Literature Meets Data: A Synergistic Approach to Hypothesis Generation

Haokun Liu♣*, **Yangqiaoyu Zhou**♣*, **Mingxuan Li**♣*, **Chenfei Yuan**†
& **Chenhao Tan**♣
Department of Computer Science
University of Chicago♣, Tsinghua University†
Chicago, IL 60637, USA
{haokunliu, zhouy1, mingxuanl, chenhao}@uchicago.edu,
yuancf21@mails.tsinghua.edu.cn

Abstract

AI holds promise for transforming scientific processes, including hypothesis generation. Prior work on hypothesis generation can be broadly categorized into theory-driven and data-driven approaches. While both have proven effective in generating novel and plausible hypotheses, it remains an open question whether they can complement each other. To address this, we develop the first method that combines literature-based insights with data to perform LLM-powered hypothesis generation. We apply our method on five different datasets and demonstrate that integrating literature and data outperforms other baselines (8.97% over few-shot, 15.75% over literature-based alone, and 3.37% over data-driven alone). Additionally, we conduct the first human evaluation to assess the utility of LLM-generated hypotheses in assisting human decision-making on two challenging tasks: deception detection and AI generated content detection. Our results show that human accuracy improves significantly by 7.44% and 14.19% on these tasks, respectively. These findings suggest that integrating literature-based and data-driven approaches provides a comprehensive and nuanced framework for hypothesis generation and could open new avenues for scientific inquiry.

1 Introduction

> "It is the theory that decides what can be observed." —Albert Einstein

Hypothesis generation drives the design of experiments and determines the set of possible scientific discoveries. However, it remains largely informal and relies on human intuitions (Ludwig and Mullainathan, 2024). Large language models (LLMs) excel at synthesizing information and identifying patterns, and thus hold promise for transforming hypothesis generation. Many recent studies recognize this potential and use LLMs to generate hypotheses (e.g., Yang et al., 2024b; Batista and Ross, 2024). We broadly categorize them into *theory-driven* and *data-driven* methods.

On one hand, theory-driven approaches leverage LLMs to review existing literature and generate novel hypotheses/ideas (Yang et al., 2024b; Baek et al., 2024). These methods have shown promising results in terms of the hypotheses' novelty, validity, and usefulness to researchers, while remaining grounded in established human knowledge (Si et al., 2024). However, they come with notable limitations: they require high-quality literature, struggle to adapt to new data, and lack empirical support. Data-driven approaches, on the other hand, propose hypotheses by discovering patterns in data (Zhou et al., 2024; Qiu et al., 2024). These hypotheses are data-adaptive and can exhibit strong performance in explaining the data. However, they could be too overly tailored to the specific datasets used, which can hinder their generalizability.

We hypothesize that theory can guide the discovery from data and propose to integrate literature-based and data-driven hypothesis generation (see Figure 1). For the data-driven component, we use HYPOGENIC as the backbone (Zhou et al., 2024). HYPOGENIC leverages an LLM to initialize hypotheses from a small number of examples and then updates them iteratively to improve the quality of hypotheses. To enhance this process with literature insights, we introduce a literature-based hypothesis agent. This agent interacts with the data-driven hypothesis agent (HYPOGENIC), refining and maintaining a shared pool of hypotheses through continuous collaboration, ensuring that the hypotheses benefit from both data-driven adaptability and the grounding of existing scientific knowledge. In addition to the refinement approach, we also propose to directly unionize literature-based and data-driven hypotheses.

[*] Equal contributions.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 245–281
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

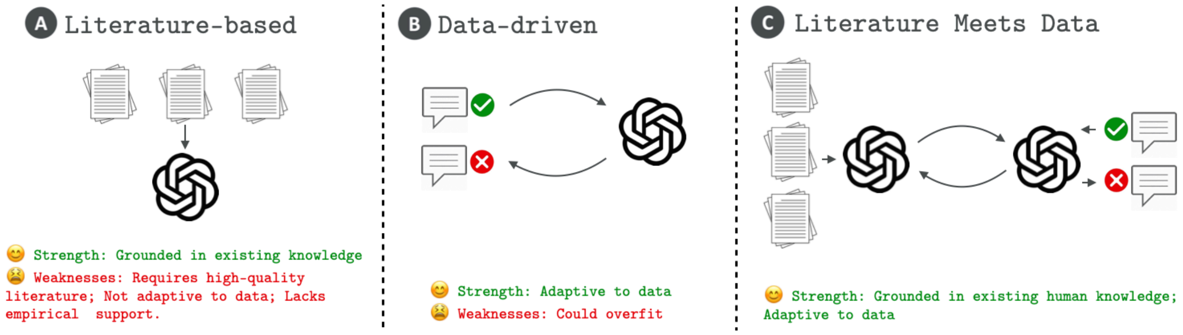

A Literature-based
Strength: Grounded in existing knowledge
Weaknesses: Requires high-quality literature; Not adaptive to data; Lacks empirical support.
B Data-driven
Strength: Adaptive to data
Weaknesses: Could overfit
C Literature Meets Data
Strength: Grounded in existing human knowledge; Adaptive to data

Figure 1: The literature-based approach (A) leverages existing human knowledge to generate hypotheses but struggles to adapt to new data and lacks empirical grounding. The data-driven approach (B) relies on large datasets to generate hypotheses, enabling adaptation to diverse scenarios but risking overfitting. The literature + data approach (C) combines the strengths of both, grounding hypotheses in human knowledge while incorporating empirical data to enhance adaptiveness. See algorithmic details in § 2.

To comprehensively evaluate these hypotheses, we conduct automatic and human evaluation to assess their generalizability, utility, and novelty. We apply our method to address research questions in social sciences: deception detection, AI generated content (AIGC) detection, mental stress detection, and persuasive argument prediction. Automatic evaluation results show that integrating literature and data outperforms other baselines: 8.97% over few-shot, 15.75% over literature-based alone, and 3.37% over data-driven alone in accuracy on out-of-distribution datasets, a measure of generalizability.

Moreover, we conduct the first study to assess the utility of AI-generated hypotheses in improving human decision-making and show that our generated hypotheses improve human accuracy by 7.44% and 14.19% on deception detection and AIGC detection. Additionally, we find that literature-based and data-driven hypotheses complement each other, as one set often contains novel information not found in the other set.

In sum, we make the following contributions:

- We propose the first approach to using both literature information and data for LLM-powered hypothesis generation.
- We conduct automatic evaluation to assess the utility of the generated hypotheses in improving LLM predictions. Experiments on five datasets demonstrate the effectiveness of our approach.
- We conduct the first human evaluation to test the utility of LLM-generated hypotheses and demonstrate consistent improvements on two challenging tasks.

The code and data are available at https://github.com/ChicagoHAI/hypothesis-generation.

2 Methods

We formulate the problem of hypothesis generation as follows. Assuming that we have access to literature $\mathcal{L}$ and observational data $\mathcal{D}$ that are relevant to a research question q. Then, we want to develop an AI-powered algorithm f with model $\mathcal{M}$ such that we can generate high-quality hypotheses for the research question q, i.e., $\mathcal{H} = f_{\mathcal{M}}(q, \mathcal{L}, \mathcal{D})$. Example research questions include what makes an argument persuasive and what signs are indicative of AI-generated texts. In this work, we consider research questions that can be formulated as classification tasks, so we use q and *task* interchangeably.

2.1 Literature-Based Hypothesis Generation

For the LITERATURE-ONLY method, we start by picking a set of papers $\mathcal{P} = \{p_1, p_2, ..., p_m\}$ for q from related papers on Semantic Scholar or Google Scholar. We also choose from papers that cited the original datasets for each task. Subsequently, we use S2ORC-doc2json to convert the raw PDF versions of the papers to a corpus of JSON files (Lo et al., 2020). We denote the converted papers as $\mathcal{C} = \{\text{doc2json}(p) : p \in \mathcal{P}\}$. Passing the full texts of all these papers to a language model would likely exceed its maximum context length. Moreover, we want to generate hypotheses from the key findings of the relevant papers because some contents, such as technical details, may not help significantly but distract the LLM. Therefore, we develop a paper summarizer $\mathcal{M}_S$ to generate paper summaries $\mathcal{S} = \{\mathcal{M}_S(p_c) : p_c \in \mathcal{C}\}$ (throughout the paper, we use subscripts to indicate $\mathcal{M}$ with different prompts). Lastly, we instruct language models to generate hypotheses $\mathcal{H}_{\mathcal{L}} = \mathcal{M}_G(\mathcal{S})$ based on the generated paper summaries, with an emphasis on usefulness for carrying out the specific tasks that our literature corpus focuses on.

In addition to our own implementation, we use commercial ones such as NOTEBOOKLM (Google, 2024) and HYPERWRITE (OthersideAI, 2024) as strong baselines.

2.2 Data-Driven Hypothesis Generation

Our data-driven hypothesis generation adopts HYPOGENIC in Zhou et al. (2024). Here we give a brief overview. Suppose we have a set of observational data in the form of input-label pairs, i.e., $\mathcal{D} = \{(x_1, y_1), ..., (x_n, y_n)\}$. During the initialization stage of HYPOGENIC, a generation agent $\mathcal{M}_G$ is prompted with a set of initial data instances $\mathcal{D}_{\text{init}} \subset \mathcal{D}$ and asked to generate initial hypotheses $\mathcal{H}^0_{\mathcal{D}} = \mathcal{M}_G(\mathcal{D}_{\text{init}})$. Then, for each of the initial hypothesis h and for every example $(x_i, y_i) \in \mathcal{D}_{\text{init}}$, h is used to prompt an inference agent $\mathcal{M}_I$ to make a prediction $\hat{y}_i = \mathcal{M}_I(x_i, h)$. The initial reward of each hypothesis is computed using:

$$r_t(h) := \text{Acc}(h, \mathbf{X}^t_h) + \alpha\sqrt{\frac{\log t}{|\mathbf{X}^t_h|}},$$
$$\text{Acc}(h, \mathbf{X}) := \frac{\sum_{(x_i,y_i)\in\mathbf{X}} \mathbb{1}(y_i = \mathcal{M}_I(x_i, h))}{|\mathbf{X}|}, \tag{1}$$

where r_t is the reward function inspired by the upper confidence bound (UCB) algorithm (Auer, 2003), $\mathbf{X}^t_h$ is the set of examples being used to evaluate hypothesis h up to time t, and α is the reward coefficient that controls the exploration term of the reward function. We also initialize $\mathcal{W} = \emptyset$, where $\mathcal{W}$ keeps track of the wrongly predicted examples.

In the update stage at time step $t + 1$, we take $(x_{t+1}, y_{t+1}) \in \mathcal{D}$, and the top k high-reward hypotheses. For each h of the selected hypotheses, we prompt $\mathcal{M}_I$ to make a prediction $\mathcal{M}_I(x_{t+1}, h)$. The accuracy and reward of the k hypotheses are updated using Eq. 1. Among the k hypotheses, if at least w_{hyp} predicted wrong, We add the example to $\mathcal{W}$. If $|\mathcal{W}| \geq w_{\max}$, a set of new hypotheses $\mathcal{H}_{\mathcal{W}} = \mathcal{M}_G(\mathcal{W})$ is generated and added to $\mathcal{H}^{t+1}_{\mathcal{D}}$. Then the top $H_{\max}$ hypotheses are kept, and $\mathcal{W}$ is reset to $\emptyset$. After all n examples are visited, we denote the final hypothesis bank as $\mathcal{H}_{\mathcal{D}}$.

2.3 Integration of Literature-based and Data-driven Hypotheses

One main contribution of our work is proposing the first approach to integrate literature-based and data-driven hypothesis generation so that we can effectively leverage the strengths of each approach, increasing the generalizability and utility of generated hypotheses. We consider two strategies.

Refinement of literature-based hypotheses. HYPOREFINE integrates paper summaries $\mathcal{S}$ from § 2.1 with HYPOGENIC. In the initialization stage, a generation agent $\mathcal{M}_G$ is asked to generate initial hypotheses based on both a set of initial examples and paper summaries, given by $\mathcal{H}^0_{\mathcal{L}+\mathcal{D}} = \mathcal{M}_G(\mathcal{S}, \mathcal{D}_{\text{init}})$.

In the update stage, we propose an iterative refinement approach to integrate patterns from data and key findings from literature into new hypotheses. Specifically, each time HYPOGENIC generates a set of new hypotheses $\mathcal{H}_{\mathcal{W}}$, these hypotheses are refined multiple rounds by a data-driven refinement agent and a literature-based refinement agent. Take $\mathcal{M}_R$ as the refinement agent, each time $\mathcal{H}_{\mathcal{W}}$ is generated from the wrong examples pool $\mathcal{W}$, it is iteratively refined as follows:

$$\mathcal{H}^i_{\mathcal{W}} = \begin{cases} \mathcal{M}_R(\mathcal{H}^{i-1}_{\mathcal{W}}, \mathcal{S}) & \text{if } i \bmod 2 = 0 \\ \mathcal{M}_R(\mathcal{H}^{i-1}_{\mathcal{W}}, \mathcal{W}) & \text{if } i \bmod 2 = 1. \end{cases}$$

After N_{refine} rounds of refinement, the final hypothesis bank $\mathcal{H}^{N_{\text{refine}}}_{\mathcal{W}}$ is fed back to the HYPOGENIC pipeline as $\mathcal{H}_{\mathcal{W}}$.

The reward function and update process for the hypothesis bank $\mathcal{H}^t_{\mathcal{L}+\mathcal{D}}$ remain consistent with those of the original HYPOGENIC.

Union and redundancy elimination. As the reward function of HYPOGENIC focuses only on the hypotheses' performance on the datasets at hand, literature-based hypotheses are sometimes undervalued during the update stage. On occasions they can even be replaced by hypotheses that have especially good performances on data but are not necessarily generalizable on real-world tasks. To counter this issue, we use a union approach to combine literature-based and data-based hypotheses. We first generate two hypothesis banks: one literature-based hypothesis bank $\mathcal{H}_{\mathcal{L}}$ and the other bank from $\mathcal{H}_{\mathcal{D}}$ or $\mathcal{H}_{\mathcal{L}+\mathcal{D}}$, using HYPOGENIC or HYPOREFINE, respectively. Then we build a redundancy checker to remove hypotheses that express overly similar or repeating information in each bank. Lastly, we construct the final hypothesis bank of size $H_{\max}$ by randomly choosing $\frac{H_{\max}}{2}$ hypotheses from the literature-based hypothesis bank and adding the top $\frac{H_{\max}}{2}$ hypotheses from the other hypothesis bank based on training accuracies. For detailed information of the implementation, please refer to Appendix B.3.

3 Experiments

In this section, we introduce our evaluation framework and the tasks to operationalize it.

3.1 Evaluation Framework

Formally evaluating hypotheses requires rigorous protocols and vast amounts of resources. In this work, we mainly evaluate our generated hypotheses along two dimensions: utility and novelty. We perform both automatic and human evaluations to show that our generated hypotheses can help models and humans in challenging real-world classification tasks and bring novel information.

Automatic evaluation on out-of-distribution (OOD) and in-distribution (IND) datasets and cross-model inference. Since we work with classification tasks, a natural way of evaluating the hypotheses is prompting the LLMs to do inference with the hypotheses. For all methods that generate hypotheses $\mathcal{H}$, with every test example (x, y), we prompt $\mathcal{M}_I$ to first extract the most relevant hypotheses to the example and make inference using the hypotheses, denoted as $\mathcal{M}_I(\mathcal{H}, x)$. For detailed information about the prompts, see Appendix A. Then we compute $\text{Acc}(\mathcal{H}, \mathcal{D}_{\text{test}})$, defined in eq. 1, for a held-out set $\mathcal{D}_{\text{test}}$. For each task, we report average accuracy and F1 scores on held-out OOD and IND sets for 5 different random seeds. Since we are most interested in the generalizability of the generated hypotheses, we focus on performance on the OOD set in the main paper.

In addition to predicting on out-of-distribution datasets, we test our hypotheses' generalizability by taking the hypotheses generated by one model and performing inference with another model.

Human evaluation on utility and novelty. We design human studies to assess the practical utility of the generated hypotheses on Deception Detection and AIGC Detection. In addition, we evaluate the perceived clarity, novelty, and plausibility through surveys. Screenshots and details of the studies are in Appendix C. We pay participants at an average hourly rate of $12.

Human Study I: Utility in human decision-making. We recruit 60 Prolific participants and randomly assign them into experimental and control groups. The control group performs the task without hypotheses, while the experiment group is given a set of three generated hypotheses to complete the same task. Specifically, each participant is randomly assigned 14 instances, and we include attention check questions to ensure the quality of the collected responses. We evaluate the practical utility of our generated hypotheses by comparing the performance of the two groups.

We pick the hypotheses based on their impact on performance in an ablation setting. Specifically, we choose the top three hypotheses that cause the greatest drop in performance when removed from the hypotheses pool during multi-hypothesis inference. In addition, to motivate participants to perform at their best, we offer a bonus of $0.1 for each correctly predicted instance.

At the end of the study, participants in the experiment group are also asked to give overall ratings and an assessment of the given hypotheses. There are five scales: "Not at all helpful", "Slightly helpful", "Moderately helpful", "Very helpful", and "Extremely helpful".

Human study II: Clarity, novelty, and plausibility. We recruit 30 participants with graduate-level degrees in social sciences from *prolific.com* to evaluate hypotheses generated by HYPOREFINE, NOTEBOOKLM, and HYPERWRITE for Deception Detection. Using a 5-point Likert scale, participants assess each hypothesis along three dimensions: clarity, novelty, and plausibility. See details in Appendix C.2.

Human study III: Novelty and nuance. To compare data-driven hypotheses and literature-driven hypotheses, we present one hypothesis of each type to participants and ask them to judge whether the second hypothesis provides meaningfully novel information that is not covered in the first hypothesis.

We sample 50 pairs of hypotheses (h_1, h_2), one from literature-based and one from data-driven, with duplications removed within each group. We recruit 10 Prolific participants to annotate whether h_2 provides new information to h_1 for each pair. Each participant is randomly assigned to annotate 15 pairs. For each pair, we take the majority vote to determine the final novelty label.

3.2 Tasks

We consider four tasks in social sciences.

Deception Detection is a widely studied problem in psychology and other social sciences (Granhag and Vrij, 2005). We use the dataset introduced by

Ott et al. (2013) (DECEPTIVE REVIEWS), which consists of 800 genuine hotel reviews and 800 fake hotel reviews, as our IND dataset. For the OOD dataset, we use hotel reviews from different source websites and different cities (Li et al., 2013).

AI-Generated Content (AIGC) Detection has attracted significant attention in recent years (Tang et al., 2023). Most existing works focus on developing black-box detection methods and rarely take interpretability into account (Wu et al., 2024). We thus build our own dataset for this task. We take 800 distinct prompts and human-written stories in the WRITINGPROMPTS dataset (Fan et al., 2018). Then we use the same prompts to generate AI-written stories with LLAMA-3.1-70B-INSTRUCT (Dubey et al., 2024) and GPT-4O-MINI (OpenAI, 2023), constituting our LLAMAGC and GPTGC datasets. The IND data contains stories generated by the corresponding model. The stories generated by the other model are treated as OOD data.

