D}_P \cup \mathcal{D}_N$ represent the labeled data for a real-world classification task $\mathcal{T}$, where $\mathcal{D}_P$ and $\mathcal{D}_N$ correspond to demonstrations of the positive (P) and negative (N) classes, respectively. Each sample in $\mathcal{Z}$ is a pair (x, y), where x denotes the example and $y \in \{P, N\}$ represents the label. A valid natural language hypothesis h, as introduced by Zhong et al. (2022), is expressed as a natural language string. For any example x, h is capable of determining whether x belongs to the positive or negative class.

Natural language hypothesis generation involves

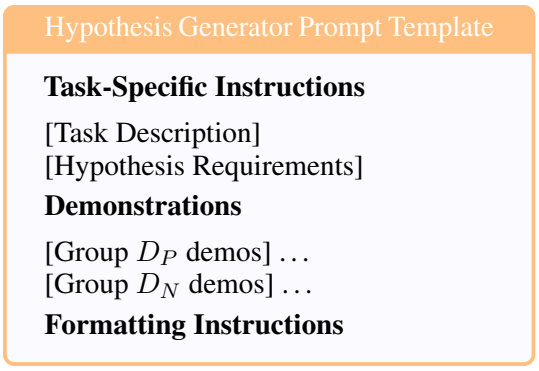

Figure 1: Prompt template for hypothesis generation.

prompting LLMs to produce a set of valid hypotheses $\mathcal{H} = \{h_1, h_2, \ldots, h_m\}$ using in-context demonstrations tailored to task $\mathcal{T}$. In this paper, we consider the setting where the input to LLMs can be divided into two parts, as shown in Figure 1: (1) **Task-Specific Instructions:** a set of natural language to describe the task and the requirements for the hypothesis. (2) **Demonstrations:** a set of exemplars from different groups structured in a specified way to show the patterns of each group. Ideally, we aim to prompt LLMs to generate a list of valid hypothesis to maximize the downstream task performance, by carefully selecting instructions and demonstrations. There are two factors contributing to the hypothesis generation:

- **Task-Specific Model Prior:** LLMs are pretrained on a diverse set of datasets, allowing them to accumulate extensive background knowledge across a wide range of domains. When provided with a task description, the model leverages its priors to infer relevant patterns, generating hypotheses based on this internalized knowledge.

- **Input-Label Mappings in Demonstrations:** The demonstrations provided serve as a specific guidance, offering cues about how to approach the task. The model may use these demonstrations to refine its hypothesis generation, aligning its output more closely with the intended task requirements.

4 Experimental Settings

4.1 Hypothesis Generation Baselines

In this paper, we evaluate three commonly-used hypothesis generation baselines.

Input-Output Prompting. Input-output prompting (IO-Prompting) represents the most common approach to prompting LLMs (Qiu et al., 2023). In this standard IO-Prompting framework, we directly provide the LLMs with a set of in-context demonstrations within the prompt context. The objective is to generate m hypotheses that effectively captures the patterns of positive class P. This approach is a single-step method, utilizing the in-context demonstrations once to guide the model's hypothesis generation.

Iterative Refinement with Ranking. Standard IO-prompting utilizes in-context demonstrations only once, potentially under utilizing their full capacity. To address this limitation, various methods have been proposed to iteratively refine hypotheses, thereby enhancing model performance (Wang et al., 2023; Qiu et al., 2023; Shojaee et al., 2024; Xiao et al., 2024). In our approach, we iteratively refine hypotheses using ranking information as a feedback signal.

The refinement process begins with an initial set of m hypotheses generated via IO-prompting. At each iteration, hypotheses in the bank are ranked based on their performance on a validation set. The top-ranked m hypotheses are then fed back to the model, along with in-context demonstrations, guiding it to generate hypotheses with improved performance. In cases where no demonstrations are available, only the ranked hypotheses with their accuracies are provided in the iterative refinement process. This approach thus augments data utilization by continuously leveraging feedback to generate higher-quality hypotheses.

Update from Mistakes: HypoGeniC. The previous methods leverage data within one single prompt to generate hypotheses, yet using all demonstrations in a single prompt may not be optimal for performance. Therefore, we also evaluate a strategy that updates hypotheses from mistakes made by current hypothesis. We largely follow an established approach, HypoGeniC (Zhou et al., 2024; Liu et al., 2024), which iteratively generate new hypotheses from incorrect prediction examples.

In our evaluation, we initialize the hypothesis bank using standard IO-prompting as well as the reward scores as in Zhou et al. (2024); Liu et al. (2024). During the update phase, if the number of incorrect examples for each group reaches a predefined number, these incorrect examples are

Dataset	Demos	IO-Prompting		Iterative-Refinement		HypoGeniC	
		Best	Average	Best	Average	Best	Average
Hallucination	w/o	63.7 ± 2.3	59.4 ± 1.1	66.9 ± 0.5	62.1 ± 0.3	61.7 ± 0.3	57.9 ± 0.5
	w/	63.8 ± 0.3	58.3 ± 0.1	63.7 ± 2.0	59.7 ± 2.6	65.6 ± 2.0	59.2 ± 1.1
Unhealthy Comments	w/o	70.3 ± 0.7	63.2 ± 0.9	71.4 ± 1.2	68.8 ± 1.4	71.2 ± 0.4	67.3 ± 1.3
	w/	70.0 ± 0.3	66.9 ± 1.2	71.8 ± 0.7	69.8 ± 0.3	71.1 ± 1.2	67.6 ± 0.8
Funny Reddit	w/o	64.1 ± 2.3	58.6 ± 0.3	67.0 ± 1.6	63.5 ± 0.4	64.4 ± 1.7	60.6 ± 1.3
	w/	65.8 ± 2.4	59.0 ± 1.4	69.8 ± 1.7	66.1 ± 0.8	62.2 ± 3.5	57.6 ± 1.0
Truthful Review	w/o	69.1 ± 0.6	57.0 ± 0.5	69.0 ± 0.7	63.8 ± 1.0	69.2 ± 0.7	59.6 ± 1.3
	w/	68.5 ± 0.9	59.7 ± 0.8	69.5 ± 1.6	63.6 ± 0.4	62.4 ± 5.1	59.4 ± 3.7
PneumoniaMNIST	w/o	75.9 ± 0.5	72.4 ± 0.4	77.6 ± 0.5	75.6 ± 0.2	76.8 ± 0.8	73.4 ± 0.3
	w/	74.7 ± 1.1	69.7 ± 0.5	76.2 ± 1.7	74.2 ± 1.1	74.6 ± 0.5	71.4 ± 0.8
Overall Average	w/o	**68.62**	62.12	**70.38**	**66.76**	**68.66**	**63.76**
	w/	68.56	**62.72**	70.20	66.68	67.18	63.04

Table 1: Accuracy comparison of *single hypothesis-based classification* across five datasets of three baselines: accuracy (*mean* ± *standard deviation*) for the best single hypothesis and the average across five hypotheses, with (**w/**) and without (**w/o**) demonstrations. The better overall average between (**w/**) and (**w/o**) is highlighted in **bold**.

	Demos	Qwen2-VL	Gemini-1.5-pro
Best	w/o	**64.98**	**64.40**
	w/	63.80	64.08
Average	w/o	**58.16**	59.08
	w/	57.64	**59.20**

Table 2: Accuracy comparison of *single hypothesis-based classification* with **Qwen2-VL-72B** and **Gemini-1.5-pro**: overall average across five datasets for the best single hypothesis and the average across five hypotheses, with (**w/**) and without (**w/o**) demonstrations. The better overall average between (**w/**) and (**w/o**) is highlighted in **bold**. Detailed results can be found in Appendix C.3.

employed to guide the generation of new hypotheses. In each update, m hypotheses with highest reward scores are kept in the hypothesis bank. This iterative updating approach enables the model to adapt hypotheses progressively, making better use of feedback from misclassifications. For a fair comparison, when demonstrations are absent, we update the hypothesis by iterative refinement, using reward scores for ranking.

4.2 Evaluation of Hypothesis

After generating a set of hypotheses $\mathcal{H} = \{h_1, h_2, \ldots, h_m\}$, it is crucial to evaluate their quality to ensure that the generated hypotheses are both functional and interpretable. We perform this evaluation from three perspectives: **hypothesis-based classification**, **LLM-based evaluation** and **human evaluation**. These complementary methods allow for a robust assessment, combining quantitative performance metrics with qualitative assessments from domain experts.

Hypothesis-Based Inference. In hypothesis-based inference (Liu et al., 2024; Zhou et al., 2024), the goal is to assess how well the generated hypotheses support downstream decision-making tasks. We measure the predictive performance of the hypothesis on a test dataset $\mathcal{D}_{\text{test}} = \{(x_j, y_j)\}_{j=1}^{N_{test}}$. The hypothesis is evaluated based on how accurately it assigns the correct label to each input x_j. Predictions are made by comparing test examples x_j with learned patterns, which can consist of a single hypothesis or multiple hypotheses. If a test example satisfies the pattern, it is assigned the corresponding class. Unless otherwise stated, the results reported in this work are based on patterns formed from single hypothesis. We also do extensive experiments to verify that LLMs can follow the provided hypothesis in the inference, which can be found in Appendix C.

LLM-Based Evaluation. In addition to assessing the effectiveness of hypotheses in downstream task usage, we also evaluate their *helpfulness* (Liu et al., 2024) and *novelty* (Liu et al., 2024; Si et al., 2024) through LLM-based metrics. Specifically: (1) *Helpfulness* measures the extent to which a hypothesis accurately captures the underlying patterns of the data and generalizes effectively to unseen samples. (2) *Novelty* assesses whether the hypothesis introduces new insights or unique perspectives relevant to the task.

Our LLM-based evaluation incorporates both scoring and pairwise comparison assessments. For scoring, LLMs assign a rating on a 5-point scale to reflect each hypothesis's quality. For pairwise comparison, we randomly pair hypotheses gener-

Dataset	IO-Prompting		Iterative-Refinement		HypoGeniC	
	w/o demos	w/ demos	w/o demos	w/ demos	w/o demos	w/ demos
Hallucination	62.2 ± 1.0	61.1 ± 0.3	60.1 ± 4.5	61.1 ± 1.3	58.6 ± 4.0	60.1 ± 0.5
Unhealthy Comments	71.5 ± 0.7	70.9 ± 0.5	71.0 ± 0.4	70.9 ± 0.3	70.9 ± 1.0	70.7 ± 2.3
Funny Reddit	58.3 ± 0.4	59.2 ± 0.3	63.9 ± 2.7	67.3 ± 1.2	58.8 ± 0.7	58.4 ± 0.5
Truthful Reviews	63.8 ± 1.4	65.3 ± 0.9	68.5 ± 0.3	69.1 ± 1.3	67.7 ± 1.5	62.1 ± 4.6
PneumoniaMNIST	75.8 ± 0.9	72.2 ± 1.2	76.0 ± 2.5	74.1 ± 1.7	74.9 ± 1.7	74.6 ± 1.0
Overall Average	**66.32**	65.74	67.90	**68.50**	**66.18**	65.18

Table 3: Accuracy comparison of *multiple hypotheses-based classification* across five datasets of three baselines: accuracy (*mean* ± *standard deviation*) with (**w/**) and without (**w/o**) demonstrations. The better overall average between (**w/**) and (**w/o**) is highlighted in **bold**.

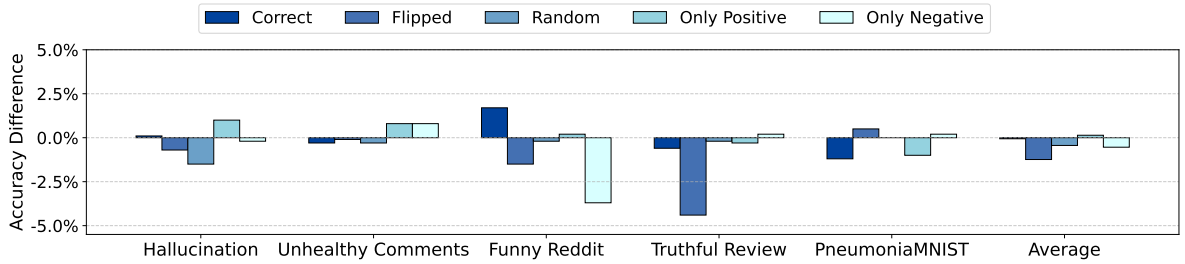

Figure 2: Accuracy difference comparison of *single hypothesis-based classification* under different label settings: Accuracy difference (*accuracy of different label settings - accuracy without demos*) across five datasets with IO-Prompting.

ated with and without demonstrations, and prompt the LLMs to select the better hypothesis in each pair. This pairwise evaluation provides insights into relative performance, while scoring offers an absolute measure of quality.

Human Evaluation. To validate the effectiveness of LLM-based evaluation, we also conduct a human evaluation to assess the quality of the generated hypotheses. Our goal is to examine the degree of alignment between LLM-based evaluation results and those obtained from human experts. Given that scoring may be challenging for human evaluators, we employ a pairwise comparison format, allowing experts to select the higher-quality hypothesis or indicate if the difference is difficult to discern. A total of nine participants are recruited for this evaluation, ensuring diverse perspectives in assessing the hypotheses.

For further details in both LLM-based and human evaluations, refer to Appendix D.

4.3 Other Settings

Models. We conduct experiments with GPT-4o, Qwen2-VL-72B[1] and gemini-1.5-pro-002, leveraging both open-source models and API-accessible models to ensure diverse evaluation. Unless otherwise stated, we use GPT-4o[2] in experiments.

Datasets. We conduct evaluations on five real-world inductive reasoning datasets: hallucination pattern induction (Li et al., 2023), unhealthy comments (Zhong et al., 2023), funny Reddit posts (Zhong et al., 2023), pneumoniaMNIST (Xiao et al., 2024), and truthful hotel reviews (Zhou et al., 2024).

Our selection of datasets is motivated by three key factors: (1) their coverage of three distinct modalities—text (unhealthy comments, funny Reddit posts, and truthful hotel reviews), image (pneumoniaMNIST), and image-text (hallucination pattern induction), (2) diverse domains, including model behavior analysis (hallucination pattern induction), medical diagnosis (pneumoniaMNIST), and social media content (unhealthy comments, funny Reddit posts, and truthful hotel reviews), and (3) their status as widely studied problems in real-world inductive reasoning tasks. Further details and more references for these datasets are provided in Appendix A.

[1]https://huggingface.co/Qwen/Qwen2-VL-72B-Instruct-AWQ

[2]By default, we use GPT-4o-2024-08-06. However, if a request is rejected due to safety reasons, we will switch to GPT-4o-2024-05-13.

	IO-Prompting		Iterative-Refinement		HypoGeniC		Overall Average	
Criteria	**w/o**	**w/**	**w/o**	**w/**	**w/o**	**w/**	**w/o**	**w/**
Helpfulness	4.00 ± 0.000	3.96 ± 0.195	4.00 ± 0.000	3.80 ± 0.400	4.04 ± 0.195	4.08 ± 0.271	**4.01**	3.95
Novelty	2.56 ± 0.571	2.40 ± 0.566	2.60 ± 0.693	2.60 ± 0.748	2.84 ± 0.674	2.36 ± 0.741	**2.67**	2.45

Table 4: LLM-based scoring: Comparison of *Helpfulness* and *Novelty* scores across three baselines, with and without demonstrations (*w/ demos* vs. *w/o demos*). The better overall average between (**w/**) and (**w/o**) is highlighted in **bold**.

Other Parameters. The number of in-context demonstrations is set to $N = 30$ for IO-prompting and iterative-refinement, and $N = 50$ for HypoGenic to encourage more updates. Examples are randomly sampled from the training set. For each dataset, we generate five candidate hypotheses. Main results are averaged over three random seeds to ensure robustness. (See Appendix B for further implementation details.)

5 Task-Specific Model Prior Dominates Hypothesis Generation

5.1 LLMs Are Zero-Shot Hypothesis Generators

To see the impact of the model prior in hypothesis generation, we compare the hypothesis generation in the following two settings.

Model Prior Only is a typical zero-shot hypothesis generation scenario without the use of demonstrations, relying primarily on prior for generation.

Demos with Ground Truth Labels is used in a typical real-world inductive reasoning tasks, with demonstrations as a specific guidance.

Results for single hypothesis-based and multiple hypotheses-based classification are shown in Table 1 and Table 3. From the results, We find that removing in-context demonstrations cause little degradation for the downstream task performance. The trend is consistent across five different datasets on three baselines. In some cases, LLMs can even generate better hypothesis using only model prior. Additionally, iterative refinement outperforms the other two baselines, showing that data still helps for hypothesis selection, but not as in-context demonstrations for hypothesis generation.

Resutls with Qwen2-VL and Gemini-1.5-pro. The results for single hypothesis-based classification on Qwen2-VL-72B and Gemini-1.5-pro-002, with IO-prompting, are provided in Table 2. The results similarly show a negligible performance drop without demonstrations, underscoring the universality of our findings across different models.