Persuasive Argument Prediction examines persuasion and social interactions to reveal predictive cues of persuasiveness (Tan et al., 2016). We use PERSUASIVE PAIRS, a dataset with pairs of short texts constructed by Pauli et al. (2024). Within each pair of texts, one is from existing corpora with signals of persuasiveness, while the other one is generated by an LLM with instructions for it to be more/less persuasive than the one from existing corpora. We formulate this task as predicting the more persuasive one of each pair of texts. The dataset contains human-annotated ground-truth labels and is pre-processed by removing examples where there exists disagreement among annotators. The IND and OOD datasets are then created based on different original sources of texts.

Mental Stress Detection from social media content is an important task in mental health (Lupien et al., 2009). We use DREADDIT, a corpus of lengthy Reddit posts with stress status labels developed by Turcan and McKeown (2019). The dataset contains 3.5k post segments annotated using Amazon Mechanical Turk, with labels indicating the presence or absence of stress in posts. Our IND and OOD sets are separated based on subreddits that the posts come from.

For each task, we split the IND dataset with at least 200 examples in train set, 300 in test set (on which we perform inference), 300 in validation set, and sample at least 300 instances from OOD (see Appendix B.1 for more details).

3.3 Implementation and Baselines

Our method works with any LLM ($\mathcal{M}$). We use GPT-4O-MINI and LLAMA-3.1-70B-INSTRUCT in the main paper. We refer to GPT-4O-MINI as "GPT-4-MINI" and LLAMA-3.1-70B-INSTRUCT as "LLAMA-70B-I". We compare our method with the following baselines.

1. **Zero-shot and few-shot prompting.** We give the LLMs detailed task instructions (zero-shot) and optionally provide three demonstrating examples (few-shot). This approach does not involve any hypothesis.

2. **Zero-shot hypothesis generation.** Inspired by Qi et al. (2023), we provide specific task descriptions and instructions, and then we prompt the LLMs to generate hypotheses directly without incorporating literature or data.

3. **Literature-driven hypothesis generation.** We use the implementation in §2.1. In addition to our own implementation, we compare two of the recently released agent frameworks for scientific writing, NOTEBOOKLM (Google, 2024) and HYPERWRITE (OthersideAI, 2024). We use the same prompt for NOTEBOOKLM and HYPERWRITE as what we apply in our methods. See details in Appendix B.5. These methods only use literature in hypothesis generation.

4. **Data-driven hypothesis generation**. We use HYPOGENIC. See details in § 2.2.

For all the hypothesis generation methods we use, we keep the size of the hypothesis bank $\mathcal{H}$ to be 20 (i.e., $H_{\max} = 20$.)

4 Results

We first present automatic evaluation results to demonstrate the utility of generated hypotheses for model inference. We then show that the generated hypotheses are novel and useful, and can improve human decision-making in challenging tasks.

4.1 Automatic Evaluation

Hypotheses generated by combining information from literature and data achieves the best performance across all task and model configurations (Table 1). First, few-shot inference outperforms zero-shot inference for all task and model configurations, with an average improvement of 6.84% in accuracy. In addition, few-shot inference surpasses zero-shot generation and the best of literature-based methods on average accuracy by

Model	Methods	DECEPTIVE REVIEWS	LLAMAGC	GPTGC	PERSUASIVE PAIRS	DREADDIT
GPT-4 MINI	**No hypothesis**					
	Zero-shot	55.47	50.00	56.33	81.24	64.60
	Few-shot k=3	65.56	51.11	64.22	83.64	75.00
	Zero-shot generation	68.69	49.00	53.00	86.08	65.00
	Literature-based					
	LITERATURE-ONLY	59.22	49.00	54.00	78.80	67.68
	HYPERWRITE	61.63	49.67	52.67	82.36	68.76
	NOTEBOOKLM	53.03	49.33	51.67	68.96	62.28
	Data-driven					
	HYPOGENIC	75.22	81.67	68.56	82.20	76.56
	Literature + Data (This work)					
	HYPOREFINE	**77.78**	55.33	63.33	89.04	78.04
	Literature ∪ HYPOGENIC	72.41	**83.00**	**69.22**	**89.88**	78.20
	Literature ∪ HYPOREFINE	77.19	55.33	63.00	89.52	**79.24**
LLAMA 70B-I	**No hypothesis**					
	Zero-shot	62.87	58.67	63.00	85.60	64.56
	Few-shot k=3	68.56	70.45	76.00	86.80	69.44
	Zero-shot generation	56.28	50.67	55.67	88.16	66.16
	Literature-based					
	LITERATURE-ONLY	64.25	50.00	49.67	80.56	66.04
	HYPERWRITE	58.62	50.67	54.00	83.24	74.40
	NOTEBOOKLM	57.81	49.33	50.67	67.64	66.56
	Data-driven					
	HYPOGENIC	62.06	78.67	78.00	88.44	75.48
	Literature + Data (This work)					
	HYPOREFINE	72.16	67.00	66.67	87.52	**78.92**
	Literature ∪ HYPOGENIC	**73.72**	**81.33**	**78.67**	86.72	72.56
	Literature ∪ HYPOREFINE	71.75	66.67	65.67	**88.76**	74.80

Table 1: Accuracy scores on the held-out OOD datasets. Literature + Data outperforms all other methods in every model and task configurations. The bolded numbers outperform the few-shot method ($p < 0.05$), as determined by a paired t-test using five random seeds. We show the full results with F1 scores in Table 8.

7.21% and 6.78%, respectively, suggesting off-the-shelve LLMs or literature alone does not generate effective hypotheses for predictive purposes. In fact, NOTEBOOKLM and HYPERWRITE can generate some invalid or irrelevant hypotheses, which degrades their inference performance (see Table 17 in Appendix E.2).

In contrast, HYPOGENIC consistently outperforms few-shot inference, improving average accuracy by 5.61%, highlighting the advantage of data-driven hypotheses. Compared to few-shot inference, the hypotheses also offer more interpretable insights. Furthermore, our best hypothesis generation method combining literature and data outperforms HYPOGENIC by 3.37% on average (i.e., an improvement of 11.92% over few-shot methods and 16.54% over literature-based methods for GPT-4-MINI, and 6.03% over few-shot methods and 14.97% over literature-based methods for LLAMA-70B-I), demonstrating the benefit of incorporating literature with data.

For DECEPTIVE REVIEWS, PERSUASIVE PAIRS, and DREADDIT, refining the hypotheses with literature consistently improves inference accuracy compared to HYPOGENIC, with a 3.92% improvement on average. On the other hand, refining the hypotheses with literature does not help with GPTGC and LLAMAGC, but the union of HYPOGENIC and hypotheses generated from literature consistently performs the best. Comparing with HYPOGENIC for these two tasks, refining the hypotheses with literature actually results in an accuracy drop by 13.64%. This is likely due to that the literature for AIGC detection has relatively few insights on interpretable features to detect AI generated contents, and refining the data-driven hypotheses with that information degrades performance.

To further illustrate our approach, we present

Case I: LITERATURE-ONLY and HypoGeniC generate **different** hypotheses

LITERATURE-ONLY: Deceptive reviews often contain a higher frequency of first-person singular pronouns, while truthful reviews may use these pronouns less frequently.
HypoGeniC: Reviews that reference the reviewer's previous experiences with the hotel brand or similar hotels are more likely to be truthful, while reviews that do not provide any context or comparison to past experiences are more likely to be deceptive.

Case II: LITERATURE-ONLY and HypoGeniC generate **similar** hypotheses

LITERATURE-ONLY: Truthful reviews often provide a balanced perspective, while deceptive reviews may seem overly promotional or biased towards a competitor.
HypoGeniC: Reviews that express a balanced perspective, mentioning both positive and negative aspects of the stay, are more likely to be truthful, whereas reviews that are overly positive or negative without nuance tend to be deceptive.
HypoRefine: Reviews that present a balanced perspective by discussing both positive and negative aspects of the stay, particularly with specific examples (e.g., "The location was fantastic, but the air conditioning was broken"), are more likely to be truthful, while reviews that are excessively positive or negative without acknowledging any redeeming qualities (e.g., "This is the best hotel ever!" or "I will never stay here again!") tend to be more deceptive, as they may reflect an attempt to manipulate reader emotions rather than provide an honest assessment.

Table 2: Examples of generated hypotheses from different methods. We show cases where LITERATURE-ONLY and HypoGeniC generate different hypotheses or similar hypotheses, and how HypoRefine combines them in the case if they express unifiable ideas.

a case study of our generated hypotheses in Table 2. For most cases, LITERATURE-ONLY and HypoGeniC generate different hypotheses as in Case I: one is about first-person singular pronouns, while the other one is about past experiences. We include more details on the differences between hypotheses generated by different methods in § 4.2. More examples of hypotheses generated using Literature∪HypoRefine are in Table 16. Under some cases, the methods can generate similar hypotheses, and HypoRefine improves the quality of the hypothesis. In Case II, all three hypotheses focus on balanced perspectives being indicative of truthful reviews. HypoRefine incorporates the "reviews that seem to be promoting a competitor" insight from LITERATURE-ONLY, while also capturing the emphasis on "lack of nuance" from HypoGeniC. By doing so, HypoRefine offers a more nuanced hypothesis that not only explains how deceptive reviews may manipulate reader emotions, but also provides specific examples to illustrate how balanced perspectives can contribute to truthful assessments. This combination of insights from literature and data allows HypoRefine to offer a more comprehensive and explanatory hypothesis. We include another case study to compare the generated hypotheses from SciMON (Wang et al., 2024) and HypoRefine in Appendix E.1, demonstrating the difference between research idea generation and the hypothesis generation task we focus on.

Performance on IND held-out datasets. Similar to Table 1, combining literature and data achieves the best accuracy and F1 scores in most cases on the held-out IND datasets (see Table 9 in the Appendix). For some cases, such as using Llama on the IND datasets for GPTGC, LlamaGC, and Persuasive Pairs, HypoGeniC gets the top performance compared to other methods. This is not surprising, since HypoGeniC generates hypotheses by looking at IND data only. In contrast, our methods that take information from both literature and data may generate hypotheses that are more generally applicable but with slightly worse performance on the IND data. Thus in Table 1, the hypotheses generated from both literature and data performs the best on all methods for OOD datasets.

Generated hypotheses can be effectively transferred to a different model. To further check the generalization behavior of our generated hypotheses, we take the hypotheses from the best-performing method with our literature+data approach and then use the other model to perform inference. Table 3 shows that the generated hypotheses from one model remain effective for the other model, the performance exhibits no significant change in most cases (drop <5% in 11 out of 20 cases). Even with this performance drop, our methods still outperform the few-shot baseline by 3.76% and 3.66% in OOD and IND settings. This finding further demonstrates the robustness of our approach to hypothesis generation and hypothesis-based inference.

A significant outlier case is for LlamaGC OOD: when using Llama-70B-I-generated hypotheses for GPTGC OOD and ask GPT-4-mini to perform hypothesis-based inference, the inference performance can degrade significantly. This can be due to innate deficits in the task setting, as LLMs tend

Generation Model	Inference Model	DECEPTIVE REVIEWS	LLAMAGC	GPTGC	PERSUASIVE PAIRS	DREADDIT
		OOD Accuracy	OOD Accuracy	OOD Accuracy	OOD Accuracy	OOD Accuracy
GPT-4-MINI	GPT-4-MINI	77.78	83.00	69.22	89.88	79.24
	LLAMA-70B-I	72.53 (↓5.25)	71.67 (↓11.33)	76.33 (↑7.11)	86.88 (↓3.00)	72.36 (↓6.88)
LLAMA-70B-I	LLAMA-70B-I	73.72	81.33	78.67	88.76	78.92
	GPT-4-MINI	70.31 (↓3.41)	57.00 (↓24.33)	74.67 (↓4.00)	89.36 (↑0.60)	77.28 (↓1.64)
Generation Model	Inference Model	IND Accuracy	IND Accuracy	IND Accuracy	IND Accuracy	IND Accuracy
GPT-4-MINI	GPT-4-MINI	70.76	73.00	68.33	90.52	70.88
	LLAMA-70B-I	62.20 (↓8.56)	69.00 (↓4.00)	81.67 (↑13.34)	90.40 (↓0.12)	74.88 (↑4.00)
LLAMA-70B-I	LLAMA-70B-I	72.60	78.33	80.67	91.24	78.68
	GPT-4-MINI	66.28 (↓6.32)	68.00 (↓10.33)	68.33 (↓12.34)	89.88 (↓1.36)	68.60 (↓10.08)

Table 3: Cross-model inference performance.

to favor and better detect their own writing (Panickssery et al., 2024).

Our hypothesis generation method is robust to different prompts and hyperparameters, and it is effective for smaller models. We conduct additional experiments to test the robustness of our method with different prompts, hyperparameters, and with LLAMA-3.1-8B-INSTRUCT. For the OOD setting, in Table 11 and Table 13, we show that the average performance of our method only drop by 0.20% and 0.07% with different prompts and hyperparameters, respectively. In Table 10, we further show that with LLAMA-3.1-8B-INSTRUCT, our literature + data approach outperforms all baselines, with an average accuracy improvement of 15.27%, 13.04%, and 4.88% over few-shot, literature-based methods, and HYPOGENIC, respectively. These results highlight the robustness of our method and its scalability and reproducibility with smaller models. Full details are in Appendix D.

4.2 Human Evaluation

Generated hypotheses improve human decision-making in both AIGC Detection and Deception Detection. We experiment with the GPTGC task for AIGC Detection, and the average human accuracy improves by 14.19% (58.86% → 73.05%) when we provide hypotheses as assistance. We perform a statistical t-test and obtain a p-value of 0.01, indicating that the improvement is significant. In Deception Detection, the introduction of hypotheses boosts human accuracy by 7.44% (57.14% → 64.58%), with a p-value of 0.04.

When hypotheses are present, participants would use them to assist decision-making for over 90% of the time. All three presented hypotheses are selected to be used with frequency greater than 30% (Table 5, Table 6 in the Appendix). For example, the most used hypothesis, with frequency of 44.55%, in AIGC detection is "Human-written texts tend to have a more conversational tone and colloquial language, while AI-generated texts tend to be more formal and lack idiomatic expressions." For both tasks, 100% of the participants find the hypotheses to be helpful, and over 40% find them to be "Very helpful" or "Extremely helpful".

	Clarity	Novelty	Plausibility
HYPOREFINE	4.22 ± 0.07	2.85 ± 0.10	4.21 ± 0.08
NOTEBOOKLM	3.71 ± 0.09*	3.01 ± 0.10	3.65 ± 0.10*
HYPERWRITE	3.08 ± 0.12*	2.15 ± 0.11*	3.32 ± 0.13*

Table 4: Human ratings on hypotheses generated by HYPOREFINE, NOTEBOOKLM, and HYPERWRITE. * indicates that the difference between the rating with HYPOREFINE is statistically significant (p <0.001).

HYPOREFINE hypotheses are rated higher in clarity and plausibility compared to those generated by existing methods (Table 4). Hypotheses generated by HYPOREFINE achieve statistically significantly higher clarity and plausibility scores than those generated by NOTEBOOKLM and HYPERWRITE. In terms of novelty, NOTEBOOKLM receives slightly higher ratings; however, the difference between NOTEBOOKLM and HYPOREFINE is not statistically significant. Recall that hypotheses generated by NOTEBOOKLM do not have strong predictive power. In other words, generating "novel" hypotheses is easier if they are not constrained by plausibility.

Humans rate literature-based and data-driven hypotheses as distinct. We assign novelty labels to hypothesis pairs based on a majority vote from three human annotators, who evaluate whether a hypothesis give meaningfully different information ("novel") from another. 84% and 80% of the hypotheses are rated novel for Deception Detection and AIGC Detection, respectively, demonstrating

Hypotheses	Frequency of Selection
Hypothesis 1: AI-generated texts tend to use more elaborate and descriptive language, including adjectives and adverbs, to create a sense of atmosphere and immersion. Human-written texts, on the other hand, tend to be more concise and straightforward in their language use.	38.79%
Hypothesis 2: Human-written texts are more likely to contain errors or idiosyncrasies in grammar and punctuation, reflecting the natural imperfections of human writing, while AI-generated texts typically maintain a higher level of grammatical accuracy.	34.55%
Hypothesis 3: Human-written texts tend to have a more conversational tone and colloquial language, while AI-generated texts tend to be more formal and lack idiomatic expressions.	44.55%
No hypothesis selected	3.94%

Table 5: How often participants use hypotheses in AIGC Detection. We allow users to select multiple hypotheses for each instance they make prediction on, so the total frequency can exceed 100%.

the complementarity between literature-based and data-driven approaches.

5 Related Work

Literature-based research idea generation. Baek et al. (2024), Wang et al. (2024), and Ghafarollahi and Buehler (2024) use LLMs to build knowledge graphs from literature and generate research ideas, such as proposing new problem setups, methodologies, or evaluation frameworks. Unlike their focus on ideation, our work generates hypotheses to explain a phenomenon with real observations (see Appendix E.1 for detailed comparison). Yang et al. (2024b) generates hypotheses from raw web data but relies on annotated hypotheses from literature. These methods require extensive adaptation, so we developed our own literature-based approach.

Data-driven hypothesis generation. Besides HYPOGENIC, we review additional works on discovering unseen patterns from data. Zhong et al. (2023) discovers patterns by analyzing difference between large corpora. Pham et al. (2024) makes discovery by generating and refining interpretable topics. Romera-Paredes et al. (2024) uncovers new solutions in open math problems by iteratively updating programs. Qiu et al. (2024) and Yang et al. (2024a) evaluate LLMs' ability in performing inductive reasoning in synthetic settings. Batista and Ross (2024) uses LLMs to generate hypotheses and conducts comprehensive experiments to study human engagements with headlines. Recent work has also shown that LLM-generated hypotheses can meaningfully impact real-world applications beyond scientific discoveries (Li et al., 2025; Garbacea and Tan, 2025). We choose HYPOGENIC as the backbone for data-driven hypothesis generation as their tasks are most similar to ours, and their approach to hypothesis updates integrates naturally into our refinement process.

Automated scientific research with LLMs. There is growing interest in developing LLM-powered methods and multi-agent frameworks to assist scientific research. Lu et al. (2024) designs an LLM agent to generate full research papers. Li et al. (2024) proposes a method to generate research ideas from existing literature and automatically implement and execute experiments. In contrast, our work focuses primarily on hypothesis generation, as we believe it is crucial to preserve human agency and oversight in the scientific research process.

To evaluate LLM generated hypotheses, Qi et al. (2023) examines whether they contain novel information not found in existing literature. Si et al. (2024) asks experts to rate the novelty of LLM-proposed research ideas in the NLP domain. While these studies highlight LLMs' ability to generate novel hypotheses, they do not conduct human subject experiments to validate the effectiveness of hypotheses. To this end, we conduct the first human study to test the utility of LLM-generated hypotheses in supporting human decision-making.