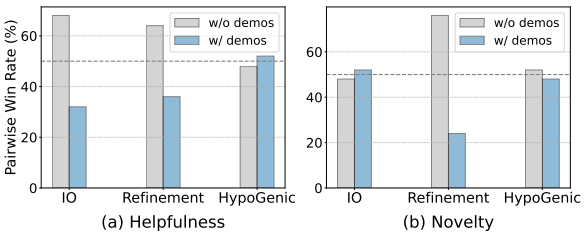

Figure 3: LLM-based pairwise comparison: Pairwise win rate (%) of three baselines. The left plot shows the comparison of *Helpfulness*, while the right plot presents *Novelty*. The dashed line indicates a tie where "w/ demos" and "w/o demos" perform equally well.

These results indicates LLMs are good zero-shot hypothesis proposers under strong prior, and in-context demonstrations with ground truth labels are not necessary to achieve acceptable hypothesis. This is a counter-intuitive phenomenon, given that labeled data is very important in in-context learning (Brown, 2020), which can inform the model of corresponding data distribution (Min et al., 2022).

5.2 Input-Label Mappings in Demonstrations Cannot Override Strong Model Prior

To further explore the interaction between model prior and input-label mappings in demonstrations in hypothesis generation, we use in-context demonstrations with different label settings:

(1) *Demos with **ground truth (correct) labels**.*
(2) *Demos with **flipped labels**.*
(3) *Demos with **random labels**.*
(4) ***Only positive** group demos.*
(5) ***Only negative** group demos.*

Figure 2 illustrates the relative accuracy difference between various label settings and without demonstrations. From the result, there is quite limited difference (mostly smaller than 3%) of performance among different settings, with the flipped label setting in truthful review as an exception, which has a performance degradation about

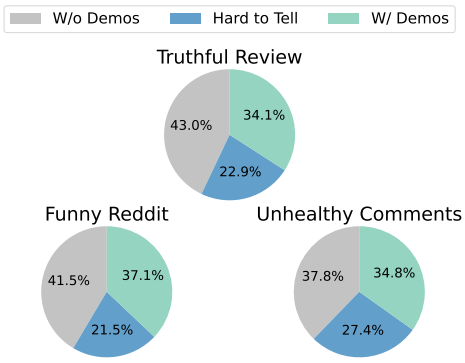

Figure 4: Human pairwise comparison results on three datasets, showing preferences for hypotheses with, without demos, and cases where it was hard to tell the difference.

4.5%. Although only positive settings surpass without demos in 3 datasets, the observed performance gap is not statically significant, with p-values $0.485, 0.460, 0.899$, respectively.

These findings suggest that while demonstrations can provide some guidance, the models' hypothesis generation abilities are ultimately shaped more by its pre-trained priors than by any superficial label configurations. Furthermore, the prior is too strong to be overridden by the patterns in demonstrations, even with totally flipped labels.

5.3 LLM-based Evaluation Results

LLM-based Scoring. Table 4 summarizes the helpfulness and novelty scores for various approaches. Each score represents the average of 25 hypotheses generated across five datasets. For helpfulness, hypotheses generated without demonstrations achieve higher scores when using IO-prompting and iterative-refinement. Regarding novelty, hypotheses generated without demonstrations score higher on IO-prompting and Hypogenic, while iterative-refinement yields a tie between the two settings.

LLM-based Pairwise Comparison. Figure 3 presents the pairwise comparison results for three baselines, evaluating hypotheses generated with and without demonstrations. The comparisons involve randomly paired hypotheses, with win rates aggregated across all datasets. For Helpfulness, IO prompting and iterative refinement perform better without demonstrations, while HypoGenic demonstrates improved performance with them. For Novelty, iterative refinement excels in the absence of demonstrations, whereas IO prompting and HypoGenic exhibit minimal differences between the two settings.

Format	Correct Label		Flipped Label	
	Best	Average	Best	Average
Label Format1	68.56	62.72	65.15	59.96
Label Format2	67.88	62.78	67.49	61.90
w/o demos	Best: 68.62		Average: 62.12	

Table 5: Accuracy comparison of different label formats in correct and flipped label settings with IO-prompting. Each number is the average over five datasets.

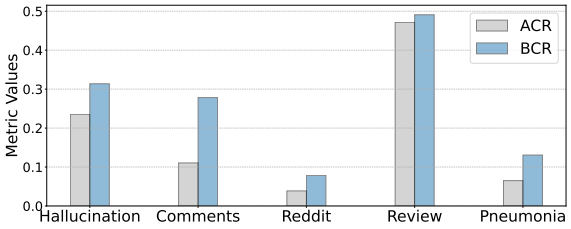

Figure 5: Difference of predictions between correct label and flipped label demos: *Adverse Correction Rate (ACR)* and *Beneficial Correction Rate (BCR)* values under multiple hypotheses-based classification.

These results highlight that LLMs can produce highly helpful and novel hypotheses even without in-context demonstrations.

5.4 Human Evaluation Results

We conduct a human evaluation on Funny Reddit, Truthful Reviews, and Unhealthy Comments datasets, as the other datasets require more specialized expertise. The results are illustrated in Figure 4. Across the three datasets, hypotheses generated without demonstrations received the highest percentage of preference. These findings indicate a slight overall preference for hypotheses generated using only the model's prior, though the extent of this preference varies by dataset.

6 Analysis

6.1 Is the result consistent with different in-context demonstration label formats?

To evaluate the consistency of results across different label formats, we compare two label formats: *Label Format 1:* Demonstrations are provided as examples for positive and negative classes as in Figure 1. *Label Format 2:* Demonstrations are presented in the format of *(Example, Label)*.

The average accuracy across all datasets for the correct and flipped label settings is presented in

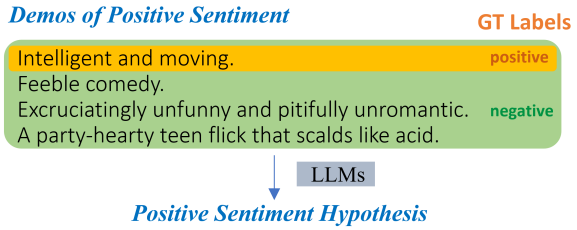

Figure 6: An illustration of the case study: positive sentiment hypothesis generation. The highlighted text with a green background represents flipped label demos.

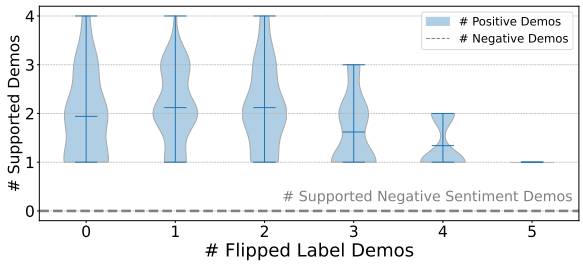

Figure 7: Distribution of the number of supported true positive and negative demos with different number of flipped label demos.

Figure 5. With correct labels, the performance of the two label formats is very similar. However, in the flipped label settings, *Label Format 2* shows almost no performance drop, which differs slightly from *Label Format 1*. Notably, neither label format outperforms the hypotheses generated without demonstrations. This finding highlights the dominant role of the strong model prior, regardless of the presentation style of the demonstrations.

6.2 What's the difference between correct label and flipped label settings?

To get an deep understanding for the impact of flipping labels and provide a more fine-grained evaluation, we adopt two additional metrics introduced by Wu et al. (2024), Adverse Correction Rate (ACR) and Beneficial Correction Rate (BCR):

$$\text{ACR} = \frac{\sum_{i=1}^{n} \mathbb{I}\left(y_{\text{correct}}(x_i) = y_i \wedge y_{\text{flipped}}(x_i) \neq y_i\right)}{\sum_{i=1}^{n} \mathbb{I}\left(y_{\text{correct}}(x_i) = y_i\right)}, \quad (1)$$

$$\text{BCR} = \frac{\sum_{i=1}^{n} \mathbb{I}\left(y_{\text{correct}}(x_i) \neq y_i \wedge y_{\text{flipped}}(x_i) = y_i\right)}{\sum_{i=1}^{n} \mathbb{I}\left(y_{\text{correct}}(x_i) \neq y_i\right)}, \quad (2)$$

where $y_{\text{correct}}(x_i)$ and $y_{\text{flipped}}(x_i)$ represents the prediction results using the hypothesis generated with ground truth label and flipped label demonstrations, x_i, y_i are input and ground truth label, respectively. These metrics offer a comprehensive evaluation of how flipping labels of the demonstrations influence the prediction results in downstream tasks.

Results for multiple hypothesis-based classification prediction difference are shown in Table 5. The results indicate that flipping the labels of in-context demonstrations does lead to some shifts in prediction outcomes, particularly notable in the truthful hotel review dataset, where nearly half of the predictions are affected. In contrast, for the other four datasets, label flipping only minimally alters prediction results. This suggests that while the model leverages the input-label mappings in provided demonstrations to inform its hypothesis generation, the inherent task-specific knowledge remains predominant, preventing the provided patterns from overriding its established priors.

6.3 A Case Study: Hypothesis Generation for Positive Sentiment Pattern

This case study highlights that large language models (LLMs) heavily rely on prior knowledge when generating hypotheses, often ignoring patterns introduced in demonstrations. As shown in Figure 6, we replace true positive demonstrations with flipped label demonstrations (negative examples) to test whether the model adjusts its hypothesis or adheres to its prior.

Using IO-prompting, we provide six demonstrations, varying the number of flipped label demos from 0 to 5, and prompt the model to generate a hypothesis and corresponding supporting demonstrations. Repeating the experiment across 50 random seeds, we track the distribution of true positive and negative examples within the model's supported demonstrations for its hypothesis.

The results, shown in Figure 7, reveal notable patterns. The distribution of positive examples in the supported demonstrations begins to shift when three flipped label demonstrations are introduced. When five flipped demonstrations are provided, the mean number of positive examples converges to one. However, the model consistently avoids using flipped label demonstrations in its hypothesis generation, even when five demonstrations are flipped. This indicates that the model's hypotheses are predominantly influenced by prior knowledge rather than the provided demonstrations.

7 Conclusion

In this paper, we explore the role of task-specific priors in a real-world inductive reasoning sce-

nario—hypothesis generation from labeled data. Experiments reveal that LLMs rely heavily on strong priors, which are difficult to override with demonstrations, offering insights into hypothesis generation mechanisms and future research directions.

Limitations

Beyond Classification Problems. Our experiments are limited to classification problems. Extensions to multi-choice or other tasks requires better representation of the hypothesis. We leave extensions to non-classification tasks for future work.

Better Application of Generated Hypotheses. We think future can explore better application of generated hypotheses. For instance, this paper uses hypotheses to construct patterns for classification problems. Better application of hypotheses can improve downstream task performance, which we leave for future work.

Model Dependence and Generalization. We rely heavily on proprietary LLMs such as GPT-4 and Gemini for hypothesis generation. These models carry strong pretrained priors, which may overshadow the effect of demonstrations. It remains an open question whether smaller or less capable models, where prior knowledge is limited, might benefit more from explicit demonstrations. Additionally, future work may explore finetuning or continual learning approaches to better align model behavior with task-specific goals.

References

Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. 2024. Researchagent: Iterative research idea generation over scientific literature with large language models. *arXiv preprint arXiv:2404.07738*.

Chen Bowen, Rune Sætre, and Yusuke Miyao. 2024. A comprehensive evaluation of inductive reasoning capabilities and problem solving in large language models. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 323–339.

Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Kewei Cheng, Jingfeng Yang, Haoming Jiang, Zhengyang Wang, Binxuan Huang, Ruirui Li, Shiyang Li, Zheng Li, Yifan Gao, Xian Li, et al. 2024. Inductive or deductive? rethinking the fundamental reasoning abilities of llms. *arXiv preprint arXiv:2408.00114*.

Kevin Du, Vésteinn Snæbjarnarson, Niklas Stoehr, Jennifer C White, Aaron Schein, and Ryan Cotterell. 2024. Context versus prior knowledge in language models. *arXiv preprint arXiv:2404.04633*.

Lisa Dunlap, Yuhui Zhang, Xiaohan Wang, Ruiqi Zhong, Trevor Darrell, Jacob Steinhardt, Joseph E Gonzalez, and Serena Yeung-Levy. 2024. Describing differences in image sets with natural language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24199–24208.

Evan Heit. 2000. Properties of inductive reasoning. *Psychonomic bulletin & review*, 7:569–592.

Or Honovich, Uri Shaham, Samuel R Bowman, and Omer Levy. 2022. Instruction induction: From few examples to natural language task descriptions. *arXiv preprint arXiv:2205.10782*.

Younghyun Kim, Sangwoo Mo, Minkyu Kim, Kyungmin Lee, Jaeho Lee, and Jinwoo Shin. 2024. Discovering and mitigating visual biases through keyword explanation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11082–11092.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Jiachun Li, Pengfei Cao, Zhuoran Jin, Yubo Chen, Kang Liu, and Jun Zhao. 2024. Mirage: Evaluating and explaining inductive reasoning process in language models. *arXiv preprint arXiv:2410.09542*.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.

Haokun Liu, Yangqiaoyu Zhou, Mingxuan Li, Chenfei Yuan, and Chenhao Tan. 2024. Literature meets data: A synergistic approach to hypothesis generation. *arXiv preprint arXiv:2410.17309*.

Bodhisattwa Prasad Majumder, Harshit Surana, Dhruv Agarwal, Sanchaita Hazra, Ashish Sabharwal, and Peter Clark. 2024a. Data-driven discovery with large generative models. *arXiv preprint arXiv:2402.13610*.

Bodhisattwa Prasad Majumder, Harshit Surana, Dhruv Agarwal, Bhavana Dalvi Mishra, Abhijeetsingh Meena, Aryan Prakhar, Tirth Vora, Tushar Khot, Ashish Sabharwal, and Peter Clark. 2024b. Discoverybench: Towards data-driven discovery with large language models. *arXiv preprint arXiv:2407.01725*.

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

Chau Minh Pham, Alexander Hoyle, Simeng Sun, Philip Resnik, and Mohit Iyyer. 2023. Topicgpt: A prompt-based topic modeling framework. *arXiv preprint arXiv:2311.01449*.

Ilan Price, Jordan Gifford-Moore, Jory Fleming, Saul Musker, Maayan Roichman, Guillaume Sylvain, Nithum Thain, Lucas Dixon, and Jeffrey Sorensen. 2020. Six attributes of unhealthy conversation. *arXiv preprint arXiv:2010.07410*.

Biqing Qi, Kaiyan Zhang, Haoxiang Li, Kai Tian, Sihang Zeng, Zhang-Ren Chen, and Bowen Zhou. 2023. Large language models are zero shot hypothesis proposers. *arXiv preprint arXiv:2311.05965*.

Linlu Qiu, Liwei Jiang, Ximing Lu, Melanie Sclar, Valentina Pyatkin, Chandra Bhagavatula, Bailin Wang, Yoon Kim, Yejin Choi, Nouha Dziri, et al. 2023. Phenomenal yet puzzling: Testing inductive reasoning capabilities of language models with hypothesis refinement. *arXiv preprint arXiv:2310.08559*.

Parshin Shojaee, Kazem Meidani, Shashank Gupta, Amir Barati Farimani, and Chandan K Reddy. 2024. Llm-sr: Scientific equation discovery via programming with large language models. *arXiv preprint arXiv:2404.18400*.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. 2024. Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers. *arXiv preprint arXiv:2409.04109*.

Suzanna Sia, David Mueller, and Kevin Duh. 2024. Where does in-context learning happen in large language models? *Advances in Neural Information Processing Systems*, 37:32761–32786.

Chandan Singh, John X Morris, Jyoti Aneja, Alexander M Rush, and Jianfeng Gao. 2022. iprompt: Explaining data patterns in natural language via interpretable autoprompting. *ArXiv preprint*, 2210.

Ruocheng Wang, Eric Zelikman, Gabriel Poesia, Yewen Pu, Nick Haber, and Noah D Goodman. 2023. Hypothesis search: Inductive reasoning with language models. *arXiv preprint arXiv:2309.05660*.

Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, et al. 2023. Larger language models do in-context learning differently. *arXiv preprint arXiv:2303.03846*.

Yurong Wu, Yan Gao, Bin Benjamin Zhu, Zineng Zhou, Xiaodi Sun, Sheng Yang, Jian-Guang Lou, Zhiming Ding, and Linjun Yang. 2024. Strago: Harnessing strategic guidance for prompt optimization. *arXiv preprint arXiv:2410.08601*.

Tim Z Xiao, Robert Bamler, Bernhard Schölkopf, and Weiyang Liu. 2024. Verbalized machine learning: Revisiting machine learning with language models. *arXiv preprint arXiv:2406.04344*.

Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. 2023a. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41.

Zonglin Yang, Li Dong, Xinya Du, Hao Cheng, Erik Cambria, Xiaodong Liu, Jianfeng Gao, and Furu Wei. 2022. Language models as inductive reasoners. *arXiv preprint arXiv:2212.10923*.

Zonglin Yang, Xinya Du, Junxian Li, Jie Zheng, Soujanya Poria, and Erik Cambria. 2023b. Large language models for automated open-domain scientific hypotheses discovery. *arXiv preprint arXiv:2309.02726*.

Ruiqi Zhong, Charlie Snell, Dan Klein, and Jacob Steinhardt. 2022. Describing differences between text distributions with natural language. In *International Conference on Machine Learning*, pages 27099–27116. PMLR.