Significant efforts have also been made to evaluate and benchmark multi-agent frameworks on data analysis tasks (Majumder et al., 2024; Gu et al., 2024; Hu et al., 2024; Chen et al., 2024; Huang et al., 2024; Guo et al., 2024), literature processing and information retrieval tasks (Press et al., 2024; Ajith et al., 2024; Kang and Xiong, 2024; Zhang et al., 2024), and more general research tasks (Tian et al., 2024; Jansen et al., 2024).

6 Conclusion

We propose a novel approach that integrates literature and data to generate hypotheses, with extensive and systematic evaluations. Our method consistently outperforms all baselines, including existing literature-based and data-driven approaches. Furthermore, human evaluations reveal that our generated hypotheses also improve human decision-making in challenging tasks.

7 Limitations

Our automated evaluation uses two recent models on datasets across various domains, showing the effectiveness of our method across diverse settings. However, we did not further evaluate our hypotheses on some tasks that require representations beyond natural language, such as math problem solving and code generation.

The literature corpus used for literature-based hypothesis generation is limited in terms of size and collection method. The collection is carried out by manually searching and collecting up to 10 papers on Semantic Scholar or Google Scholar. Though with the limited literature corpus we already show that our methods yield competent performance, a natural future direction is to enhance the literature component with automatic and scalable retrieval.

Similarly, we achieved satisfactory performance across different models and tasks with the initial set of hyperparameters. However, we did not perform an exhaustive hyperparameter search, which may have yielded further enhancements to the performance of our methods. This represents a limitation of our study that could be addressed in future work.

Our experiments with human subjects is a proof of concept. The number of participants in our human evaluation is relatively small. As a result, we do not believe that we have the statistical power to distinguish, for example, the difference between HYPOGENIC and HYPOREFINE. Although this is not the focus of our study, we encourage future work to conduct large-scale experiments in focused domains to validate the hypotheses generated through human-AI collaboration.

Last but not least, we manually chose three hypotheses through ablation-style study and subjective judgment for experiments with human subjects. We believe this process is the essence of human-AI collaboration in future scientific processes. It requires future exploration to identify the optimal collaboration regime.

Acknowledgments

We thank members of the Chicago Human+AI Lab for their helpful comments. We also thank the anonymous participants on Prolific for participating in our study. This work is supported in part by NSF grants IIS-2126602, 2302785 and an award from the Sloan Foundation.

References

Anirudh Ajith, Mengzhou Xia, Alexis Chevalier, Tanya Goyal, Danqi Chen, and Tianyu Gao. 2024. LitSearch: A retrieval benchmark for scientific literature search. *Preprint*, arXiv:2407.18940.

Peter Auer. 2003. Using confidence bounds for exploitation-exploration trade-offs. *J. Mach. Learn. Res.*, 3:397–422.

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

Rafael M. Batista and James Ross. 2024. Words that work: Using language to generate hypotheses.

Ziru Chen, Shijie Chen, Yuting Ning, Qianheng Zhang, Boshi Wang, Botao Yu, Yifei Li, Zeyi Liao, Chen Wei, Zitong Lu, Vishal Dey, Mingyi Xue, Frazier N. Baker, Benjamin Burns, Daniel Adu-Ampratwum, Xuhui Huang, Xia Ning, Song Gao, Yu Su, and Huan Sun. 2024. ScienceAgentBench: Toward rigorous assessment of language agents for data-driven scientific discovery. *Preprint*, arXiv:2410.05080.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019. Fine-grained analysis of propaganda in news article. In *Proceedings of EMNLP-IJCNLP*.

Son Doan, Amanda Ritchart, Nicholas Perry, Juan D Chaparro, and Mike Conway. 2017. How do you #relax when you're #stressed? a content analysis and infodemiology study of stress-related tweets. *JMIR Public Health and Surveillance*, 3(2):e35.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, and et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of ACL*.

Cristina Garbacea and Chenhao Tan. 2025. Hyperalign: Hypotheses-driven personalized alignment. *Preprint*, arXiv:2505.00038.

Alireza Ghafarollahi and Markus J. Buehler. 2024. SciAgents: Automating scientific discovery through multi-agent intelligent graph reasoning. *Preprint*, arXiv:2409.05556.

Pierpaolo Goffredo, Mariana Chaves, Serena Villata, and Elena Cabrio. 2023. Argument-based detection and classification of fallacies in political debates. In *Proceedings of EMNLP*.

Google. 2024. NotebookLM.

Pär Anders Granhag and Aldert Vrij. 2005. Deception detection. *Psychology and law: An empirical perspective*, pages 43–92.

Ken Gu, Ruoxi Shang, Ruien Jiang, Keying Kuang, Richard-John Lin, Donghe Lyu, Yue Mao, Youran Pan, Teng Wu, Jiaqian Yu, Yikun Zhang, Tianmai M. Zhang, Lanyi Zhu, Mike A. Merrill, Jeffrey Heer, and Tim Althoff. 2024. BLADE: Benchmarking language model agents for data-driven science. *Preprint*, arXiv:2408.09667.

Siyuan Guo, Cheng Deng, Ying Wen, Hechang Chen, Yi Chang, and Jun Wang. 2024. DS-Agent: Automated data science by empowering large language models with case-based reasoning. In *Proceedings of ICML*.

Xueyu Hu, Ziyu Zhao, Shuang Wei, Ziwei Chai, Qianli Ma, Guoyin Wang, Xuwu Wang, Jing Su, Jingjing Xu, Ming Zhu, Yao Cheng, Jianbo Yuan, Jiwei Li, Kun Kuang, Yang Yang, Hongxia Yang, and Fei Wu. 2024. InfiAgent-DABench: Evaluating agents on data analysis tasks. In *Proceedings of ICML*, Proceedings of Machine Learning Research.

Qian Huang, Jian Vora, Percy Liang, and Jure Leskovec. 2024. MLAgentBench: Evaluating language agents on machine learning experimentation. In *Proceedings of ICML*.

Peter Jansen, Marc-Alexandre Côté, Tushar Khot, Erin Bransom, Bhavana Dalvi Mishra, Bodhisattwa Prasad Majumder, Oyvind Tafjord, and Peter Clark. 2024. DISCOVERYWORLD: A virtual environment for developing and evaluating automated scientific discovery agents. *Preprint*, arXiv:2406.06769.

Hao Kang and Chenyan Xiong. 2024. ResearchArena: Benchmarking llms' ability to collect and organize information as research agents. *Preprint*, arXiv:2406.10291.

Jiwei Li, Myle Ott, and Claire Cardie. 2013. Identifying manipulated offerings on review portals. In *Proceedings of EMNLP*.

Jiwei Li, Myle Ott, Claire Cardie, and Eduard Hovy. 2014. Towards a general rule for identifying deceptive opinion spam. In *Proceedings of ACL*.

Mingxuan Li, Hanchen Li, and Chenhao Tan. 2025. Hypoeval: Hypothesis-guided evaluation for natural language generation. *Preprint*, arXiv:2504.07174.

Ruochen Li, Teerth Patel, Qingyun Wang, and Xinya Du. 2024. MLR-Copilot: Autonomous machine learning research based on large language models agents. *Preprint*, arXiv:2408.14033.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The semantic scholar open research corpus. In *Proceedings of ACL*, Online.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. 2024. The AI scientist: Towards fully automated open-ended scientific discovery. *Preprint*, arXiv:2408.06292.

Jens Ludwig and Sendhil Mullainathan. 2024. Machine learning as a tool for hypothesis generation*. *The Quarterly Journal of Economics*, page qjad055.

Sonia J Lupien, Bruce S McEwen, Megan R Gunnar, and Christine Heim. 2009. Effects of stress throughout the lifespan on the brain, behaviour and cognition. *Nature Reviews Neuroscience*, pages 434–445.

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

Sushil Kumar Maurya, Dinesh Singh, and Ashish Kumar Maurya. 2022. Deceptive opinion spam detection approaches: a literature survey. *Applied Intelligence*, 53(2):2189–2234.

OpenAI. 2023. GPT-4 technical report.

OthersideAI. 2024. HyperWrite.

Myle Ott, Claire Cardie, and Jeffrey T. Hancock. 2013. Negative deceptive opinion spam. In *Proceedings of NAACL*.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. 2024. Llm evaluators recognize and favor their own generations. *Preprint*, arXiv:2404.13076.

Amalie Brogaard Pauli, Isabelle Augenstein, and Ira Assent. 2024. Measuring and benchmarking large language models' capabilities to generate persuasive language. *Preprint*, arXiv:2406.17753.

Chau Minh Pham, Alexander Hoyle, Simeng Sun, and Mohit Iyyer. 2024. Topicgpt: A prompt-based topic modeling framework. In *Proceedings of NAACL*.

Martin Potthast, Tim Gollub, Kristof Komlossy, Sebastian Schuster, Matti Wiegmann, Erika Patricia Garces Fernandez, Matthias Hagen, and Benno Stein. 2018. Crowdsourcing a large corpus of clickbait on Twitter. In *Proceedings of COLING*.

Ori Press, Andreas Hochlehnert, Ameya Prabhu, Vishaal Udandarao, Ofir Press, and Matthias Bethge. 2024. CiteME: Can language models accurately cite scientific claims? *Preprint*, arXiv:2407.12861.

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. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.

Linlu Qiu, Liwei Jiang, Ximing Lu, Melanie Sclar, Valentina Pyatkin, Chandra Bhagavatula, Bailin Wang, Yoon Kim, Yejin Choi, Nouha Dziri, and Xiang Ren. 2024. Phenomenal yet puzzling: Testing inductive reasoning capabilities of language models with hypothesis refinement. In *Proceedings of ICLR*.

Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, M Pawan Kumar, Emilien Dupont, Francisco JR Ruiz, Jordan S Ellenberg, Pengming Wang, Omar Fawzi, et al. 2024. Mathematical discoveries from program search with large language models. *Nature*, 625(7995):468–475.

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

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of WWW*.

Ruixiang Tang, Yu-Neng Chuang, and Xia Hu. 2023. The science of detecting llm-generated texts. *Preprint*, arXiv:2303.07205.

Minyang Tian, Luyu Gao, Shizhuo Dylan Zhang, Xinan Chen, Cunwei Fan, Xuefei Guo, Roland Haas, Pan Ji, Kittithat Krongchon, Yao Li, Shengyan Liu, Di Luo, Yutao Ma, Hao Tong, Kha Trinh, Chenyu Tian, Zihan Wang, Bohao Wu, Yanyu Xiong, Shengzhu Yin, Minhui Zhu, Kilian Lieret, Yanxin Lu, Genglin Liu, Yufeng Du, Tianhua Tao, Ofir Press, Jamie Callan, Eliu Huerta, and Hao Peng. 2024. SciCode: A research coding benchmark curated by scientists. *Preprint*, arXiv:2407.13168.

Elsbeth Turcan and Kathleen McKeown. 2019. Dreaddit: A reddit dataset for stress analysis in social media. *Preprint*, arXiv:1911.00133.

Jean H.M. Wagemans. 2023. How to identify an argument type? on the hermeneutics of persuasive discourse. *Journal of Pragmatics*, 203:117–129.

Xiangxuan Wan and Li Tian. 2024. User stress detection using social media text: A novel machine learning approach. *International Journal of Computers Communications & Control*, 19.

Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. 2024. SciMON: Scientific inspiration machines optimized for novelty. In *Proceedings of ACL*.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2020. Persuasion for good: Towards a personalized persuasive dialogue system for social good. *Preprint*, arXiv:1906.06725.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Proceedings of NeurIPS*.

Junchao Wu, Shu Yang, Runzhe Zhan, Yulin Yuan, Derek F. Wong, and Lidia S. Chao. 2024. A survey on llm-generated text detection: Necessity, methods, and future directions. *Preprint*, arXiv:2310.14724.

Zonglin Yang, Li Dong, Xinya Du, Hao Cheng, Erik Cambria, Xiaodong Liu, Jianfeng Gao, and Furu Wei. 2024a. Language models as inductive reasoners. In *Proceedings of EACL*.

Zonglin Yang, Xinya Du, Junxian Li, Jie Zheng, Soujanya Poria, and Erik Cambria. 2024b. Large language models for automated open-domain scientific hypotheses discovery. In *Proceedings of ACL*.

Xingjian Zhang, Yutong Xie, Jin Huang, Jinge Ma, Zhaoying Pan, Qijia Liu, Ziyang Xiong, Tolga Ergen, Dongsub Shim, Honglak Lee, and Qiaozhu Mei. 2024. MASSW: A new dataset and benchmark tasks for ai-assisted scientific workflows. *Preprint*, arXiv:2406.06357.

Ruiqi Zhong, Peter Zhang, Steve Li, Jinwoo Ahn, Dan Klein, and Jacob Steinhardt. 2023. Goal driven discovery of distributional differences via language descriptions. In *Proceedings of NeurIPS*.

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

A Prompts

All our prompts for LLMs are separated into system prompts and user prompts. System prompts contain role and tone information, followed by detailed descriptions of the task and the expected response format. User prompts contain useful information for hypothesis generation, refinement, or inference, including information from literature, instances from datasets, and previously generated hypotheses. Below are some examples of the prompts that we use for each task.

A.1 Deception Detection

```
System Prompt
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 or deceptive. In other words,
 we want to know whether the review is written by
a someone who actually lived in the hotel.

Using the given examples, please propose
<num_hypotheses> possible hypothesis pairs.
These hypotheses should identify specific patterns
 that occur across the provided reviews.

Each hypothesis should contain a pair of the
following:
a. A hypothesis about what makes reviews more
likely to be truthful
b. The opposite hypothesis about what makes
reviews more likely to be deceptive

Generate them in the format of 1. [hypothesis], 2.
 [hypothesis], ... <num_hypotheses>. [hypothesis].
The hypotheses should analyze what kind of reviews
 are likely to be truthful or deceptive.

User Prompt
We have seen some hotel reviews:
··· more examples here ···
Please generate hypotheses that are useful for
predicting whether a review is truthful or
deceptive.
Propose <num_hypotheses> possible hypotheses.
Generate them in the format of 1. [hypothesis], 2.
 [hypothesis], ... <num_hypotheses>. [hypothesis].
Proposed hypotheses:
```

Example 1: Data-Based Hypothesis Generation with HypoGeniC.

```
System Prompt
You're a professional hotel review analyst.
Given some key findings from a series of research
papers, 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 a someone who
actually lived in the hotel.

Using the given relevant literatures, please
propose <num_hypotheses> possible hypothesis pairs.

These hypotheses should identify specific patterns
 that occur across the provided reviews.

Each hypothesis should contain a pair of the
following:
a. A hypothesis about what makes reviews more
likely to be truthful
b. The opposite hypothesis about what makes
reviews more likely to be deceptive

Generate them in the format of 1. [hypothesis], 2.
 [hypothesis], ... <num_hypotheses>. [hypothesis].
The hypotheses should analyze what kind of reviews
 are likely to be truthful or deceptive.

User Prompt
We have some key findings from a series of
research papers that might be useful for
generating the required <num_hypotheses>.
hypotheses:
··· information from literature here ···
Please generate hypotheses that are useful for
predicting whether a review is truthful or
deceptive.
When generating hypotheses, remember not to
overuse your own knowledge. Always refer to the
key findings from research papers provided.
Directly cite passages in the key findings when
generating a hypothesis.
Propose <num_hypotheses> possible hypotheses.
Remember to generate <num_hypotheses> hypotheses!
Generate them in the format of 1. [hypothesis], 2.
 [hypothesis], ... <num_hypotheses>. [hypothesis].
Proposed hypotheses:
```

Example 2: Literature-Based Hypothesis Generation.

```
System Prompt
You are a helpful assistant for summarizing key
findings in research papers on a given topic.

User Prompt
Summarize the following research paper, focusing
ONLY on this question: What is useful for one to
decide whether a review is truthful or deceptive
in real life?
Focus on hypotheses of what kind of reviews tend
to be deceptive, do not include technical details
in the paper.
...  literature texts here ...
```

Example 3: Paper Summarization.

```
System Prompt
You're a social scientist working on a project to
identify deceptive hotel reviews.
Given a set of hotel reviews, 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
a someone who actually lived in the hotel.

Using the given examples, refine the hypothesis
pairs provided.
The desired hypotheses should identify specific
patterns that occur across the provided reviews.

Each hypothesis should contain a pair of the
following:
a. A hypothesis about what makes reviews more
likely to be truthful
b. The opposite hypothesis about what makes
reviews more likely to be deceptive

Generate refined hypotheses in the format of 1. [
hypothesis], 2. [hypothesis], ... <num_hypotheses>.
 [hypothesis].
The hypotheses should analyze what kind of reviews
 are likely to be truthful or deceptive.

User Prompt
We have seen some hotel reviews:
··· more examples here ···
We have some hypotheses need to be refined:
...  hypotheses to be refined here ...
Please refine these hypotheses to make them more
specific and useful for predicting whether a
review is truthful or deceptive.
When refining the hypotheses, feel free to change
the key information or topic of a hypothesis based
 on the provided prevailing patterns in data if
you think it is necessary.
Generate refined hypotheses in the format of 1. [
hypothesis], 2. [hypothesis], ... <num_hypotheses>.
 [hypothesis].
```

```
Refined hypotheses:
```

Example 4: Hypothesis Refinement Based on Data.

```
System Prompt
You're a social scientist working on a project to
identify deceptive hotel reviews.
Given a set of hotel reviews, 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
a someone who actually lived in the hotel.

Using the given relevant literatures, refine the
hypothesis pairs provided.
The desired hypotheses should identify specific
patterns that occur across the provided reviews.

Each hypothesis should contain a pair of the
following:
a. A hypothesis about what makes reviews more
likely to be truthful
b. The opposite hypothesis about what makes
reviews more likely to be deceptive

Generate refined hypotheses in the format of 1. [
hypothesis], 2. [hypothesis], ... <num_hypotheses>.
 [hypothesis].
The hypotheses should analyze what kind of reviews
 are likely to be truthful or deceptive.

User Prompt
We have some key findings from a series of
research papers that might be useful for
generating hypotheses:
··· information from literature here ···
We have some hypotheses need to be refined:
...  hypotheses to be refined here ...
Please refine these hypotheses to make them more
specific and useful for predicting whether a
review is truthful or deceptive.
When refining the hypotheses, feel free to change
the key information or topic of a hypothesis based
 on the provided key findings if you think it is
necessary.
Generate refined hypotheses in the format of 1. [
hypothesis], 2. [hypothesis], ... <num_hypotheses>.
 [hypothesis].
Refined hypotheses:
```

Example 5: Hypothesis Refinement Based on Literature.

```
System Prompt
You are a professional deceptive detection agent
and your job is to determine whether a hotel
review is truthful or deceptive.
In other words, we want to know whether the review
 is written by someone who had real experiences
with the hotel.
From past experiences, you learned some patterns.
You need to determine whether each of the patterns
 holds for the current hotel review, and also
predict whether the current hotel review is
truthful or deceptive.
Give an answer. The answer should be one word (
truthful or deceptive).
Give your final answer in the format of {Final
answer: answer}

User Prompt
Our learned patterns: <generated_hypotheses>
A hotel review is the following: <review>

Given the pattern you learned above, give an
answer of whether the hotel review above is
deceptive or truthful.
Think step by step.
First step: Think about which pattern can be
applied to the hotel review.
Second step: Based on the pattern, is this hotel
review deceptive or truthful?
Final step: give your final answer in the format
of {Final answer: answer}
```

Example 6: Multiple-Hypothesis-Based Inference.

A.2 AI-Generated Content (AIGC) Detection

```
System Prompt
You are a professional writer.
You will be given a story prompt. Your task is to
write a coherent and creative story based on the
prompt. Ensure the story has a clear structure,
engaging characters, and stays true to the prompt'
s core idea. Keep the language clear and
appropriate to the story's tone.

User Prompt
...  story-writing prompt here ...
example:
[ WP ] You 've been able to read minds since you
turned 7 . Mostly you watch people 's thoughts
passively and undetected but one day someone talks
 back .\n
```

Example 7: AIGC Detection Dataset Generation.

```
System Prompt
You're a an AI generated content detection expert.
 You are great at detecting what type of text is
generated by AI.
Given a set of texts, we want to generate
hypotheses that are useful for predicting whether
a piece of text is generated by AI. In other words,
 we want to know whether the text is written by a
human or generated by AI.

Your task is to identify what patterns or traits
show up more in AI generated texts, and what shows
 up more in human written texts.  Focus on the
generalizable insight that can be applied in other
 contexts. Ignore things that are specific to this
 story. Do not make references this story they may
 not be for others.

Using the given examples, please propose
<num_hypotheses> possible hypothesis pairs.
When proposing hypothesis, look closely into the
given examples and identify specific patterns that
 occur across the provided text examples.
The hypotheses should be clear, easy to understand,
 and have specific details such that one can apply
 the hypotheses to predict whether a piece of text
 is written by human or AI.

Generate them in the format of 1. [hypothesis], 2.
 [hypothesis], ... <num_hypotheses>. [hypothesis].
The hypotheses should analyze what kind of text is
 likely to be written by human or AI.

User Prompt
We have seen some texts:
...  more examples here ...
Please generate hypotheses that are useful for
predicting predicting whether a piece of text is
written by human or AI.
Propose <num_hypotheses> possible hypotheses.
Generate them in the format of 1. [hypothesis], 2.
 [hypothesis], ... <num_hypotheses>. [hypothesis].

When proposing hypothesis, look closely into the
given examples and identify specific patterns that
 occur across the provided text examples.

Please make sure that the hypotheses are:
i. clear (i.e., precise , not too wordy , and easy
 to understand);
ii. generalizable to novel situations (i.e., they
would make sense if applied to other AI generated
content detection experiments or other messaging
contexts);
iii. empirically plausible (i.e., this is a
dimension on which messages can vary on);
iv. unidimensional (i.e., avoid hypotheses that
list multiple constructs so if there are many
things changing , pick one);
v. usable (i.e., a human equipped with this
insight could use it to predict if a new piece of
text is generated AI in a similar way)

Proposed hypotheses:
```

Example 8: Data-Based Hypothesis Generation with HypoGeniC.

```
System Prompt
You're a professional AI content detector.
Given some key findings from a series of research
papers, we want to generate hypotheses that are
useful for detecting whether a piece of text is
written by human or AI.

Your task is to identify what patterns or traits
show up more in AI generated texts, and what shows
 up more in human written texts.  Focus on the
generalizable insight that can be applied in other
 contexts. Ignore things that are specific to this
 story. Do not make references this story they may
 not be for others.

Using the given relevant literatures, please
propose <num_hypotheses> possible hypothesis pairs.

These hypotheses should identify specific patterns
 that occur across the provided texts.

Generate them in the format of 1. [hypothesis], 2.
 [hypothesis], ... <num_hypotheses>. [hypothesis].
The hypotheses should analyze what kind of text is
 likely to be written by human or AI.

User Prompt
We have some key findings from a series of
research papers that might be useful for
generating the required <num_hypotheses>
hypotheses:
··· information from literature here ···
Please generate hypotheses that are useful for
predicting whether a piece of text is written of
human or AI.
Propose <num_hypotheses> possible hypotheses.
Remember to generate <num_hypotheses> hypotheses!
Generate them in the format of 1. [hypothesis], 2.
 [hypothesis], ... <num_hypotheses>. [hypothesis].
Proposed hypotheses:
```

Example 9: Literature-Based Hypothesis Generation.

```
System Prompt
You are a helpful assistant for summarizing key
findings in research papers on a given topic.

User Prompt
Summarize the following research paper, focusing
ONLY on this question: What is useful for one to
detect whether some text is generated by AI?
Focus on hypotheses of what kind of text tend to
be generated by AI, do not include technical
details in the paper.
...  literature texts here ...
```

Example 10: Paper Summarization.

```
System Prompt
You're a an AI generated content detection expert.
 You are great at detecting what type of text is
generated by AI.
Given a set of texts, we want to generate
hypotheses that are useful for predicting whether
a piece of text is generated by AI. In other words,
 we want to know whether the text is written by a
human or generated by AI.

Using the given examples, refine the hypothesis
pairs provided.
The desired hypotheses should identify specific
patterns that occur across the provided text
examples.

Generate refined hypotheses in the format of 1. [
hypothesis], 2. [hypothesis], ... <num_hypotheses>.
 [hypothesis].
The hypotheses should analyze what kind of text is
 likely to be written by human or AI.

User Prompt
We have seen some texts:
··· more examples here ···
We have some hypotheses need to be refined:
...  hypotheses to be refined here ...
Please refine these hypotheses to make them more
specific and useful for predicting whether a piece
 of text is written by human or AI.
When refining the hypotheses, feel free to change
the key information or topic of a hypothesis based
 on the provided prevailing patterns in data if
you think it is necessary.
Generate refined hypotheses in the format of 1. [
hypothesis], 2. [hypothesis], ... <num_hypotheses>.
 [hypothesis].
Refined hypotheses:
```

Example 11: Hypothesis Refinement Based on Data.

```
System Prompt
You're a an AI generated content detection expert.
 You are great at detecting what type of text is
generated by AI.
Given a set of texts, we want to generate
hypotheses that are useful for predicting whether
a piece of text is generated by AI. In other words,
 we want to know whether the text is written by a
human or generated by AI.

Using the given relevant literatures, refine the
hypothesis pairs provided.
The desired hypotheses should identify specific
patterns that occur across the provided text
examples.

Generate refined hypotheses in the format of 1. [
hypothesis], 2. [hypothesis], ... <num_hypotheses>.
 [hypothesis].
The hypotheses should analyze what kind of text is
 likely to be written by human or AI.

User Prompt
We have some key findings from a series of
research papers that might be useful for
generating hypotheses:
··· information from literature here ···
We have some hypotheses need to be refined:
...  hypotheses to be refined here ...
Please refine these hypotheses to make them more
specific and useful for predicting whether a piece
 of text is written by human or AI.
When refining the hypotheses, feel free to change
the key information or topic of a hypothesis based
 on the provided prevailing patterns in data if
you think it is necessary.
Generate refined hypotheses in the format of 1. [
hypothesis], 2. [hypothesis], ... <num_hypotheses>.
 [hypothesis].
Refined hypotheses:
```

Example 12: Hypothesis Refinement Based on Literature.

```
System Prompt
You are an AI generated content detection agent
and want to determine whether a piece of text is
written by a human or generated by an AI. In other
 words, we want to know whether the text is
written by a human or generated by AI.
From past experiences, you learned some patterns.
You need to determine whether each of the patterns
 holds for the current text, and also predict
whether the current text is written by human or AI.

Give an answer. The answer should be one word (AI
or HUMAN).
```

```
Give your final answer in the format of "Final
answer: ANSWER"

User Prompt
Our learned patterns: <generated_hypotheses>
New text:
Here is a story: <story>

Given the patterns you learned above, give an
answer of whether the current text is written by
human or AI.
Think step by step.
First step: Think about which pattern can be
applied to the story.
Second step: Based on the pattern, is this story
written by human or AI?
You must give your final answer in the format of "
Final answer: ANSWER".
```

Example 13: Multiple-Hypothesis-Based Inference.

A.3 Mental Stress Detection

```
System Prompt
You're a psychologist and social scientist
studying people's stress and their online posts.
given a set of reddit posts, we want to generate
hypotheses that are useful for deciding people's
stress status (has stress or no stress) based on
reddit post.

Using the given examples, please propose
<num_hypotheses> possible hypothesis pairs.
These hypotheses should identify specific patterns
 that occur across the provided posts.

Each hypothesis should contain a pair of the
following:
a. A hypothesis about what makes the post more
likely to indicate that the poster has stress
b. The opposite hypothesis about what makes the
post more likely to indicate that the poster does
not have stress

Generate them in the format of 1. [hypothesis], 2.
 [hypothesis], ... <num_hypotheses>. [hypothesis].
The hypotheses should analyze what kind of posts
are likely to indicate stress or no stress.

User Prompt
We have seen some reddit posts:
··· more examples here ···
Please generate hypotheses that are useful for
deciding people's stress status (has stress or no
stress) based on reddit post.
Propose <num_hypotheses> possible hypotheses.
Generate them in the format of 1. [hypothesis], 2.
 [hypothesis], ... <num_hypotheses>. [hypothesis].
Proposed hypotheses:
```

Example 14: Data-Based Hypothesis Generation with HypoGeniC.

```
System Prompt
You're a psychologist and social scientist
studying people's stress and their online posts.
Given some key findings from a series of research
papers, we want to generate hypotheses that are
useful for deciding people's stress status (has
stress or no stress) based on reddit post.

Using the given relevant literatures, please
propose <num_hypotheses> possible hypothesis pairs.

These hypotheses should identify specific patterns
 that occur across the provided posts.

Each hypothesis should contain a pair of the
following:
a. A hypothesis about what makes the post more
likely to indicate that the poster has stress
b. The opposite hypothesis about what makes the
post more likely to indicate that the poster does
not have stress

Generate them in the format of 1. [hypothesis], 2.
 [hypothesis], ... <num_hypotheses>. [hypothesis].
The hypotheses should analyze what kind of posts
are likely to indicate stress or no stress.

User Prompt
We have some key findings from a series of
research papers that might be useful for
generating the required <num_hypotheses>
hypotheses:
··· information from literature here ···
Please generate hypotheses that are useful for
deciding people's stress status (has stress or no
stress) based on reddit post.
Propose <num_hypotheses> possible hypotheses.
Remember to generate <num_hypotheses> hypotheses!
Generate them in the format of 1. [hypothesis], 2.
 [hypothesis], ... <num_hypotheses>. [hypothesis].
Proposed hypotheses:
```

Example 15: Literature-Based Hypothesis Generation.

```
System Prompt
You are a helpful assistant for summarizing key
findings in research papers on a given topic.

User Prompt
Summarize the following research paper, focusing
ONLY on this question: What is useful for one to
judge whether a reddit poster has stress based on
one of their reddit post content?
Focus on hypotheses of what kind of posts indicate
 stress, do not include technical details in the
paper.
...  literature texts here ...
```

Example 16: Paper Summarization.

```
System Prompt
You're a psychologist and social scientist working
 on a project to identify whether a person has
stress based on reddit posts.
given a set of reddit posts, we want to generate
hypotheses that are useful for deciding people's
stress status (has stress or no stress) based on
reddit post.

Using the given examples, refine the hypothesis
pairs provided.
The desired hypotheses should identify specific
patterns that occur across the provided posts.

Each hypothesis should contain a pair of the
following:
a. A hypothesis about what makes the post more
likely to indicate that the poster has stress
b. The opposite hypothesis about what makes the
post more likely to indicate that the poster does
not have stress

Generate refined hypotheses in the format of 1. [
hypothesis], 2. [hypothesis], ... <num_hypotheses>.
 [hypothesis].
The hypotheses should analyze what kind of posts
are likely to indicate stress or no stress.

User Prompt
We have seen some reddit posts:
··· more examples here ···
We have some hypotheses need to be refined:
...  hypotheses to be refined here ...
Please refine these hypotheses to make them more
specific and useful for deciding people's stress
status (has stress or no stress) based on reddit
post.
Generate refined hypotheses in the format of 1. [
hypothesis], 2. [hypothesis], ... <num_hypotheses>.
 [hypothesis].
Refined hypotheses:
```

Example 17: Hypothesis Refinement Based on Data.

```
System Prompt
You're a psychologist and social scientist working
 on a project to identify whether a person has
stress based on reddit posts.
given a set of reddit posts, we want to generate
hypotheses that are useful for deciding people's
stress status (has stress or no stress) based on
reddit post.

Using the given relevant literatures, refine the
hypothesis pairs provided.
The desired hypotheses should identify specific
patterns that occur across the provided posts.

Each hypothesis should contain a pair of the
following:
a. A hypothesis about what makes the post more
likely to indicate that the poster has stress
b. The opposite hypothesis about what makes the
post more likely to indicate that the poster does
not have stress

Generate refined hypotheses in the format of 1. [
hypothesis], 2. [hypothesis], ... <num_hypotheses>.
 [hypothesis].
The hypotheses should analyze what kind of posts
are likely to indicate stress or no stress.

User Prompt
We have some key findings from a series of
research papers that might be useful for
generating hypotheses:
··· information from literature here ···
We have some hypotheses need to be refined:
...  hypotheses to be refined here ...
Please refine these hypotheses to make them more
specific and useful for deciding people's stress
status (has stress or no stress) based on reddit
post.
Generate refined hypotheses in the format of 1. [
hypothesis], 2. [hypothesis], ... <num_hypotheses>.
 [hypothesis].
Refined hypotheses:
```

Example 18: Hypothesis Refinement Based on Literature.

```
System Prompt
You're a psychologist and social scientist working
 on a project to identify whether a person has
stress based on reddit posts.
From past experiences, you learned some patterns.
You need to determine whether each of the patterns
 holds for the current reddit post, and also
predict whether the poster of the reddit post has
stress or not based on the content of the post.
Give an answer. The answer should be "has stress"
or "no stress".
Give your final answer in the format of {Final
answer: answer}

User Prompt
Our learned patterns: <generated_hypotheses>
A reddit post is the following: <post>

Given the pattern you learned above, give an
answer of whether the poster of the reddit post
has stress or not based on the content of the post.

Think step by step.
First step: Think about which pattern can be
applied to the reddit post.
Second step: Based on the pattern, does the poster
 of a reddit post has stress or not? Answer should
 be "has stress" or "no stress".
Final step: give your final answer in the format
of {Final answer: answer}
```

Example 19: Multiple-Hypothesis-Based Inference.

A.4 Persuasive Argument Prediction

```
System Prompt
You are an intelligent rhetorician and debater who
 masters persuasiveness in language.
Given a pair of arguments, you are asked to
determine which one of them uses more persuasive
language. The two arguments are often on the same
topic and are similar, so focus on their
differences.
What difference between the two arguments makes
one more persuasive than the other?
You will be given a set of observations of the
format:
Argument 1: [argument_1]
Argument 2: [argument_2]
Observation: The first/second argument uses more
persuasive language.
Based on the observations, please generate
hypotheses that are useful for explaining why one
argument uses more persuasive language than the
other.
These hypotheses should identify patterns, phrases,
 wordings etc. that occur across the provided
examples. They should also be generalizable to new
 instances.
Please propose <num_hypotheses> possible
hypotheses and generate them in the format of 1. [
hypothesis], 2. [hypothesis], ... <num_hypotheses>.
 [hypothesis].

User Prompt
Here are the Observations:
··· more examples here ···

Please generate hypotheses that can help determine
 which argument uses more persuasive language.
Please propose <num_hypotheses> possible
hypotheses.

Generate them in the format of 1. [hypothesis], 2.
 [hypothesis], ... <num_hypotheses>. [hypothesis].

Proposed hypotheses:
```

Example 20: Data-Based Hypothesis Generation with HypoGeniC.

```
System Prompt
You are an intelligent rhetorician and debater who
 masters persuasiveness in language.
Given a pair of arguments, you are asked to
determine which one of them uses more persuasive
language. The two arguments are often on the same
topic and are similar, so focus on their
differences.
What difference between the two arguments makes
one more persuasive than the other?
You will be given a set of literature of the
format:
Title: [title]
Key Findings: [summary]
Based on the literature, please generate
hypotheses that are useful for explaining why one
argument uses more persuasive language than the
other.
These hypotheses should identify patterns, phrases,
 wordings etc. that you can find in the literature.
 They should also be generalizable to new
instances.
Please propose <num_hypotheses> refined hypotheses
 and generate them in the format of 1. [hypothesis
], 2. [hypothesis], ... <num_hypotheses>. [
hypothesis].

User Prompt
Here are some key findings from a series of
research papers that might be useful for
generating hypotheses:
··· information from literature here ···

Please generate hypotheses that can help determine
 which argument uses more persuasive language.
Please propose <num_hypotheses> possible
hypotheses.

Generate them in the format of 1. [hypothesis], 2.
```

```
 [hypothesis], ... <num_hypotheses>. [hypothesis].

Proposed hypotheses:
```

Example 21: Literature-Based Hypothesis Generation.

```
System Prompt
You are a helpful assistant for summarizing key
findings in research papers on a given topic.

User Prompt
Summarize the following research paper, focusing
ONLY on this question: What characterizes texts
that use more persuasive language? In other words,
 how can one determine which one of two sentences
uses more persuasive language?
Focus on hypotheses of what characterizes texts
that use more persuasive language, do not include
technical details in the paper.
...  literature texts here ...
```

Example 22: Paper Summarization.

```
System Prompt
You are an intelligent rhetorician and debater who
 masters persuasiveness in language.
Given a pair of arguments, you are asked to
determine which one of them uses more persuasive
language. The two arguments are often on the same
topic and are similar, so focus on their
differences.
What difference between the two arguments makes
one more persuasive than the other?
You will be given a set of observations of the
format:
Argument 1: [argument_1]
Argument 2: [argument_2]
Observation: The first/second argument uses more
persuasive language.
Based on the observations, please refine
hypotheses provided to make them more useful for
explaining why one argument uses more persuasive
language than the other.
These hypotheses should identify patterns, phrases,
 wordings etc. that occur across the provided
examples. They should also be generalizable to new
 instances.
Please propose <num_hypotheses> refined hypotheses
 and generate them in the format of 1. [hypothesis
], 2. [hypothesis], ... <num_hypotheses>. [
hypothesis].

User Prompt
Here are the Observations:
··· more examples here ···

And here are the previous hypotheses:
...  hypotheses to be refined here ...

Please generate refined hypotheses that can help
determine which argument uses more persuasive
language.
Please propose <num_hypotheses> refined hypotheses.


Generate them in the format of 1. [hypothesis], 2.
 [hypothesis], ... <num_hypotheses>. [hypothesis].

Refined hypotheses:
```

Example 23: Hypothesis Refinement Based on Data.

```
System Prompt
You are an intelligent rhetorician and debater who
 masters persuasiveness in language.
Given a pair of arguments, you are asked to
determine which one of them uses more persuasive
language. The two arguments are often on the same
topic and are similar, so focus on their
differences.
What difference between the two arguments makes
one more persuasive than the other?
You will be given a set of literature of the
format:
··· information from literature here ···
Based on the literature, please refine hypotheses
provided to make them more useful for explaining
why one argument uses more persuasive language
than the other.
These hypotheses should identify patterns, phrases,
 wordings etc. that you can find in the literature.
 They should also be generalizable to new
instances.
Please propose <num_hypotheses> refined hypotheses
 and generate them in the format of 1. [hypothesis
], 2. [hypothesis], ... <num_hypotheses>. [
hypothesis].

User Prompt
Here are some key findings from a series of
research papers that might be useful for
generating hypotheses:
··· information from literature here ···

And here are the previous hypotheses:
...  hypotheses to be refined here ...

Please generate refined hypotheses that can help
determine which argument uses more persuasive
language.
Please propose <num_hypotheses> refined hypotheses.


Generate them in the format of 1. [hypothesis], 2.
 [hypothesis], ... <num_hypotheses>. [hypothesis].

Refined hypotheses:
```

Example 24: Hypothesis Refinement Based on Literature.

```
System Prompt
You are an intelligent rhetorician and debater who
 masters persuasiveness in language.
Given a pair of arguments, you are asked to
determine which one of them uses more persuasive
language. The two arguments are often on the same
topic and are similar, so focus on their
differences.
From past experiences, you learned some patterns.
Now, at each time, you should apply the learned
patterns to a new pair of arguments and determine
which one uses more persuasive language.
The answer for the more persuasive language should
 be of the form "the _ argument" where _ is either
 first or second.
Please give your final answer in the format of {
Final answer: the _ argument uses more persuasive
language}

User Prompt
Our learned patterns: <generated_hypotheses>
Given the patterns you learned above, determine
which of the following arguments uses more
persuasive language:
Argument 1: <first_argument>
Argument 2: <second_argument>

Think step by step.
Step 1: Think about which learned patterns can be
applied to the arguments.
Step 2: Analyze the difference between "Argument
1" and "Argument 2".
Step 3: Based on the pattern, which argument uses
more persuasive language?
You MUST give your final answer in the following
format:
Final answer: the _ argument uses more persuasive
language.
```

Example 25: Multiple-Hypothesis-Based Inference.