Ruiqi Zhong, Peter Zhang, Steve Li, Jinwoo Ahn, Dan Klein, and Jacob Steinhardt. 2023. Goal driven discovery of distributional differences via language descriptions. *Advances in Neural Information Processing Systems*, 36:40204–40237.

Yangqiaoyu Zhou, Haokun Liu, Tejes Srivastava, Hongyuan Mei, and Chenhao Tan. 2024. Hypothesis generation with large language models. *arXiv preprint arXiv:2404.04326*.

Zhaocheng Zhu, Yuan Xue, Xinyun Chen, Denny Zhou, Jian Tang, Dale Schuurmans, and Hanjun Dai. 2023. Large language models can learn rules. *arXiv preprint arXiv:2310.07064*.

A Dataset Details

In this paper, we include 5 real-world datasets: hallucination, unhealthy comments in conversation, truthful hotel review, pneumonia MNIST and funny reddit post.

Dataset	Train	Validation	Test
Hallucination	400	100	374
Pneumonia MNIST	800	270	468
Unhealthy Conversation	800	400	800
Funny Reddit	200	100	308
Truthful Hotel Review	800	300	500

Table 6: Dataset Split for Train, Validation, and Test Sets.

Hallucination Pattern. The dataset is first introduced in (Li et al., 2023). We use its adversarial sampling version, which can be found in https://github.com/RUCAIBox/POPE. To build our hallucination dataset, we prompt GPT-4o with each image-question pair once and see if the model hallucinates the object presence. As a result, we get 437 hallucinated image-question pairs and randomly sample another 437 image-question pairs as non-hallucination cases.

Unhealthy Comments. Expert-annotated unhealthy conversations are from (Price et al., 2020), and we use the version from (Zhong et al., 2023), which can be downloaded from https://github.com/ruiqi-zhong/D5. We sample longest 1000 samples for unhealthy and healthy comments from the dataset in our evaluation.

Truthful Hotel Reviews. Truthful review detection is an instance of deception. The dataset we use is from (Zhou et al., 2024). The dataset includes 800 genuine reviews and 800 fictitious reviews for 20 hotels in Chicago, which can be downloaded from https://github.com/ChicagoHAI/hypothesis-generation.

Funny Reddit Posts. We collect jokes posted on the Reddit forum r/Jokes and cleaned by (Zhong et al., 2023). This dataset can be downloaded from https://github.com/ruiqi-zhong/D5. We also remove all the duplicate samples for better dataset quality.

Pneumonia MNIST. Pneumonia recognition via chest X-ray image is an important problem. The dataset is from (Yang et al., 2023a), and can be downloaded from https://medmnist.com/.

For each dataset, we have at least 200 samples for training, 100 samples for validation and 300 samples for test. For each dataset, we keep a balance between positive and negative class. Detailed statistics is shown in Table 6.

B Implementation Details

Model Parameters. For API usage, the temperature and top-p are set to a small number 1×10^{-15} and 1×10^{-10}, respectively. Our choices were made to encourage deterministic responses, ensuring the reproducibility of results across multiple runs. Lower temperature and controlled top-p values help minimize randomness, allowing for a more stable evaluation.

Iterative Refinement. We initialize the hypothesis bank with 5 hypotheses generated using IO-prompting. In refinement process, for each iteration, we select 5 hypotheses achieving highest accuracy on the validation set to LLMs for refinement and hope to get hypothesis with better quality. We evaluate 5 hypotheses with the best performance on validation dataset. We set refinement iteration to 3 in the paper.

HypoGeniC. We set the hypothesis bank size to 5. Throughout the experiment, we use the reward efficient $\alpha = 0.5$, the number of initialized examples $num\_init = 10$, and maximum number of wrong examples for each group to 2 for more updates. For each iteration, we select top 3 hypotheses to evaluate. For each update, we generate 1 new hypothesis with incorrect examples. When there are no demonstrations, we rank the hypotheses in the bank by reward scores and use this ranking as feedback to get better hypothesis.

C Additional Results

C.1 Does the model rely on generated hypothesis for downstream classification?

To verify whether LLMs truly follow the in-context hypothesis for downstream classification, or whether they rely primarily on their internal prior knowledge, we adopt three complementary evaluation strategies.

C.1.1 Zero-shot Prompting Performance

We first perform zero-shot prompting as a baseline to roughly estimate the model's inherent prior

Method	Hallucination	PneumoniaMNIST	Truthful Review	Unhealthy Comments	Funny Reddit
Zero-shot	57.8	56.0	51.4	67.5	58.8
Best Hypothesis	**66.9**	**77.6**	**69.0**	**71.4**	**67.0**

Table 7: Accuracy comparison between zero-shot prompting and best hypothesis-based classification across five datasets.

	Demos	Hallucination	Unhealthy Comments	Funny Reddit	Truthful Review	PneumoniaMNIST	Overall Average
Best	w/o	63.1	70.1	61.6	64.0	75.6	66.9
	w/	57.5	68.0	59.1	64.6	80.8	66.0
Average	w/o	54.4	60.3	54.1	56.7	69.8	59.1
	w/	53.6	63.3	54.8	51.8	73.1	59.3

Table 8: Accuracy comparison of *single hypothesis-based classification* without task-specific knowledge in inference: accuracy for the single hypothesis and the average across five hypotheses, with (**w/**) and without (**w/o**) demonstrations.

knowledge about the task. No hypothesis is provided in this setting. As shown in Table 7, the zero-shot accuracy is consistently lower than the accuracy achieved using the best hypothesis, indicating that LLMs can leverage hypotheses to improve task performance beyond their prior.

C.1.2 Hypothesis-based Inference without Task-Specific Knowledge

To minimize the impact of prior knowledge in hypothesis-based inference, we eliminate task-specific knowledge from the evaluation prompt and remove learned patterns from the hypothesis. Instead, we reformulate the task into its corresponding modalities, prompting large language models (LLMs) with: *"Does the provided text/image/image-question align with the given text/image/image-question patterns?"* This approach isolates the quality of the hypothesis, ensuring that inference is not influenced by prior knowledge.

The results are shown as Table 8. On average, there is limited difference between the hypotheses generated with and without demonstrations. The findings demonstrate again that LLMs are able to generate hypothesis with high quality only with task-specific prior.

C.1.3 Ablation: Accuracy under Invalid Hypotheses

To further verify whether the model is truly following the hypothesis or merely relying on its prior, we introduce systematically constructed invalid hypotheses across four types: (1) **Semantically Misaligned**, which include irrelevant content not grounded in task semantics; (2) **Over-general**, which are vacuous or trivially true; (3) **Superficial Feature-Based**, which depend on shallow surface patterns like punctuation or length; and (4) **Misleading**, which appear reasonable but wrong, even with inverted logic.

As shown in Table 9, these hypotheses degrade model performance to near-random or even below-random levels, confirming that LLMs do follow the structure and logic of the hypothesis—even when it is deceptive or invalid.

C.2 Results of Different Datasets with Label Format 2

We provide results on each dataset with *Label Format 2*. The results are shown as Table 10. From the results, we can see that the results vary by dataset. However, there is quite limited difference (smaller than 3%) between correct and flipped label settings, showing the prior is too strong to be overridden by provided demonstrations.

C.3 Results with Qwen and Gemini Model

We test IO-prompting with and without demonstrations on model **Qwen2-VL-72B** and **gemini-1.5-pro-002**. We report the average over different random seeds. The results are shown as Table 11 and 12. On average, there is quite limited performance difference with and without demonstrations, demonstrating that with only prior, LLMs can generate good hypotheses.

C.4 The Impact of the Number of Demonstrations

We study the effect of the number of demonstrations on two datasets: a text dataset (Funny Reddit) and a multi-modal dataset (PneumoniaMNIST). For each task, we use IO-prompting to generate five

Type	Dataset	Hypothesis	Acc (%)
Semantically Irrelevant	Hallucination	If the image contains trees and the question mentions music, the model will hallucinate.	50.0
	PneumoniaMNIST	If the X-ray image contains a bright top-left corner, then it is pneumonia.	50.0
	Truthful Review	If the review mentions animals and musical instruments, then it is truthful.	50.2
	Funny Reddit	If the post includes colors and geographical names, then it is funny.	50.0
	Unhealthy Comment	If the comment contains numbers and food items, then it is unhealthy.	49.5
Over-general	Hallucination	If there is an image and a question, the model will hallucinate.	50.0
	PneumoniaMNIST	If the image contains pixels, then it is pneumonia.	50.0
	Truthful Review	If the review is written in English, then it is truthful.	50.0
	Funny Reddit	If the post contains at least one sentence, then it is funny.	51.1
	Unhealthy Comment	If the comment includes any text, then it is unhealthy.	50.0
Superficial Feature	Hallucination	If the question contains more than 15 words, the model will hallucinate.	50.0
	PneumoniaMNIST	If the top-left pixel is not black, then it is pneumonia.	50.6
	Truthful Review	If the review contains more than two commas, then it is truthful.	53.6
	Funny Reddit	If the post ends with a question mark, then it is funny.	44.8
	Unhealthy Comment	If the comment has more than 3 exclamation marks, then it is unhealthy.	50.6
Misleading	Hallucination	If the asked object is the main focus of the image, the model will hallucinate.	42.7
	PneumoniaMNIST	If the lung fields are symmetric and occupy more than 70% of the image width, then it is pneumonia.	48.2
	Truthful Review	If the review contains only positive sentiment words, then it is truthful.	41.6
	Funny Reddit	If the post has a question format, then it is funny.	45.1
	Unhealthy Comment	If the comment uses hedging language, then it is unhealthy.	46.2

Table 9: Accuracy under intentionally invalid hypotheses to assess hypothesis dependence.