B Automated Experiments Implementation Details

B.1 Partitioning of IND and OOD datasets

Deception Detection As stated in Section 4.1, our IND datasets are from DECEPTIVE REVIEWS, which contains 800 truthful hotel reviews from the web and 800 deceptive reviews gathered from Mechanical Turk (Ott et al., 2013). The OOD dataset, FOUR-CITIES (Li et al., 2013) consists of 640 hotel reviews from four cities and different web sources following the same procedure as DECEPTIVE REVIEWS.

AI-Generated Content (AIGC) Detection As discussed in Section 4.1, our AIGC Detection task consists of two subtasks, GPTGC and LlamaGC. The IND dataset for GPTGC contains GPT generated stories and human-written stories, while the one for LlamaGC includes Llama-generated stories and human-written stories. The OOD dataset of GPTGC is the IND dataset of LlamaGC and vice versa.

Mental Stress Detection IND and OOD datasets are separated based on the source subreddits, or topic-specific communities, from which the Reddit posts are collected. Instances of the IND dataset are from ptsd, anxiety, and domestic violence subreddits, while those of OOD dataset are from relationships and homeless subreddits.

Persuasive Argument Prediction IND and OOD datasets are partitioned according to the source corpora of the non-LLM-generated texts. Examples from IND datasets are from ElecDeb60to20 (Goffredo et al., 2023), Persuasion For Good (Wang et al., 2020), and Webis-Clickbait-17 (Potthast et al., 2018), while OOD dataset is from PT-Corpus (Da San Martino et al., 2019).

B.2 Specificity Boost

We further observed that sometimes the solely literature-based hypotheses generated by gpt-4o-mini are often too short and brief, making it harder to apply during inference. To address this, we add a LLM-based specificity booster after the literature-based hypothesis generation that adds more concrete illustrations and examples to each of the hypotheses based solely on its pre-training knowledge. Specifically we apply the specificity booster on our Deception Detection, Mental Stress Detection, and Persuasive Argument Prediction tasks. The specificity booster is not applied to Llama-3.1-70B-Instruct because it can already generate reasonably specific hypotheses.

B.3 Refinement and Union Implementation

Refinement is implemented as an extension based on the original HypoGeniC pipeline. During the initialization stage, an LLM $\mathcal{M}_G$ is instructed to generate an initial hypothesis bank $\mathcal{H}^0_{\mathcal{L}+\mathcal{D}}$ based on a set of initial examples $\mathcal{D}_{\text{init}}$ and a series of generated paper summaries $\mathcal{S}$, i.e., $\mathcal{H}^0_{\mathcal{L}+\mathcal{D}} = \mathcal{M}_G(\mathcal{S}, \mathcal{D}_{\text{init}})$. These initial hypotheses are then evaluated and re-ranked using the same reward function as in HypoGeniC. In the update stage at time t, if the size of the wrong examples bank $\mathcal{W}$ reaches $w_{\max}$, a set of new hypotheses is generated by feeding both the wrong examples bank and paper summaries to $\mathcal{M}_G$. $\mathcal{H}^t_{\mathcal{L}+\mathcal{D}}$ is then updated with the new hypothesis according to the reward, following the same procedure as HypoGeniC.

Union and Redundancy Elimination is implemented by combining the hypothesis bank generated using HYPOGENIC $\mathcal{H}_{\mathcal{D}}$ or HYPOREFINE $\mathcal{H}_{\mathcal{L}+\mathcal{D}}$and the bank generated by our literature-based hypothesis generation method. We first generate the two hypothesis banks separately using HYPOGENIC, HYPOREFINE, and LITERATURE-ONLY, following the procedures described above and in Section 3. Each hypothesis bank is then fed to a redundancy checker module. For a hypothesis bank of size 20, the LLM-based redundancy checker checks each pair of hypotheses and see if one entails the other, with results recorded as a 20×20 matrix $\mathcal{A}$ of 1 (redundant) or 0 (not redundant). To create the new no-redundancy hypothesis bank $\mathcal{H}_{\text{new}}$, we first rank the hypotheses based on their training accuracy. Each time we take the best-performing hypothesis h out of the original hypothesis bank $\mathcal{H}$ and check if there exists a hypothesis h_{new} in $\mathcal{H}_{\text{new}}$ such that redundancy is recorded in $\mathcal{A}$ for the pair h and h_{new}, i.e., $\mathcal{A}_{h,h_{\text{new}}} = 1$ or $\mathcal{A}_{h_{\text{new}},h} = 1$. If yes, h is moved out of the original bank $\mathcal{H}$ and skipped; if not, h is moved to $\mathcal{H}_{\text{new}}$ with a rank determined by its training accuracy.

After removing redundancies of hypothesis banks, we unite two hypothesis banks to create a final bank $\mathcal{H}_{\text{final}}$ with a balanced prioritization strategy. We first move the top 10 hypotheses from the HYPOGENIC or HYPOREFINE hypothesis bank to $\mathcal{H}_{\text{final}}$. If there is less than 10 hypotheses in the banks, we move all hypotheses to $\mathcal{H}_{\text{final}}$.

Then we randomly choose hypotheses from the literature-based hypothesis bank until the size of $\mathcal{H}_{\text{final}}$ reaches 20.

B.4 Multiple-Hypothesis Inference Implementation

During multiple-hypothesis based inference, each time we feed a LLM with our final hypothesis bank $\mathcal{H}$ of size 20 (see Appendix B.5) and an instance $(x, \_)$ of our IND or OOD datasets with labels removed. The LLM $\mathcal{M}_I$ is asked to generate an answer for the given instance using Chain-of-Thought prompting (Wei et al., 2022) that considers both the relevance of the hypotheses to the given instance and the utility of the hypothesis bank (see Appendix A for the exact prompts we used). The prediction is denoted as $\hat{y} = \mathcal{M}_I(\mathcal{H}, x)$. We then compute the average accuracy for all data instances in the held-out IND and OOD sets with the model predictions. For F1 scores, we report the macro-averaged F1 scores.

B.5 Technical Details of NotebookLM and HyperWrite

NotebookLM is an LLM-powered research assistance tool that generates source-grounding responses to user prompts. Specifically in our case, collected literature are uploaded in the NotebookLM interface, followed by a hypothesis generation prompt asking to generate hypotheses based on given literature. Given its functionality and our usage, it is placed under the literature-based hypothesis generation category in our evaluations.

For HyperWrite, we use its Hypothesis Maker function, which is an AI-driven tool that generates hypotheses based on a given research question. Though there is no publicly available technical report for this tool, it generally leverages LLM's pretraining knowledge and literature information to produce hypotheses.

B.6 Hyperparameters

We use the same set of hyperparameters across all tasks, models, and methods.

During the training stage of HypoGeniC, the limit of the hypothesis bank size, $H_{\max}$, is set to 20, and the size of training set is set to 200. In the initialization stage, we set `num_init` $= 10$. In the update stage, we use reward coefficient $\alpha = 0.5$, $w_{\max} = 10$, $k = 10$, and generate 10 hypothesis per update.

In our HYPOREFINE method, the round of refinement N_{refine} is set to 6.

We use 5 random seeds for multiple-hypothesis inference: 11376, 8271, 39660, 543, 3.

Across all tasks and methods and for both GPT-4o-mini and Llama-3.1-70B-Instruct, we use `temperature` $= 1 \times 10^{-5}$ and `max_tokens` $=$ 4000.

B.7 Licensing Details

DECEPTIVE REVIEWS is released under CC BY-NC-SA 3.0, and PERSUASIVE PAIRS is released under CC BY-NC 4.0. The WRITINGPROMPTS dataset which we use to create the AIGC Detection datasets are under MIT License. The LLAMAGC and GPTGC datasets will be released under the same licensing as this work, CC BY 4.0 License, should it be accepted. DREADDIT and FOUR-CITIES do not have licenses specified in their original papers, but are considered under CC BY 4.0 and CC BY-NC-SA 3.0 license respectively as they are ACL materials.

For the LLMs, GPT-4-MINI is a proprietary and not released under any open-source license, while LLAMA-70B-I is released under Llama 3.1 Community License Agreement.

Throughout our study, we find that we are in compliance with the licensing agreements of all the datasets and models used in this work.

B.8 Estimated Cost

For LLAMA-70B-I, we run all of our experiments with 4 NVIDIA A100s, and it takes on average 1.5 hours to run all of our hypothesis generation pipelines, including HYPOGENIC, HYPOREFINE, LITERATURE∪HYPOGENIC , and LITERATURE∪HYPOREFINE . With GPT-4-MINI, the average cost for running the same pipelines is $0.6.

C Human Study Details

C.1 Decision-making Utility Study Details

The instructions of the practical relevance study can be found in Figure 6 and Figure 8. For the interface, we present an example of the control group interface for Deception Detection in Figure 7, and examples of the experiment group interface in Figure 9.

The subjects of the control group are instructed to perform deception detection or AIGC (GPTGC) detection tasks without any assistance from the hypotheses. Subjects in the experiment group are

Hypotheses	Frequency of Selection
Hypothesis 1: Reviews that present a balanced perspective by detailing both positive and negative experiences with specific examples (e.g., "the room was spacious and clean, but the noise from the street was disruptive at night") are more likely to be truthful, whereas reviews that express extreme sentiments without acknowledging any redeeming qualities (e.g., "everything was perfect" or "it was a total disaster") are more likely to be deceptive.	50.00%
Hypothesis 2: Reviews that mention specific dates of stay or unique circumstances surrounding the visit (e.g., "We stayed during the busy Memorial Day weekend and faced long lines") are more likely to be truthful, while reviews that use vague temporal references (e.g., "I stayed recently") without concrete details are more likely to be deceptive, as they often lack the specificity that suggests a real and engaged experience.	34.44%
Hypothesis 3: Reviews that provide detailed sensory descriptions of the hotel experience, such as the specific decor of the room, the quality of bedding, and the overall ambiance (e.g., "the room featured luxurious furnishings, high-thread-count sheets, and soft lighting that created a relaxing atmosphere") are more likely to be truthful, while reviews that use vague or overly simplistic descriptors (e.g., "the hotel was nice and comfortable") are more likely to be deceptive.	46.39%
No hypothesis selected	7.50%

Table 6: How often humans use hypotheses in Deception Detection human study. We allow users to select multiple hypotheses for each instance they make prediction on, so the total frequency can exceed 100%.

asked to first read the presented 3 hypotheses and then make their predictions on the given instance. They are then required to choose which ones, if any, of the hypotheses that were used in their prediction. At the end of the study, participants in the experiment group are also asked to give overall rating and assessment of the helpfulness of the given hypotheses. There are five scales: "Not at all helpful", "Slightly helpful", "Moderately helpful", "Very helpful", and "Extremely helpful".

We choose top 3 hypotheses from the hypothesis bank generated using LITERATURE∪HYPOREFINE that cause the greatest drop in performance when removed from the hypotheses pool during multi-hypothesis inference. The chosen hypotheses for Deception Detection and AIGC Detection can be found in Table 5 and Table 6.

We recruit 30 participants for the control group and 30 for the experimental group. For the control group, 4 people timed out, and 25 out of the remaining 26 participants passed attention checks. For the experimental group, 3 people timed out, and 22 out of the remaining 27 passed attention checks. We compute human accuracy based on responses from people who finished tasks in time and passed attention checks. The average time spent is around 25 minutes and participants are timed out by the system if they spend more than 60 minutes in the study, which can happen when they accidentally leave the study website tab open but forget to do the task.

C.2 Likert Rating Survey Details

We provide summaries of existing findings in Deception Detection to annotators and ask them to rate the hypotheses in terms of clarity, novelty, and plausibility. Each metric has five scales, which we include in Table 7. In particular, we manually select five representative hypotheses from the set of hypotheses generated by the different methods. We report the average human ratings in Table 4. The instructions can be found in figure Figure 2, and the interface for annotation can be found in Figure 3.

C.3 Novelty and Nuance Study Details

For the Novelty and Nuance Study, we present the instructions for AIGC Detection in Figure 4. We showcase the interfaces for AIGC Detection in Figure 5.

For both Deception Detection and AIGC Detection, the two hypothesis banks compared are generated using LITERATURE-ONLY and HYPOGENIC respectively.

We recruit 10 participants each task and all particpants passed attention the check question.

Criteria	Texts
Clarity	1. Highly ambiguous (The hypothesis is presented in a highly ambiguous manner, lacking clear definition and leaving significant room for interpretation or confusion.) 2. Somewhat clear but vague (The hypothesis is somewhat defined but suffers from vague terms and insufficient detail, making it challenging to grasp its meaning or how it could be tested.) 3. Moderately clear (The hypothesis is stated in a straightforward manner, but lacks the depth or specificity needed to fully convey its nuances, assumptions, or boundaries.) 4. Clear and precise (The hypothesis is clearly articulated with precise terminology and sufficient detail, providing a solid understanding of its assumptions and boundaries with minimal ambiguity.) 5. Exceptionally clear (The hypothesis is exceptionally clear, concise, and specific, with every term and aspect well-defined, leaving no room for misinterpretation and fully encapsulating its assumptions, scope, and testability.)
Novelty	1. Not novel (The hypothesis has already been shown, proven, or is widely known, closely mirroring existing ideas without introducing any new perspectives.) 2. Minimally novel (The hypothesis shows slight novelty, introducing minor variations or nuances that build upon known ideas but do not offer significant new insights.) 3. Moderately novel (The hypothesis demonstrates moderate novelty, presenting some new perspectives or angles that provide meaningful, but not groundbreaking, avenues for exploration.) 4. Notably novel (The hypothesis is notably novel, offering unique nuances or perspectives that are well-differentiated from existing ideas, representing valuable and fresh contributions to the field.) 5. Highly novel (The hypothesis is highly novel, introducing a pioneering perspective or idea that has not been previously explored, opening entirely new directions for future research.)
Plausibility	1. Not plausible (The hypothesis does not make sense at all, lacking logical or empirical grounding and failing to align with established knowledge or principles.) 2. Minimally plausible (The hypothesis has significant plausibility challenges, making sense in limited contexts but contradicting existing evidence or lacking coherence with established theories.) 3. Moderately plausible (The hypothesis makes sense overall and aligns with general principles or existing knowledge but has notable gaps or uncertainties that raise questions about its validity.) 4. Mostly plausible (The hypothesis is mostly plausible, grounded in logical reasoning and existing evidence, with only minor uncertainties or assumptions that could reasonably be addressed.) 5. Highly plausible (The hypothesis is highly plausible, fully aligning with established knowledge and logical reasoning, will likely be supported in experiments or theoretical consistency, and highly likely to be true.)

Table 7: Criteria used for human evaluation of the generated hypotheses.

C.4 IRB

We received IRB exempt (and will provide study number in the non-anonymous version of the paper). For both of the human studies, we present a detailed description of the study, incentives, risks and benefits, confidentiality, and contacts & questions in our consent form. The study proceeds only if the participant agrees to give consent.

D Additional Experiments

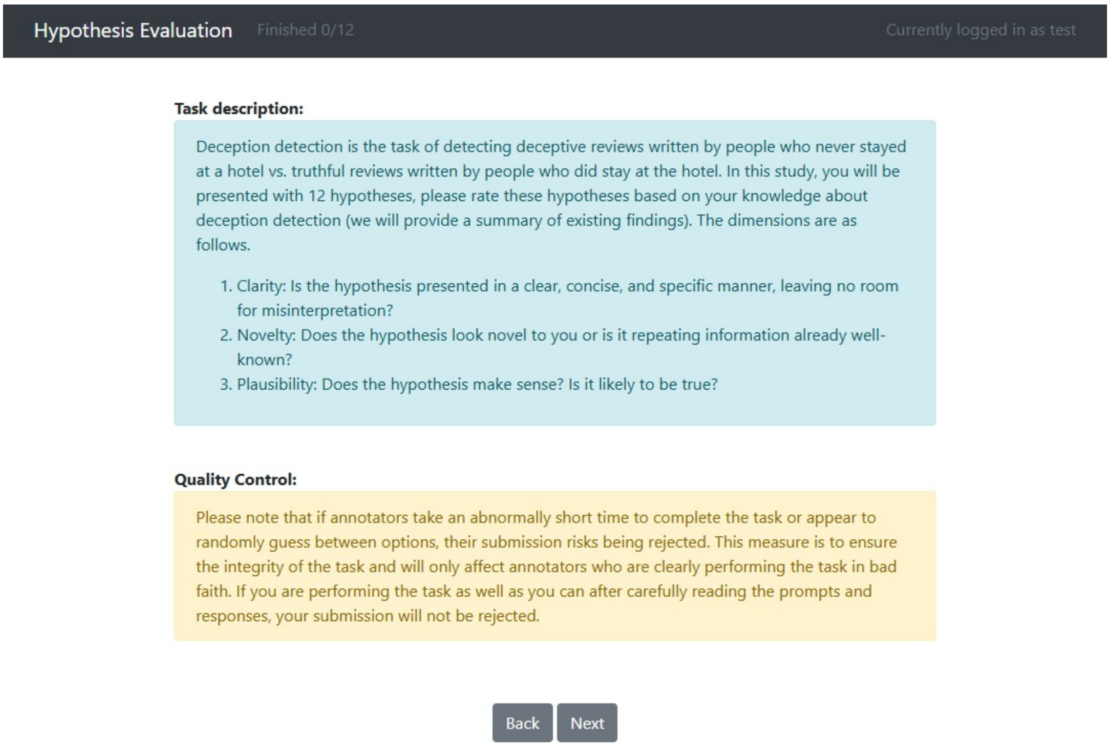
In this section, we include more analysis on the robustness of our hypothesis generation methods.

D.1 Llama-3.1-8B-Instruct Results

In Table 10, we show the performance of Llama-3.1-8B-Instruct on the OOD and IND datasets for all tasks. We show that our approach with literature + data outperforms all other methods in 8 of the 10 total configurations. Across the 10 configurations, our method outperforms few-shot inference by 15.27% on average, and it outperforms the best of literature-based methods and HYPOGENIC by 13.04% and 4.88%, respectively. This result further shows the effectiveness of our approach and provides evidence that our method can be applied with smaller models, highlighting its scalability and reproducibility.

D.2 Robustness to Prompt Variations

Since our framework heavily relies on LLMs, we perform a robustness test of our hypothesis generation method with different prompt variations. Compared with the original prompts, we consider



Task description:

Deception detection is the task of detecting deceptive reviews written by people who never stayed at a hotel vs. truthful reviews written by people who did stay at the hotel. In this study, you will be presented with 12 hypotheses, please rate these hypotheses based on your knowledge about deception detection (we will provide a summary of existing findings). The dimensions are as follows.

1. Clarity: Is the hypothesis presented in a clear, concise, and specific manner, leaving no room for misinterpretation?
2. Novelty: Does the hypothesis look novel to you or is it repeating information already well-known?
3. Plausibility: Does the hypothesis make sense? Is it likely to be true?

Quality Control:

Please note that if annotators take an abnormally short time to complete the task or appear to randomly guess between options, their submission risks being rejected. This measure is to ensure the integrity of the task and will only affect annotators who are clearly performing the task in bad faith. If you are performing the task as well as you can after carefully reading the prompts and responses, your submission will not be rejected.

Figure 2: Instruction page for likert rating.

three prompt variations: modifying the hypothesis generation prompt for $\mathcal{M}_G$, inference prompt for $\mathcal{M}_I$, and both prompts. We show performance of all model and task configurations using these prompt variations in Table 11 and Table 12 for OOD and IND settings, respectively. For the OOD datasets, the accuracy only decreases by 0.20% on average. For the 90 different configurations, 71 of them have a performance drop of less than 5%, and 48 of them get an accuracy improvement. Additionally, with the IND datasets, the average accuracy gets an increase of 0.01%. 74 out of the 90 configurations have a performance drop of less than 5%, where 50 of them get an improvement. These additional results further illustrate the robustness of our method against variations of prompts.

D.3 Hyperparameter Search

As introduced in § 2 and Appendix B.6, our hypothesis generation methods have some hyperparameters. Throughout all main experiments in § 4, we use the same set of hyperparameters, which is adopted from HypoGenic. As we show in § 4, this default choice of hyperparameters works consistently well across all different model and task configurations, highlighting the robustness of our framework. Here we conduct an additional hyperparameter search of $H_{\max}$. We show the results of using $H_{\max} = 10, 20, 30$ in Table 13 and Table 14, for the OOD and IND settings, respectively.

For the OOD datasets, changing $H_{\max} = 10, 20, 30$ results in an average accuracy decrease of only 0.07%. Out of the 60 different configurations, 51 of them get an accuracy drop of less than 5%, where 31 of them get an increase. Moreover, with the IND datasets, different choices of $H_{\max} = 10, 20, 30$ degrades average accuracy by 0.23%. In 51 out of 60 cases, we get a performance drop of less than 5%, and we get an improvement for 26 cases. These results suggest that although our default choice of hyperparameters may not be optimal for all tasks, our method is able to perform consistently well. This again highlights the robustness of our hypothesis generation framework with different hyperparameters.

E Examples of Generated Hypotheses and Qualitative Analysis

E.1 Comparing Hypotheses from SciMON and Ours

In § 5, we briefly introduce the difference between research idea generation and our hypothesis generation work. To better illustrate this difference, we in-

Model	Methods	Deceptive reviews OOD		LlamaGC OOD		GPTGC OOD		Persuasive Pairs OOD		Dreaddit OOD	
		Accuracy	F1	Accuracy	F1	Accuracy	F1	Accuracy	F1	Accuracy	F1
GPT-4 MINI	**No hypothesis**										
	Zero-shot	55.47	45.77	50.00	35.03	56.33	47.15	81.24	80.90	64.60	59.91
	Few-shot k=3	65.56	64.01	51.11	43.37	64.22	61.55	83.64	83.36	75.00	73.37
	Zero-shot generation	68.69	68.39	49.00	34.54	53.00	41.15	86.08	85.99	65.00	60.58
	Literature-based										
	literature-only	59.22	57.31	49.00	33.45	54.00	41.65	78.80	78.62	67.68	64.52
	HyperWrite	61.63	57.97	49.67	33.76	52.67	39.96	82.36	82.09	68.76	65.92
	NotebookLM	53.03	49.12	49.33	33.04	51.67	37.96	68.96	67.50	62.28	56.41
	Data-driven										
	HypoGeniC	75.22	75.14	81.67	81.61	68.56	67.62	82.20	81.71	76.56	75.71
	Literature + Data (This work)										
	HypoRefine	**77.78**	**77.71**	55.33	45.77	63.33	62.20	89.04	89.02	78.04	77.28
	Literature ∪ HypoGeniC	72.41	71.62	**83.00**	**82.96**	**69.22**	**68.39**	**89.88**	**89.87**	78.20	77.52
	Literature ∪ HypoRefine	77.19	77.17	55.33	45.77	63.00	61.81	89.52	89.51	**79.24**	**78.61**
Llama 70B-I	**No hypothesis**										
	Zero-shot	62.87	58.45	58.67	50.79	63.00	57.61	85.60	85.59	64.56	59.98
	Few-shot k=3	68.56	67.25	70.45	67.53	76.00	74.97	86.80	86.76	69.44	66.47
	Zero-shot generation	56.28	40.27	50.67	35.90	55.67	45.61	88.16	88.13	66.16	62.59
	Literature-based										
	literature-only	64.25	53.97	50.00	33.33	49.67	33.76	80.56	80.51	66.04	61.93
	HyperWrite	58.62	31.37	50.67	35.36	54.00	42.10	83.24	83.10	74.40	73.12
	NotebookLM	57.81	36.91	49.33	33.61	50.67	35.90	67.64	66.41	66.56	62.83
	Data-driven										
	HypoGeniC	62.06	56.89	78.67	78.53	78.00	77.26	88.44	88.38	75.48	74.55
	Literature + Data (This work)										
	HypoRefine	72.16	71.85	67.00	66.37	66.67	63.53	87.52	87.48	**78.92**	**78.55**
	Literature ∪ HypoGeniC	**73.72**	**73.02**	**81.33**	**81.19**	**78.67**	**78.06**	86.72	86.64	72.56	70.78
	Literature ∪ HypoRefine	71.75	71.33	66.67	65.79	65.67	62.67	**88.76**	**88.73**	74.80	73.55

Table 8: Accuracy and F1 scores on the held-out OOD datasets. Literature + data outperforms all other methods in every model and task configurations. The bolded numbers outperform the few-shot method with statistical significance, as determined by a paired t-test using five random seeds.

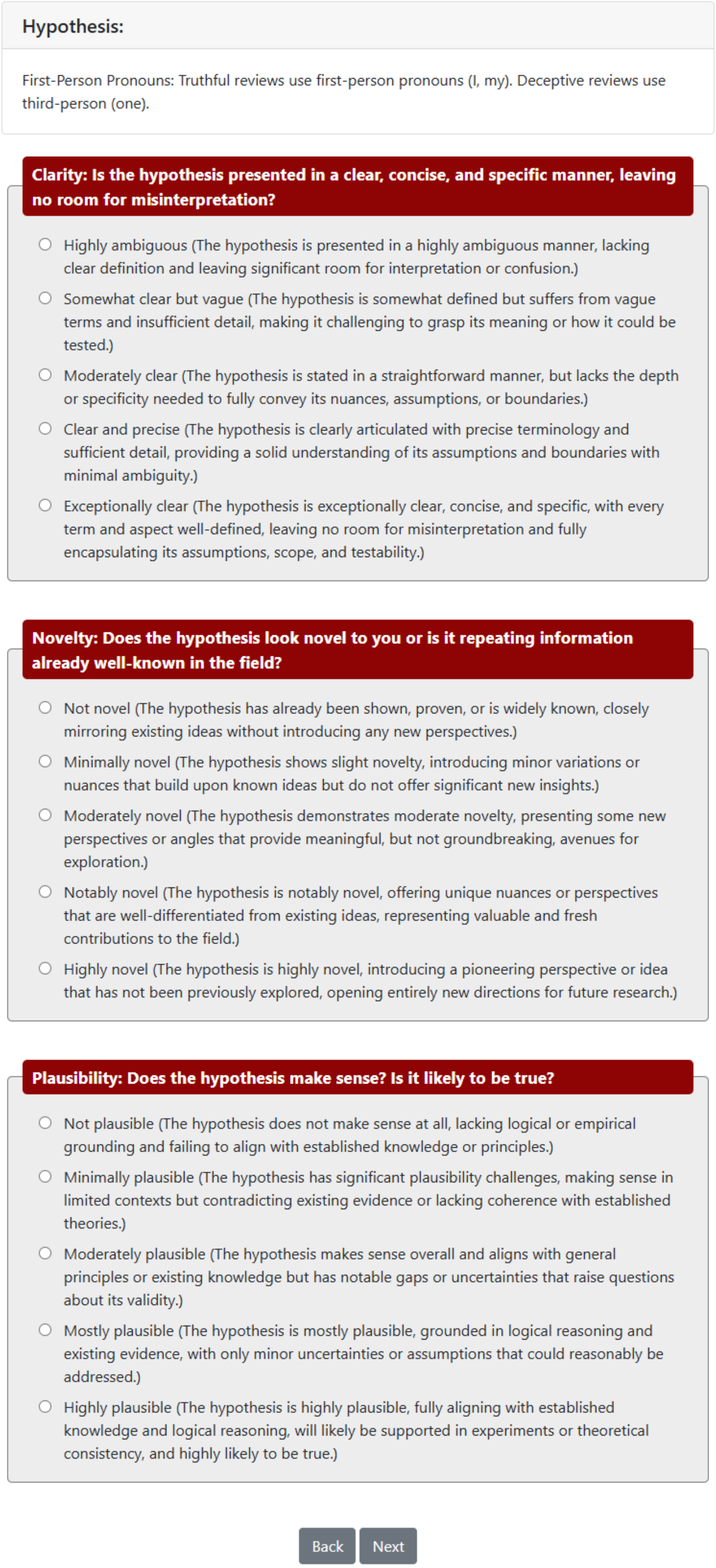
clude a detailed comparison in Table 15, consisting the generated research idea from SciMON (Wang et al., 2024), generated hypothesis with HypoRefine, and an existing finding from Li et al. (2014) on Deception Detection. For the SciMON generated idea, we adopted from Table 11 in Wang et al. (2024). These examples show that SciMON aims to generate ideas for a potential research project, where our method focus on generating possible explanations of a phenomenon. In addition, comparing with the existing finding from Li et al. (2014), our generated hypothesis is highly relevant to the field of interest, i.e., Deception Detection.

E.2 Example Hypotheses

We include examples of generated hypotheses using our Literature∪HypoRefine approach and GPT-4-mini, together with a brief qualitative analysis of its source in Table 16. We also showcase example hypotheses generated using NotebookLM and HyperWrite on Deceptive reviews that are invalid or irrelevant in Table 17. These hypotheses can lead to degraded inference performance for theses two methods.

Model	Methods	Deceptive reviews IND		LlamaGC IND		GPTGC IND		Persuasive Pairs IND		Dreaddit IND	
		Accuracy	F1	Accuracy	F1	Accuracy	F1	Accuracy	F1	Accuracy	F1
GPT-4 MINI	**No hypothesis**										
	Zero-shot	56.56	51.66	56.33	47.15	50.00	35.03	83.72	83.59	62.32	56.24
	Few-shot k=3	62.60	61.40	63.67	61.54	54.78	49.50	85.24	85.14	67.48	63.59
	Zero-shot generation	60.16	60.15	54.33	44.36	49.67	33.18	87.72	87.71	62.24	56.11
	Literature-based										
	literature-only	65.60	64.53	52.00	39.58	50.00	33.33	78.80	78.80	62.76	56.91
	HyperWrite	58.88	54.29	52.67	39.96	49.67	33.76	86.16	86.13	64.96	60.18
	NotebookLM	50.40	48.79	51.67	37.96	49.33	33.04	75.92	75.78	60.52	53.37
	Data-driven										
	HypoGeniC	68.32	67.60	72.33	71.70	66.33	64.88	84.68	84.52	70.40	67.90
	Literature + Data (This work)										
	HypoRefine	68.56	68.43	62.33	57.08	57.33	54.41	89.76	89.75	70.76	68.31
	Literature ∪ HypoGeniC	68.76	67.66	**73.00**	**72.49**	**68.33**	**67.54**	89.84	89.84	**70.88**	**68.46**
	Literature ∪ HypoRefine	**70.76**	**70.68**	60.33	54.30	55.00	51.65	**90.52**	**90.52**	69.88	67.21
Llama 70B-I	**No hypothesis**										
	Zero-shot	58.32	50.19	63.00	57.61	58.67	50.79	86.28	86.26	64.80	60.08
	Few-shot k=3	62.92	58.48	76.89	75.97	73.00	70.89	87.80	87.79	68.28	64.81
	Zero-shot generation	53.68	41.02	55.67	45.61	50.67	35.90	88.80	88.80	70.52	68.39
	Literature-based										
	literature-only	62.96	58.39	51.33	36.23	49.67	22.61	80.32	80.32	66.08	62.09
	HyperWrite	52.28	26.75	54.00	42.10	50.67	35.36	87.48	87.46	71.12	69.24
	NotebookLM	57.12	35.31	50.67	35.90	49.33	33.61	71.16	70.91	68.72	65.59
	Data-driven										
	HypoGeniC	64.44	61.24	**78.33**	**78.21**	**80.67**	**79.97**	**91.24**	**91.22**	74.68	73.48
	Literature + Data (This work)										
	HypoRefine	71.68	**71.42**	71.00	70.14	75.33	74.38	88.92	88.89	**78.68**	**78.22**
	Literature ∪ HypoGeniC	70.28	68.76	**78.33**	78.13	80.00	79.22	89.36	89.33	70.68	68.28
	Literature ∪ HypoRefine	**72.60**	67.80	71.00	70.14	74.00	72.84	90.88	90.88	73.20	71.60

Table 9: Accuracy and F1 scores on the held-out IND datasets. Literature + data outperforms all other methods in 7 out of 10 configurations. For Llama-70B-I on GPTGC, LlamaGC, and Persuasive Pairs, HypoGeniC performs the best. This is likely due to that the literature in these tasks do not offer helpful information for the IND data, but they can still provide useful information for the tasks in general. As in Table 8, our approaches with literature + data performs the best in all configurations for the OOD datasets.



Hypothesis:

First-Person Pronouns: Truthful reviews use first-person pronouns (I, my). Deceptive reviews use third-person (one).

Clarity: Is the hypothesis presented in a clear, concise, and specific manner, leaving no room for misinterpretation?

○ Highly ambiguous (The hypothesis is presented in a highly ambiguous manner, lacking clear definition and leaving significant room for interpretation or confusion.)

○ Somewhat clear but vague (The hypothesis is somewhat defined but suffers from vague terms and insufficient detail, making it challenging to grasp its meaning or how it could be tested.)

○ Moderately clear (The hypothesis is stated in a straightforward manner, but lacks the depth or specificity needed to fully convey its nuances, assumptions, or boundaries.)

○ Clear and precise (The hypothesis is clearly articulated with precise terminology and sufficient detail, providing a solid understanding of its assumptions and boundaries with minimal ambiguity.)

○ Exceptionally clear (The hypothesis is exceptionally clear, concise, and specific, with every term and aspect well-defined, leaving no room for misinterpretation and fully encapsulating its assumptions, scope, and testability.)

Novelty: Does the hypothesis look novel to you or is it repeating information already well-known in the field?

○ Not novel (The hypothesis has already been shown, proven, or is widely known, closely mirroring existing ideas without introducing any new perspectives.)

○ Minimally novel (The hypothesis shows slight novelty, introducing minor variations or nuances that build upon known ideas but do not offer significant new insights.)

○ Moderately novel (The hypothesis demonstrates moderate novelty, presenting some new perspectives or angles that provide meaningful, but not groundbreaking, avenues for exploration.)

○ Notably novel (The hypothesis is notably novel, offering unique nuances or perspectives that are well-differentiated from existing ideas, representing valuable and fresh contributions to the field.)

○ Highly novel (The hypothesis is highly novel, introducing a pioneering perspective or idea that has not been previously explored, opening entirely new directions for future research.)

Plausibility: Does the hypothesis make sense? Is it likely to be true?

○ Not plausible (The hypothesis does not make sense at all, lacking logical or empirical grounding and failing to align with established knowledge or principles.)