Label	Hallucination	Unhealthy Comments	Funny Reddit	Truthful Review	PneumoniaMNIST	Average
Correct (Best)	63.9	70.6	61.7	68.0	75.2	67.9
Flipped (Best)	61.2	71.5	62.0	68.8	73.9	67.5
Correct (Avg)	57.0	65.1	59.0	62.5	70.3	62.8
Flipped (Avg)	57.8	64.7	57.3	61.3	68.5	61.9

Table 10: Accuracy comparison across five datasets with correct and flipped label settings in the *Label Format 2*.

hypotheses. Since LLMs often struggle with long contexts—leading to issues such as the "Lost-in-the-Middle" effect—we vary the number of demonstrations from 10 to 100.

Figure 8 reports the accuracy of the best single hypothesis and the average accuracy across five hypotheses. We observe that even with only ten demonstrations, LLMs can generate strong hypotheses. Increasing the number of demonstrations from 10 to 100 does not improve performance and may even degrade it. Overall, the number of demonstrations has minimal impact on hypothesis quality, reinforcing the conclusion that LLMs primarily rely on their prior knowledge rather than in-context examples.

D Evaluation Details

LLM-based Evaluation Details. We prompt large language models (LLMs) to generate five hypotheses for each dataset across three different baselines. This results in a total of 25 hypotheses per baseline for both settings: with and without demonstrations.

For LLM-based scoring, each hypothesis is evaluated by prompting the LLMs to assign a score on a 1–5 scale. Additionally, for pairwise comparisons,

	Demos	Hallucination	Unhealthy Comments	Funny Reddit	Truthful Review	PneumoniaMNIST	Overall Average
Best	w/o	60.4 ± 0.0	68.5 ± 0.0	63.6 ± 0.0	67.0 ± 0.0	65.4 ± 0.0	**64.98**
	w/	60.1 ± 2.3	68.0 ± 0.0	62.4 ± 2.2	66.0 ± 0.4	62.5 ± 2.9	63.80
Average	w/o	57.7 ± 0.0	63.0 ± 0.0	57.1 ± 0.0	55.1 ± 0.0	57.9 ± 0.0	**58.16**
	w/	55.4 ± 1.1	63.4 ± 0.2	58.2 ± 0.8	56.5 ± 1.1	54.7 ± 2.0	57.64

Table 11: Accuracy comparison of *single hypothesis-based classification* with **Qwen2-VL-72B**: accuracy (*mean* ± *standard deviation*) for the best single hypothesis and the average across five hypotheses, with (**w/**) and without (**w/o**) demonstrations. The better overall average between (**w/**) and (**w/o**) is highlighted in **bold**.

	Demos	Hallucination	Unhealthy Comments	Funny Reddit	Truthful Review	PneumoniaMNIST	Overall Average
Best	w/o	-	67.9 ± 0.2	62.7 ± 0.3	68.8 ± 2.0	58.2 ± 1.8	**64.40**
	w/	-	67.8 ± 1.3	65.9 ± 0.3	66.9 ± 1.7	55.7 ± 1.4	64.08
Average	w/o	-	61.9 ± 0.3	56.8 ± 0.0	64.5 ± 2.3	53.1 ± 0.2	59.08
	w/	-	62.4 ± 1.1	58.0 ± 1.2	63.4 ± 1.2	53.0 ± 1.5	**59.20**

Table 12: Accuracy comparison of *single hypothesis-based classification* with **Gemini-1.5-pro-002**: accuracy (*mean* ± *standard deviation*) for the best single hypothesis and the average across five hypotheses, with (**w/**) and without (**w/o**) demonstrations. "-" means the response is prohibited due to safety reasons.

we randomly pair hypotheses generated with and without demonstrations, creating a total of 25 pairs for evaluation.

Human Evaluation Details. We randomly pair the hypotheses generated with and without demonstrations across three datasets and three baselines. We selected the datasets unhealthy comments, truthful reviews, and funny Reddit posts because their domain knowledge is accessible to non-experts.

Participants were provided with a questionnaire for evaluation. For each evaluation, we included the evaluation context, paired hypotheses, and illustrative examples to guide participants. An example of the evaluation interface is shown in Figure 9.

E Examples of Generated Hypothesis

We randomly select generated hypothesis with and without demonstrations for each dataset, shown as Table 13.

F Prompts

For prompt construction, we begin by manually crafting a prompt for hallucination pattern induction, following a format similar to that used in (Zhou et al., 2024). Subsequently, we leverage in-context learning to generate prompts for other tasks. Specifically, we provide the task name along with the manually constructed prompt to the language model, enabling it to generate prompts tailored to other tasks. Some texts are reformatted for presentation.

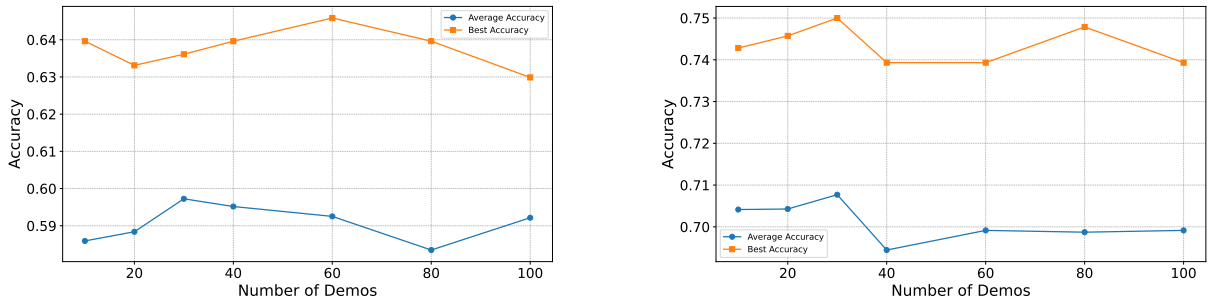

Figure 8: Accuracy comparison of *single hypothesis-based classification*: accuracy for the best single hypothesis and the average across five hypotheses, evaluated with varying numbers of demonstrations. Left: Funny Reddit dataset. Right: PneumoniaMNIST dataset.

Dataset	**Hypothesis without Demos**	**Hypothesis with Demos**
Hallucination	Hallucinations are more likely to occur when the questioned object is partially occluded or located in a cluttered environment, making it difficult for the model to accurately identify its presence or absence.	**Complex Backgrounds Hypothesis**: Images with complex or cluttered backgrounds may lead to hallucinations, as the model might misinterpret overlapping or densely packed objects as the queried item.
Unhealthy Comments	Comments containing personal attacks or insults are more likely to be unhealthy, as they often escalate conflicts and discourage constructive dialogue.	Comments that include personal attacks or derogatory language towards individuals are more likely to be unhealthy.
Funny Reddit Posts	Posts that incorporate unexpected punchlines or twists are more likely to be perceived as funny, as they play on the element of surprise and subvert reader expectations.	Posts that use wordplay or double entendres, where a phrase can be interpreted in multiple humorous ways, tend to be perceived as funny.
Pneumonia MNIST	The presence of pleural effusion, seen as blunting of the costophrenic angles or fluid layering in the pleural space, may indicate pneumonia.	Presence of air bronchograms within areas of increased opacity suggests pneumonia.
Truthful Hotel Reviews	Truthful reviews often mention both positive and negative aspects of the stay, providing a balanced perspective rather than an overly positive or negative one.	Truthful reviews often mention both positive and negative aspects of the stay, providing a balanced perspective that suggests authenticity.

Table 13: Examples of Generated Hypotheses with and without In-Context Demonstrations.