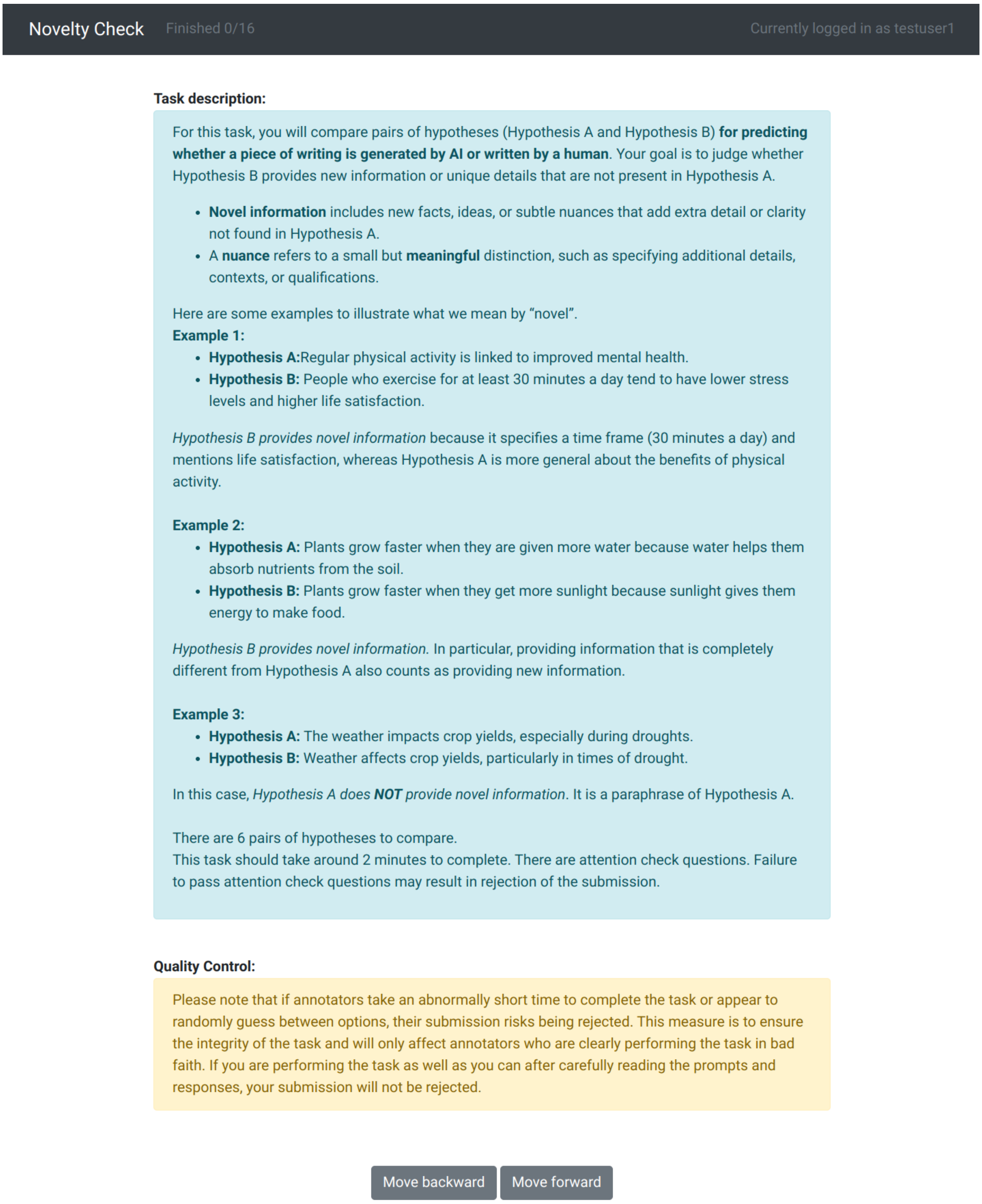
○ Minimally plausible (The hypothesis has significant plausibility challenges, making sense in limited contexts but contradicting existing evidence or lacking coherence with established theories.)

○ Moderately plausible (The hypothesis makes sense overall and aligns with general principles or existing knowledge but has notable gaps or uncertainties that raise questions about its validity.)

○ Mostly plausible (The hypothesis is mostly plausible, grounded in logical reasoning and existing evidence, with only minor uncertainties or assumptions that could reasonably be addressed.)

○ Highly plausible (The hypothesis is highly plausible, fully aligning with established knowledge and logical reasoning, will likely be supported in experiments or theoretical consistency, and highly likely to be true.)

Figure 3: Annotation page for likert rating.



Task description:

For this task, you will compare pairs of hypotheses (Hypothesis A and Hypothesis B) **for predicting whether a piece of writing is generated by AI or written by a human**. Your goal is to judge whether Hypothesis B provides new information or unique details that are not present in Hypothesis A.

- **Novel information** includes new facts, ideas, or subtle nuances that add extra detail or clarity not found in Hypothesis A.
- A **nuance** refers to a small but **meaningful** distinction, such as specifying additional details, contexts, or qualifications.

Here are some examples to illustrate what we mean by “novel”.
Example 1:

- **Hypothesis A:**Regular physical activity is linked to improved mental health.
- **Hypothesis B:** People who exercise for at least 30 minutes a day tend to have lower stress levels and higher life satisfaction.

Hypothesis B provides novel information because it specifies a time frame (30 minutes a day) and mentions life satisfaction, whereas Hypothesis A is more general about the benefits of physical activity.

Example 2:

- **Hypothesis A:** Plants grow faster when they are given more water because water helps them absorb nutrients from the soil.
- **Hypothesis B:** Plants grow faster when they get more sunlight because sunlight gives them energy to make food.

Hypothesis B provides novel information. In particular, providing information that is completely different from Hypothesis A also counts as providing new information.

Example 3:

- **Hypothesis A:** The weather impacts crop yields, especially during droughts.
- **Hypothesis B:** Weather affects crop yields, particularly in times of drought.

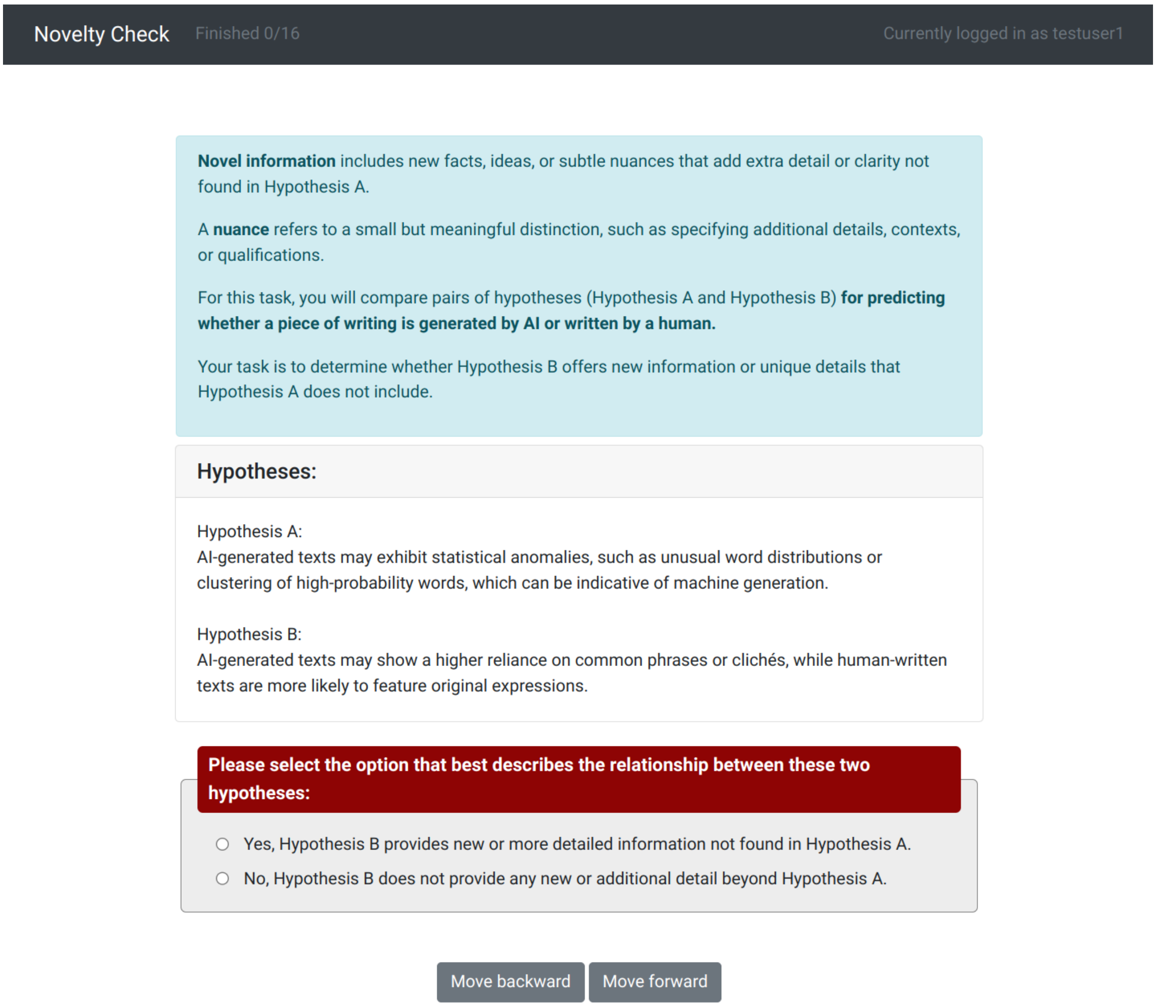
In this case, *Hypothesis A does **NOT** provide novel information*. It is a paraphrase of Hypothesis A.

There are 6 pairs of hypotheses to compare.
This task should take around 2 minutes to complete. There are attention check questions. Failure to pass attention check questions may result in rejection of the submission.

Quality Control:

Please note that if annotators take an abnormally short time to complete the task or appear to randomly guess between options, their submission risks being rejected. This measure is to ensure the integrity of the task and will only affect annotators who are clearly performing the task in bad faith. If you are performing the task as well as you can after carefully reading the prompts and responses, your submission will not be rejected.

Figure 4: Instruction page for novelty check.



Novel information includes new facts, ideas, or subtle nuances that add extra detail or clarity not found in Hypothesis A.

A **nuance** refers to a small but meaningful distinction, such as specifying additional details, contexts, or qualifications.

For this task, you will compare pairs of hypotheses (Hypothesis A and Hypothesis B) **for predicting whether a piece of writing is generated by AI or written by a human.**

Your task is to determine whether Hypothesis B offers new information or unique details that Hypothesis A does not include.

Hypotheses:

Hypothesis A:
AI-generated texts may exhibit statistical anomalies, such as unusual word distributions or clustering of high-probability words, which can be indicative of machine generation.

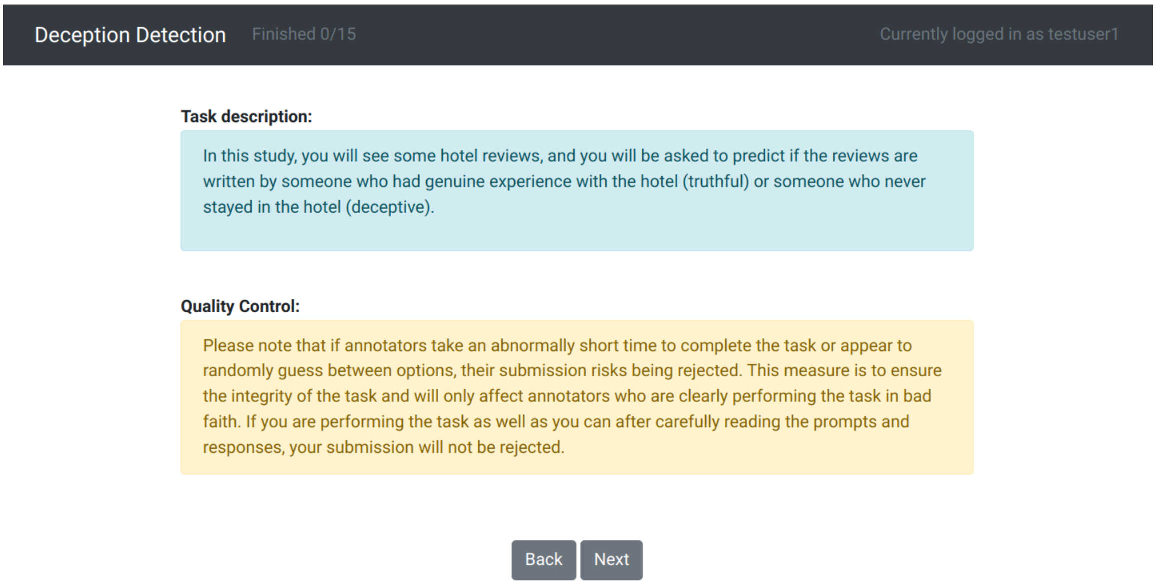
Hypothesis B:
AI-generated texts may show a higher reliance on common phrases or clichés, while human-written texts are more likely to feature original expressions.

Please select the option that best describes the relationship between these two hypotheses:

○ Yes, Hypothesis B provides new or more detailed information not found in Hypothesis A.

○ No, Hypothesis B does not provide any new or additional detail beyond Hypothesis A.

Figure 5: Annotation page for novelty check.



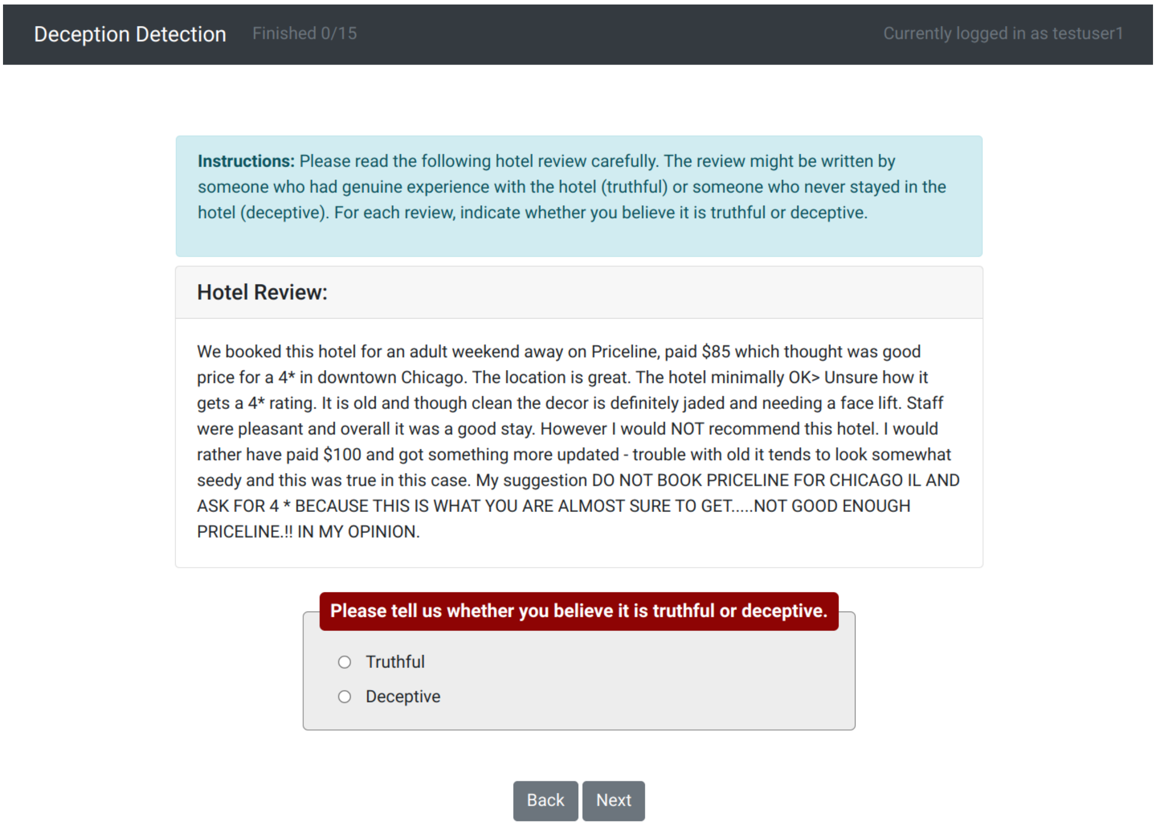
Task description:

In this study, you will see some hotel reviews, and you will be asked to predict if the reviews are written by someone who had genuine experience with the hotel (truthful) or someone who never stayed in the hotel (deceptive).

Quality Control:

Please note that if annotators take an abnormally short time to complete the task or appear to randomly guess between options, their submission risks being rejected. This measure is to ensure the integrity of the task and will only affect annotators who are clearly performing the task in bad faith. If you are performing the task as well as you can after carefully reading the prompts and responses, your submission will not be rejected.

Figure 6: Instruction page for prediction task without hypotheses.



Instructions: Please read the following hotel review carefully. The review might be written by someone who had genuine experience with the hotel (truthful) or someone who never stayed in the hotel (deceptive). For each review, indicate whether you believe it is truthful or deceptive.

Hotel Review:

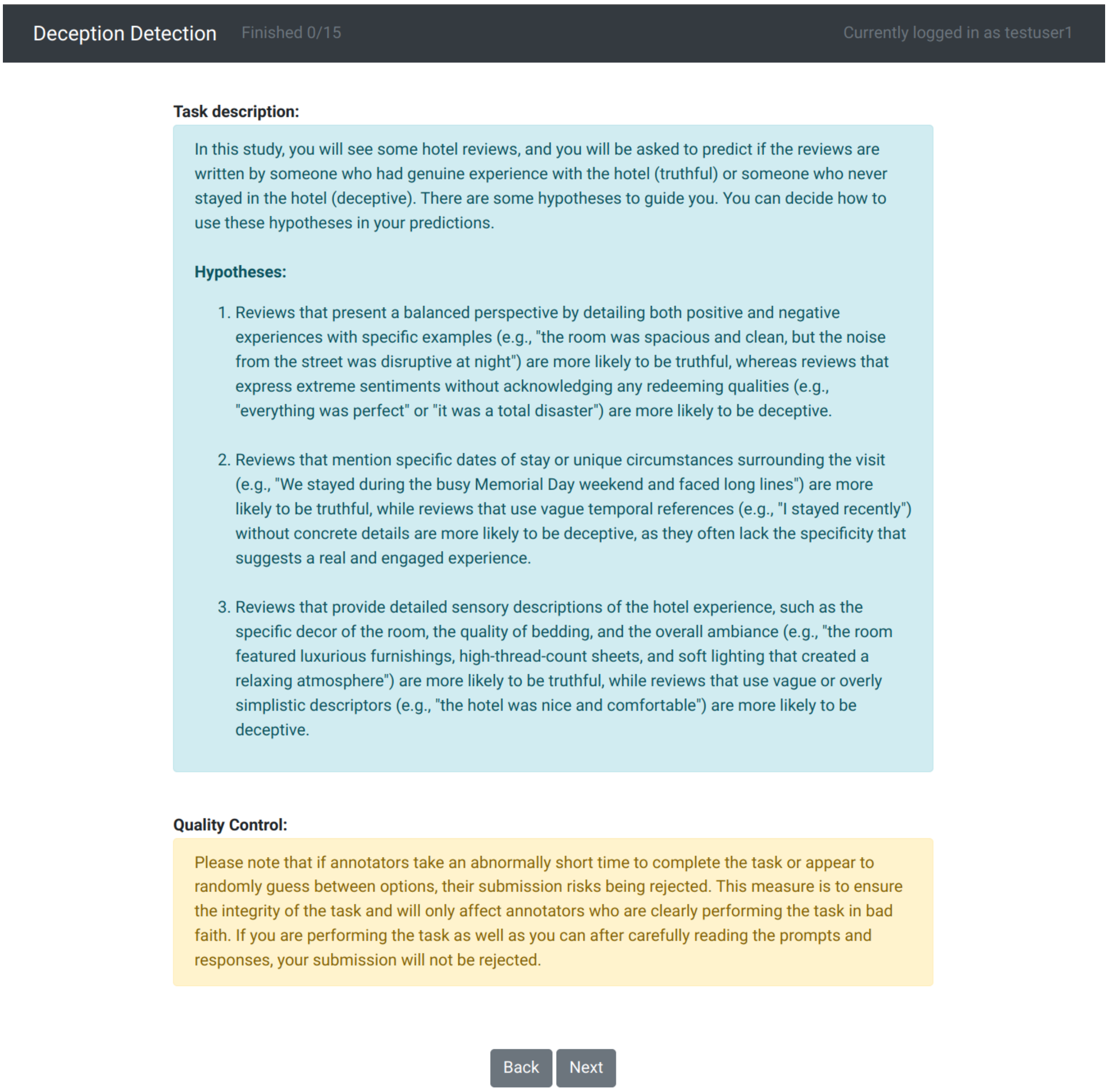
We booked this hotel for an adult weekend away on Priceline, paid $85 which thought was good price for a 4* in downtown Chicago. The location is great. The hotel minimally OK> Unsure how it gets a 4* rating. It is old and though clean the decor is definitely jaded and needing a face lift. Staff were pleasant and overall it was a good stay. However I would NOT recommend this hotel. I would rather have paid $100 and got something more updated - trouble with old it tends to look somewhat seedy and this was true in this case. My suggestion DO NOT BOOK PRICELINE FOR CHICAGO IL AND ASK FOR 4 * BECAUSE THIS IS WHAT YOU ARE ALMOST SURE TO GET.....NOT GOOD ENOUGH PRICELINE.!! IN MY OPINION.