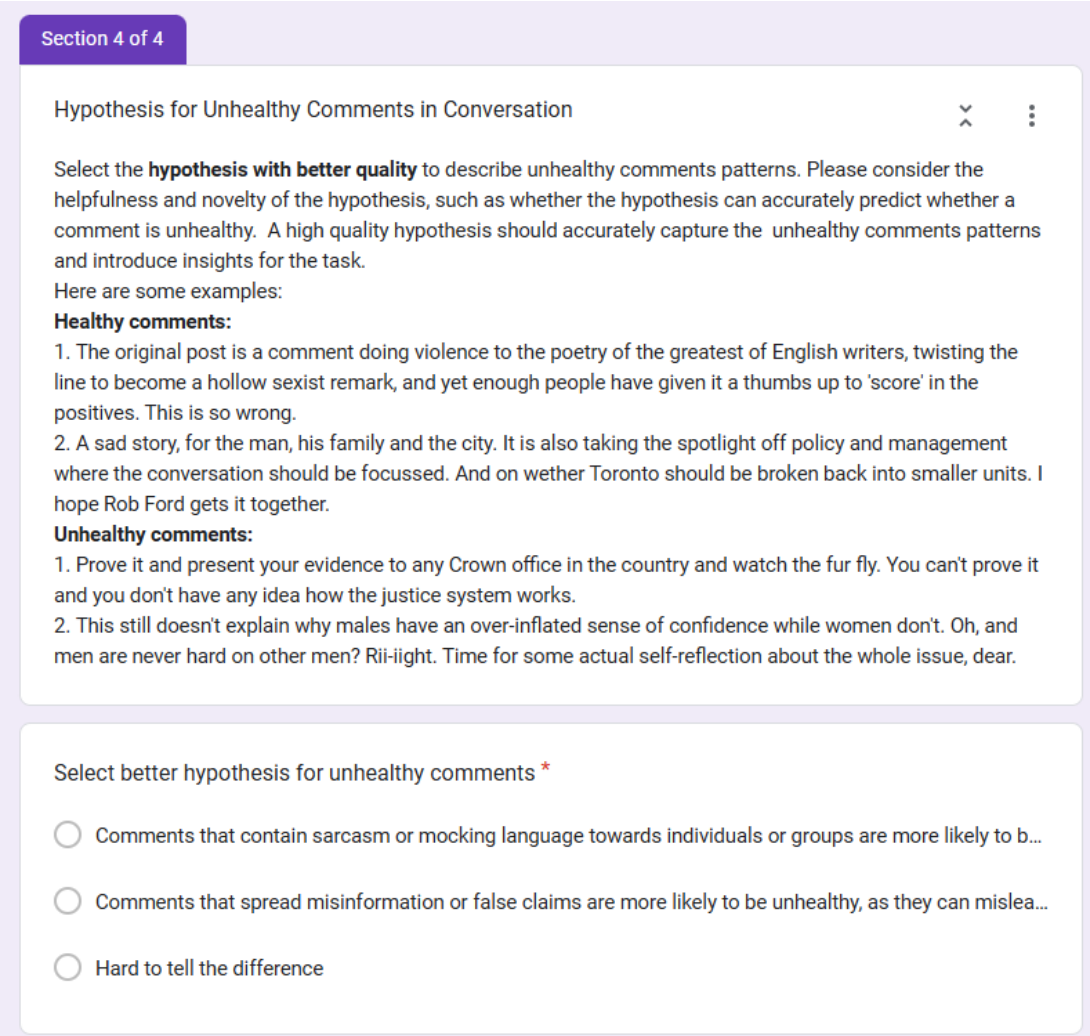

Figure 9: Example interface of human evaluation.

Prompt for hallucination with demonstrations

You're a professional vision-language model behavior analyst.
Given a set of image-question pairs, we want to generate hypotheses that are useful for predicting whether a model will hallucinate the existence of an object in response to a given question.
In other words, we want to know whether the model will falsely claim the presence of an object in the image when answering the question.
Using the given examples, please propose {{num_hypotheses}} possible hypotheses that can identify specific patterns that occur across the provided image-question pairs.
Each hypothesis should contain the following: a hypothesis about what image content features, object features, or contextual relationships make the model more likely to hallucinate.
The hypotheses should analyze what kinds of image-question pairs are more likely to trigger hallucinations.

Some examples of hallucination and non-hallucination cases are shown.
Hallucination cases:
{{positive_examples}}
Non-hallucination cases:
{{negative_examples}}

Based on provided examples, please generate hypotheses that are useful for predicting whether the model will hallucinate the existence of an object in response to a given question.
Propose {{num_hypotheses}} possible hypotheses for hallucination patterns.
Generate them in the format of 1. [hypothesis], 2. [hypothesis], ... {{num_hypotheses}}. [hypothesis].
Proposed hypotheses:

Prompt for hallucination without demonstrations

You are an expert in vision-language models, specializing in detecting and preventing hallucinations.
We want to generate hypotheses that are useful for predicting whether a vision-language model will hallucinate the existence of an object when responding to a question about an image.
In other words, we want to identify patterns that indicate when the model will incorrectly claim the presence of an object not present in the image, or the absence of an object that is present.
Please propose {{num_hypotheses}} possible hypotheses.
These hypotheses should identify specific patterns that occur across common hallucination cases and focus on the relationship between the image content and the questioned object.
Each hypothesis should contain the following: a hypothesis about what image content features, object features, or contextual relationships make the model more likely to hallucinate.
The hypotheses should analyze what kind of image-question pairs are more likely to lead to hallucinations.
Please generate {{num_hypotheses}} possible hypotheses for hallucination patterns in the given context.
Generate them in the format of 1. [hypothesis], 2. [hypothesis], ... {{num_hypotheses}}. [hypothesis].
Don't talk about any other words.
Proposed hypotheses:

Prompt for unhealthy comments with demonstrations

You're an expert comment analyst in online conversation.
Given a set of comments, we want to generate hypotheses that are useful for predicting whether a comment is unhealthy.
In other words, we want to know if the comment contributes to unhealthy conversations online.
Using the given examples, please propose {{num_hypotheses}} possible hypotheses.
These hypotheses should identify specific patterns that occur across the provided unhealthy comments. Each hypothesis should contain the following: A hypothesis about what makes comments more likely to be unhealthy. The hypotheses should analyze what kind of comments are likely to be unhealthy.
Here are some examples of unhealthy and healthy comments:

Unhealthy comments:
{{positive_examples}}
Healthy comments:
{{negative_examples}}

Based on the provided examples, please generate hypotheses that are useful for predicting whether a comment is unhealthy.
Propose {{num_hypotheses}} possible hypotheses for unhealthy comment patterns.
Generate them in the format of 1. [hypothesis], 2. [hypothesis], ... {{num_hypotheses}}. [hypothesis].
Don't include any other words.
Proposed hypotheses:

Prompt for unhealthy comments without demonstrations

You're an expert comment analyst in online conversation.
We want to generate hypotheses that are useful for predicting whether a comment is unhealthy. In other words, we want to know if the comment contributes to unhealthy conversations online.
Please propose {{num_hypotheses}} possible hypotheses.
These hypotheses should identify specific patterns that occur across common unhealthy comments.
Each hypothesis should contain the following: A hypothesis about what makes comments more likely to be unhealthy.
The hypotheses should analyze what kind of comments are likely to be unhealthy.
Please generate hypotheses that are useful for predicting whether a comment is unhealthy or healthy.
Propose {{num_hypotheses}} possible hypotheses for unhealthy comment patterns.
Generate them in the format of 1. [hypothesis], 2. [hypothesis], ... {{num_hypotheses}}. [hypothesis].
Don't talk about any other words.
Proposed hypotheses:

Prompt for truthful reviews with demonstrations

You're a professional hotel review analyst.
Given a set of hotel reviews, we want to generate hypotheses that are useful for predicting whether a review is truthful. In other words, we want to know whether the review is written by someone who actually lived in the hotel.
Using the given examples, please propose {{num_hypotheses}} possible hypotheses.
These hypotheses should identify specific patterns that occur across the provided reviews. Each hypothesis should contain the following: A hypothesis about what makes reviews more likely to be truthful. The hypotheses should analyze what kind of reviews are likely to be truthful.
Here are some examples of truthful and deceptive reviews:

Truthful reviews:
{{positive_examples}}
Deceptive reviews:
{{negative_examples}}

Based on provided examples, please generate hypotheses that are useful for predicting whether a review is truthful.
Propose {{num_hypotheses}} possible hypotheses for truthful review patterns.
Generate them in the format of 1. [hypothesis], 2. [hypothesis], ... {{num_hypotheses}}. [hypothesis].
Don't talk about any other words.
Proposed hypotheses:

Prompt for truthful reviews with demonstrations

You're a professional hotel review analyst.
We want to generate hypotheses that are useful for predicting whether a review is truthful or deceptive. In other words, we want to know whether the review is written by someone who actually lived in the hotel.
Please propose {{num_hypotheses}} possible hypotheses.
These hypotheses should identify specific patterns that occur across common truthful reviews. Each hypothesis should contain the following: A hypothesis about what makes reviews more likely to be truthful. The hypotheses should analyze what kind of reviews are likely to be truthful or deceptive.
Please generate hypotheses that are useful for predicting whether a review is truthful or deceptive.
Propose {{num_hypotheses}} possible hypotheses for truthful review patterns.
Generate them in the format of 1. [hypothesis], 2. [hypothesis], ... {{num_hypotheses}}. [hypothesis].
Don't talk about any other words.
Proposed hypotheses:

Prompt for PneumoniaMNIST with demonstrations

You're a professional radiologist specializing in chest X-rays.
Given a set of labeled chest X-ray images, we want to generate hypotheses that are useful for predicting whether a patient has pneumonia. In other words, we want to know whether the X-ray shows signs of pneumonia.
Using the given examples, please propose {{num_hypotheses}} possible hypotheses.
These hypotheses should identify specific patterns that occur across the provided X-ray images.
Each hypothesis should contain the following: A hypothesis about what makes an X-ray more likely to indicate pneumonia. The hypotheses should analyze what kind of image patterns are likely to be indicative of pneumonia or not.

Some examples of X-ray images labeled as pneumonia and non-pneumonia are shown.
Pneumonia cases:
{{positive_examples}}
Non-pneumonia cases:
{{negative_examples}}

Based on provided examples, please generate hypotheses that are useful for predicting whether an X-ray shows pneumonia or not.
Propose {{num_hypotheses}} possible hypotheses for pneumonia pattern recognition.
Generate them in the format of 1. [hypothesis], 2. [hypothesis], ... {{num_hypotheses}}. [hypothesis].
Don't include any other information.
Proposed hypotheses:

Prompt for PneumoniaMNIST without demonstrations

You're a professional radiologist.
We want to generate hypotheses that are useful for predicting whether a patient has pneumonia based on their chest X-ray image. In other words, we want to know which patterns in the image are indicative of pneumonia presence.
Please propose {{num_hypotheses}} possible hypotheses.
These hypotheses should identify specific visual patterns that occur in typical pneumonia cases.
Each hypothesis should contain the following: A hypothesis about what makes an image more likely to show signs of pneumonia.
The hypotheses should analyze what kind of visual patterns or markers are likely to indicate pneumonia.
Please generate hypotheses that are useful for predicting whether a patient has pneumonia or not based on the X-ray.
Propose {{num_hypotheses}} possible hypotheses for pneumonia-related visual patterns.
Generate them in the format of 1. [hypothesis], 2. [hypothesis], ... {{num_hypotheses}}. [hypothesis].
Don't include any additional context.
Proposed hypotheses:

Prompt for funny reddit with demonstrations

You're a professional humor analyst for Reddit posts.
Given a set of Reddit posts, we want to generate hypotheses that are useful for predicting whether a post is considered funny or not. In other words, we want to know whether a post contains humor patterns often associated with successful humorous posts.
Using the provided examples, please propose {{num_hypotheses}} possible hypotheses.
These hypotheses should identify specific patterns that occur across the provided posts.
Each hypothesis should contain the following: A hypothesis about what makes posts more likely to be considered funny. The hypotheses should analyze what kind of posts are likely to be perceived as funny or not.
Here are some examples of funny and unfunny posts:

Funny posts:
{{positive_examples}}
Unfunny posts:
{{negative_examples}}

Based on the provided examples, please generate hypotheses that are useful for predicting whether a post is funny or not.
Propose {{num_hypotheses}} possible hypotheses for funny post patterns.
Generate them in the format of 1. [hypothesis], 2. [hypothesis], ... {{num_hypotheses}}. [hypothesis].
Don't talk about any other words.
Proposed hypotheses:

Prompt for funny reddit without demonstrations

You're a professional Reddit content analyst.
We want to generate hypotheses that are useful for predicting whether a Reddit post is funny or not. In other words, we want to know what characteristics make a post likely to be perceived as humorous by the community.
Please propose {{num_hypotheses}} possible hypotheses.
These hypotheses should identify specific patterns that occur across common funny posts.
Each hypothesis should contain the following: A hypothesis about what makes posts more likely to be perceived as funny.
The hypotheses should analyze what kind of posts are likely to be considered humorous or non-humorous.
Please generate hypotheses that are useful for predicting whether a post is funny or not.
Propose {{num_hypotheses}} possible hypotheses for funny Reddit post patterns.
Generate them in the format of 1. [hypothesis], 2. [hypothesis], ... {{num_hypotheses}}. [hypothesis].
Don't talk about any other words.
Proposed hypotheses:

Evaluation prompt for hallucination

You are an expert in vision-language model behavior detection, and your job is to apply learned patterns to predict if the model will hallucinate for the given image and question.
Here are some previously learned hallucination patterns:
{{patterns}}
The image is shown and the question is: {{text}}
Based on the learned patterns, will the model hallucinate?
Let's think step by step.
Step 1: Look at the given image and question, and compare them with the provided hallucination patterns.
Step 2: If the image and question pair matches learned hallucination pattern, then it is likely that the model will hallucinate. If it does not match the pattern, then it is likely that the model will not hallucinate.
Step 3: Apply the pattern to the given image-question pair and predict whether the model will hallucinate based on the learned patterns.
Step 4: Give your final answer: yes or no. If you are unsure, respond with no.
Please give your answer strictly in the following format:

```
"""
Analysis: [your step-by-step analysis]
Answer: [your answer]
"""
```

Give your answer in the above format. Don't talk about any other words.

Evaluation prompt for unhealthy comments

You will predict whether a comment is unhealthy based on the provided unhealthy comment patterns.
Here are some previously generated unhealthy comment patterns:
{{patterns}}
A comment is the following: {{text}}
Is this comment unhealthy?
Think step-by-step.
Step 1: Look at the new comment and compare it with the provided unhealthy comment patterns.
Step 2: If the comment matches the pattern, then it is likely unhealthy. If it does not match the pattern, then it is likely healthy.
Step 3: Apply the pattern to the new comment and predict whether the new comment is unhealthy.
Step 4: Give your final answer: yes or no. If you are unsure, respond with no.
Please give your answer strictly in the following format:

```
"""
Analysis: [your step-by-step analysis]
Answer: [your answer]
"""
```

Evaluation prompt for truthful review

You will predict whether a hotel review is truthful based on the given truthful review patterns.
Here are some previously generated truthful review patterns:
{{patterns}}
A hotel review is the following: {{text}}
Is this hotel review truthful?
Think step-by-step.
Step 1: Look at the new hotel review and compare it with the provided truthful review patterns.
Step 2: If the review matches the pattern, then it is likely truthful. If it does not match the pattern, then it is likely not truthful.
Step 3: Apply the pattern to the new hotel review and predict whether the new hotel review is truthful.
Step 4: Give your final answer: yes or no. If you are unsure, respond with no.
Please give your answer strictly in the following format:
"""

```
Analysis: [your step-by-step analysis]
Answer: [your answer]
```

"""

Evaluation prompt for pneumoniaMNIST

You are an expert in pneumonia detection, and your job is to apply learned patterns to predict if a person has pneumonia.
Here are some previously generated pneumonia patterns: {{patterns}}
A chest X-ray image is shown.
Based on the learned patterns and given image, is this person likely to have pneumonia based on the learned patterns?
Think step-by-step.
Step 1: Look at the given chest X-ray image and compare it with the provided pneumonia patterns.
Step 2: If the image features match the pneumonia patterns, then the person is likely to have pneumonia. If the features do not match the patterns, then the person is likely not to have pneumonia.
Step 3: Apply the pattern to the new chest X-ray image and predict whether the person has pneumonia.
Step 4: Give your final answer: yes or no. If you are unsure, respond with no.
Please give your answer strictly in the following format:
"""

```
Analysis: [your step-by-step analysis]
Answer: [your answer]
```

"""

Give your answer in the above format. Don't talk about any other words.

Evaluation prompt for funny reddit

You will predict whether a Reddit post is funny based on the given funny Reddit post patterns.
Here are some previously generated funny Reddit post patterns:
{{patterns}}
A Reddit post is the following: {{text}}
Is this Reddit post funny?
Think step-by-step:
Step 1: Look at the new Reddit post and compare it with the provided funny post patterns.
Step 2: If the post matches the pattern, then it is likely funny. If it does not match the pattern, then it is likely not funny.
Step 3: Apply the pattern to the new Reddit post and predict whether the new post is funny.
Step 4: Give your final answer: yes or no. If you are unsure, respond with no.
Please give your answer strictly in the following format:
"""

```
Analysis: [your step-by-step analysis]
Answer: [your answer]
"""
```