Please tell us whether you believe it is truthful or deceptive.

○ Truthful

○ Deceptive

Figure 7: Annotation page for prediction task without hypotheses.



Task description:

In this study, you will see some hotel reviews, and you will be asked to predict if the reviews are written by someone who had genuine experience with the hotel (truthful) or someone who never stayed in the hotel (deceptive). There are some hypotheses to guide you. You can decide how to use these hypotheses in your predictions.

Hypotheses:

1. Reviews that present a balanced perspective by detailing both positive and negative experiences with specific examples (e.g., "the room was spacious and clean, but the noise from the street was disruptive at night") are more likely to be truthful, whereas reviews that express extreme sentiments without acknowledging any redeeming qualities (e.g., "everything was perfect" or "it was a total disaster") are more likely to be deceptive.

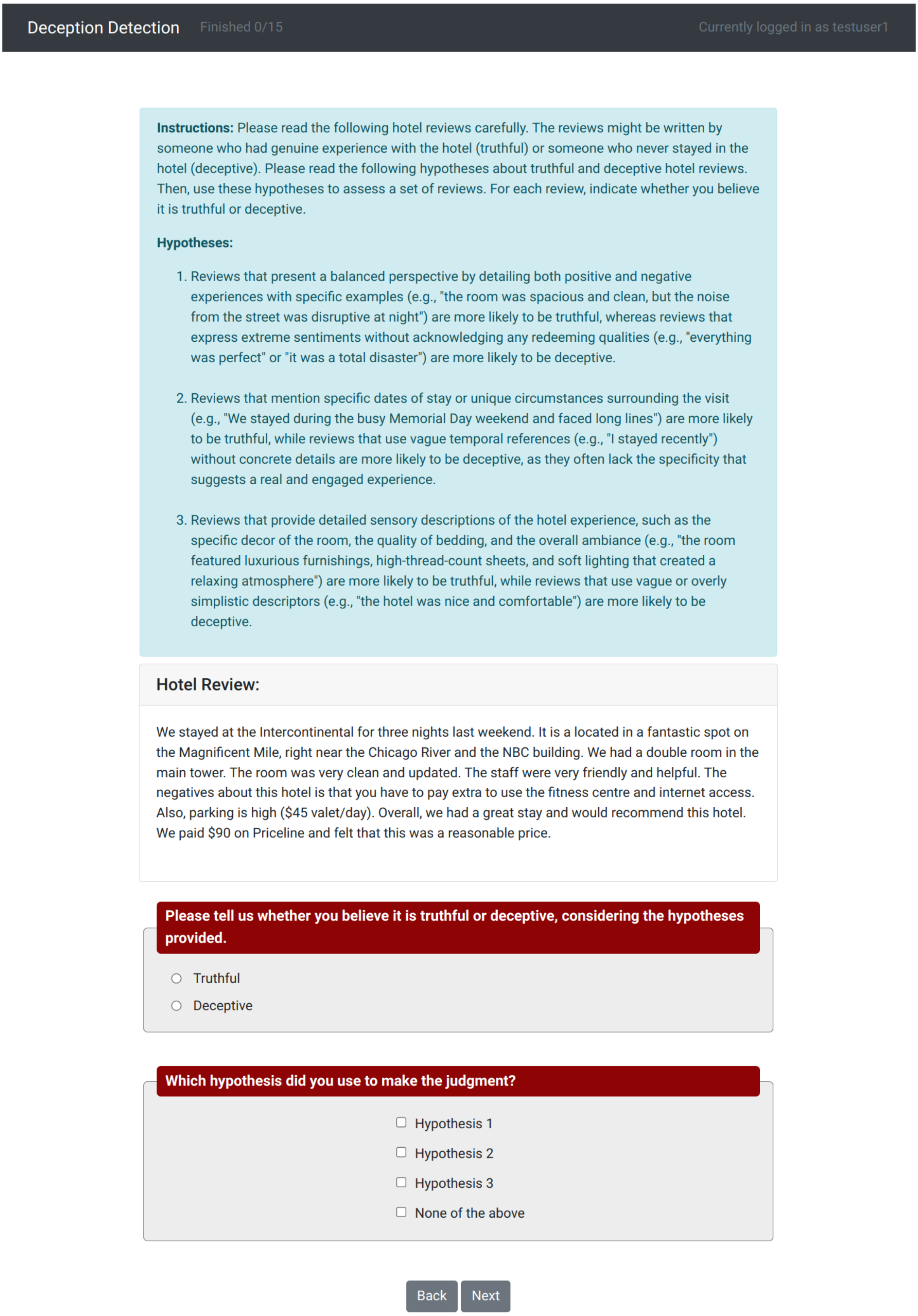
2. Reviews that mention specific dates of stay or unique circumstances surrounding the visit (e.g., "We stayed during the busy Memorial Day weekend and faced long lines") are more likely to be truthful, while reviews that use vague temporal references (e.g., "I stayed recently") without concrete details are more likely to be deceptive, as they often lack the specificity that suggests a real and engaged experience.

3. Reviews that provide detailed sensory descriptions of the hotel experience, such as the specific decor of the room, the quality of bedding, and the overall ambiance (e.g., "the room featured luxurious furnishings, high-thread-count sheets, and soft lighting that created a relaxing atmosphere") are more likely to be truthful, while reviews that use vague or overly simplistic descriptors (e.g., "the hotel was nice and comfortable") are more likely to be deceptive.

Quality Control:

Please note that if annotators take an abnormally short time to complete the task or appear to randomly guess between options, their submission risks being rejected. This measure is to ensure the integrity of the task and will only affect annotators who are clearly performing the task in bad faith. If you are performing the task as well as you can after carefully reading the prompts and responses, your submission will not be rejected.

Figure 8: Instruction page for prediction task with the guide of hypotheses.



Instructions: Please read the following hotel reviews carefully. The reviews might be written by someone who had genuine experience with the hotel (truthful) or someone who never stayed in the hotel (deceptive). Please read the following hypotheses about truthful and deceptive hotel reviews. Then, use these hypotheses to assess a set of reviews. For each review, indicate whether you believe it is truthful or deceptive.

Hypotheses:

1. Reviews that present a balanced perspective by detailing both positive and negative experiences with specific examples (e.g., "the room was spacious and clean, but the noise from the street was disruptive at night") are more likely to be truthful, whereas reviews that express extreme sentiments without acknowledging any redeeming qualities (e.g., "everything was perfect" or "it was a total disaster") are more likely to be deceptive.

2. Reviews that mention specific dates of stay or unique circumstances surrounding the visit (e.g., "We stayed during the busy Memorial Day weekend and faced long lines") are more likely to be truthful, while reviews that use vague temporal references (e.g., "I stayed recently") without concrete details are more likely to be deceptive, as they often lack the specificity that suggests a real and engaged experience.

3. Reviews that provide detailed sensory descriptions of the hotel experience, such as the specific decor of the room, the quality of bedding, and the overall ambiance (e.g., "the room featured luxurious furnishings, high-thread-count sheets, and soft lighting that created a relaxing atmosphere") are more likely to be truthful, while reviews that use vague or overly simplistic descriptors (e.g., "the hotel was nice and comfortable") are more likely to be deceptive.

Hotel Review:

We stayed at the Intercontinental for three nights last weekend. It is a located in a fantastic spot on the Magnificent Mile, right near the Chicago River and the NBC building. We had a double room in the main tower. The room was very clean and updated. The staff were very friendly and helpful. The negatives about this hotel is that you have to pay extra to use the fitness centre and internet access. Also, parking is high ($45 valet/day). Overall, we had a great stay and would recommend this hotel. We paid $90 on Priceline and felt that this was a reasonable price.

Please tell us whether you believe it is truthful or deceptive, considering the hypotheses provided.

○ Truthful

○ Deceptive

Which hypothesis did you use to make the judgment?

□ Hypothesis 1

□ Hypothesis 2

□ Hypothesis 3

□ None of the above

Figure 9: Annotation page for prediction task with the guide of hypotheses.

	Methods	DECEPTIVE REVIEWS		LLAMAGC		GPTGC		PERSUASIVE PAIRS		DREADDIT	
		Accuracy	F1	Accuracy	F1	Accuracy	F1	Accuracy	F1	Accuracy	F1
OOD	**No hypothesis**										
	Zero-shot	48.25	41.37	48.00	32.43	52.00	40.48	63.96	47.15	62.80	57.40
	Few-shot k=3	54.00	53.09	46.20	33.78	53.20	44.35	77.96	77.79	63.32	57.81
	Zero-shot generation	25.25	20.97	48.53	32.68	59.00	53.76	76.88	76.84	60.56	54.68
	Literature-based										
	LITERATURE-ONLY	55.22	55.19	47.27	32.10	54.67	48.74	84.08	84.08	62.64	57.32
	HYPERWRITE	55.28	48.30	47.33	33.30	47.00	36.29	81.00	80.99	61.88	56.19
	NOTEBOOKLM	52.09	47.01	49.00	32.89	50.67	36.42	58.84	58.83	62.20	56.89
	Data-driven										
	HYPOGENIC	**64.72**	**64.00**	63.53	60.48	70.33	68.97	78.96	78.89	68.64	66.18
	Literature + Data (This work)										
	HYPOREFINE	56.91	51.33	**65.80**	**64.92**	**81.33**	**81.14**	**87.44**	**87.44**	68.20	67.85
	Literature ∪ HYPOGENIC	54.34	44.51	63.93	61.00	70.60	69.21	83.84	83.76	**71.52**	**69.75**
	Literature ∪ HYPOREFINE	61.91	61.34	65.73	64.68	81.07	80.85	85.96	85.93	68.80	67.28
IND	**No hypothesis**										
	Zero-shot	50.88	41.34	52.26	40.98	48.00	32.43	70.92	49.05	63.76	58.39
	Few-shot k=3	58.20	54.86	51.80	43.89	47.33	33.49	71.84	71.55	63.20	57.81
	Zero-shot generation	28.64	22.60	49.20	35.16	50.27	39.36	78.00	77.79	61.60	55.51
	Literature-based										
	LITERATURE-ONLY	56.72	56.48	50.00	36.68	52.20	44.32	80.88	80.79	60.32	53.78
	HYPERWRITE	53.12	42.51	48.66	36.80	47.00	33.03	79.48	79.28	61.20	55.04
	NOTEBOOKLM	50.84	44.57	50.47	36.52	48.67	32.74	63.72	63.47	63.40	58.47
	Data-driven										
	HYPOGENIC	57.60	54.29	**71.80**	**70.61**	70.07	68.66	79.16	79.06	66.12	62.53
	Literature + Data (This work)										
	HYPOREFINE	54.88	47.44	63.73	62.87	82.27	82.01	85.44	85.44	**69.92**	**69.06**
	Literature ∪ HYPOGENIC	52.84	41.59	71.33	70.15	70.33	68.98	82.04	81.88	67.48	64.76
	Literature ∪ HYPOREFINE	**61.40**	**58.89**	64.13	63.30	**82.33**	**82.09**	**86.32**	**86.32**	67.56	65.35

Table 10: Accuracy and F1 scores of Llama-3.1-8B-Instruct on the OOD and IND datasets. Literature + data outperforms all other methods in 4 out of 5 configurations for both OOD and IND datasets. This further validates the effectiveness of our methods even on smaller models.

Model	Methods	Deceptive reviews	LlamaGC	GPTGC	Persuasive Pairs	Dreaddit
GPT-4 MINI	**HypoRefine**					
	Original prompt	77.78	55.33	63.33	89.04	78.04
	Prompt Variation 1	73.28 (↓4.50)	49.33 (↓6.00)	83.33 (↑20.00)	91.80 (↑2.76)	83.40 (↑5.36)
	Prompt Variation 2	69.69 (↓8.09)	66.67 (↑11.34)	69.33 (↑6.00)	89.40 (↑0.36)	75.60 (↓2.44)
	Prompt Variation 3	74.06 (↓3.72)	49.00 (↓6.33)	56.00 (↓7.33)	91.20 (↑2.16)	71.20 (↓6.84)
	Literature ∪ HypoGeniC					
	Original prompt	72.41	83.00	69.22	89.88	78.20
	Prompt Variation 1	74.53 (↑2.12)	80.33 (↓2.67)	61.33 (↓7.89)	90.20 (↑0.32)	74.20 (↓4.00)
	Prompt Variation 2	68.91 (↓3.50)	49.33 (↓33.67)	70.00 (↑0.78)	91.60 (↑1.78)	74.60 (↓3.60)
	Prompt Variation 3	73.75 (↑1.34)	49.00 (↓34.00)	69.00 (↓0.22)	88.60 (↓1.28)	71.60 (↓6.60)
	Literature ∪ HypoRefine					
	Original prompt	77.19	55.33	63.00	89.52	79.24
	Prompt Variation 1	69.69 (↓7.50)	49.67 (↓5.66)	82.33 (↑19.33)	90.40 (↑0.82)	81.40 (↑2.16)
	Prompt Variation 2	72.03 (↓5.16)	70.67 (↑15.34)	63.67 (↑0.67)	90.20 (↑0.68)	78.60 (↓0.64)
	Prompt Variation 3	69.38 (↓7.81)	49.00 (↓6.33)	57.33 (↓5.67)	90.00 (↑0.48)	78.20 (↓1.04)
Llama 70B-I	**HypoRefine**					
	Original prompt	72.16	67.00	66.67	87.52	78.92
	Prompt Variation 1	63.13 (↓9.03)	66.00 (↓1.00)	68.33 (↑1.66)	90.60 (↑3.08)	81.60 (↑2.68)
	Prompt Variation 2	70.47 (↓1.69)	81.00 (↑14.00)	74.67 (↑8.00)	84.20 (↓3.32)	70.80 (↓8.12)
	Prompt Variation 3	68.13 (↓4.03)	74.00 (↑7.00)	74.00 (↑7.33)	89.40 (↑1.88)	81.40 (↑2.48)
	Literature ∪ HypoGeniC					
	Original prompt	73.72	81.33	78.67	86.72	72.56
	Prompt Variation 1	74.38 (↑0.66)	70.67 (↓10.66)	80.33 (↑1.66)	90.00 (↑3.28)	75.60 (↑3.04)
	Prompt Variation 2	72.97 (↓0.75)	78.33 (↓3.00)	85.33 (↑6.66)	85.00 (↓1.72)	74.40 (↑1.84)
	Prompt Variation 3	75.63 (↑1.91)	80.00 (↓1.33)	83.67 (↑5.00)	88.40 (↑1.68)	75.00 (↑2.44)
	Literature ∪ HypoRefine					
	Original prompt	71.75	66.67	65.67	88.76	74.80
	Prompt Variation 1	66.72 (↓5.03)	66.00 (↓0.67)	68.33 (↑2.66)	89.00 (↑0.24)	75.40 (↑0.60)
	Prompt Variation 2	77.97 (↑6.22)	81.00 (↑14.33)	74.67 (↑9.00)	88.80 (↑0.04)	72.80 (↓2.00)
	Prompt Variation 3	69.53 (↓2.22)	74.00 (↑7.33)	74.00 (↑8.33)	89.60 (↑0.84)	73.80 (↓1.00)

Table 11: Accuracy numbers on OOD datasets with 4 different sets of prompts. The prompts used for the results in Table 1, Table 8, and Table 9 are indicated with "original prompt". The prompt variations contain different paraphrases of the original prompts for hypothesis generation and hypothesis-based inference. Results show the robustness of our methods to different prompts.

Model	Methods	Deceptive reviews	LlamaGC	GPTGC	Persuasive Pairs	Dreaddit
GPT-4 mini	**HypoRefine**					
	Original prompt	68.56	62.33	57.33	89.76	70.76
	Prompt Variation 1	65.40 (↓3.16)	52.00 (↓10.33)	84.67 (↑27.34)	89.80 (↑0.04)	73.20 (↑2.44)
	Prompt Variation 2	67.00 (↓1.56)	68.33 (↑6.00)	59.67 (↑2.34)	88.80 (↓0.96)	69.80 (↓0.96)
	Prompt Variation 3	68.60 (↑0.04)	53.33 (↓9.00)	53.00 (↓4.33)	88.60 (↓1.16)	69.40 (↓1.36)
	Literature ∪ HypoGenic					
	Original prompt	68.76	73.00	68.33	89.84	70.88
	Prompt Variation 1	69.20 (↑0.44)	75.00 (↑2.00)	54.67 (↓13.66)	88.40 (↓1.44)	65.80 (↓5.08)
	Prompt Variation 2	67.60 (↓1.16)	57.00 (↓16.00)	60.67 (↓7.66)	91.20 (↑1.36)	69.00 (↓1.88)
	Prompt Variation 3	70.20 (↑1.44)	54.67 (↓18.33)	59.33 (↓9.00)	87.20 (↓2.64)	66.80 (↓4.08)
	Literature ∪ HypoRefine					
	Original prompt	70.76	60.33	55.00	90.52	69.88
	Prompt Variation 1	64.00 (↓6.76)	51.67 (↓8.66)	85.67 (↑30.67)	90.20 (↓0.32)	71.20 (↑1.32)
	Prompt Variation 2	67.00 (↓3.76)	66.33 (↑6.00)	54.67 (↓0.33)	90.80 (↑0.28)	70.80 (↑0.92)
	Prompt Variation 3	65.00 (↓5.76)	54.33 (↓6.00)	52.33 (↓2.67)	89.80 (↓0.72)	70.80 (↑0.92)
Llama 70B-I	**HypoRefine**					
	Original prompt	71.68	71.00	75.33	88.92	78.68
	Prompt Variation 1	61.80 (↓9.88)	73.00 (↑2.00)	80.00 (↑4.67)	91.20 (↑2.28)	80.80 (↑2.12)
	Prompt Variation 2	70.80 (↓0.88)	77.00 (↑6.00)	79.00 (↑3.67)	89.80 (↑0.88)	74.60 (↓4.08)
	Prompt Variation 3	70.00 (↓1.68)	79.00 (↑8.00)	70.33 (↓5.00)	92.00 (↑3.08)	80.00 (↑1.32)
	Literature ∪ HypoGenic					
	Original prompt	70.28	78.33	80.00	89.36	70.68
	Prompt Variation 1	69.80 (↓0.48)	70.33 (↓8.00)	83.00 (↑3.00)	91.80 (↑2.44)	76.20 (↑5.52)
	Prompt Variation 2	72.20 (↑1.92)	79.33 (↑1.00)	92.33 (↑12.33)	89.60 (↑0.24)	74.20 (↑3.52)
	Prompt Variation 3	73.20 (↑2.92)	81.00 (↑2.67)	82.00 (↑2.00)	92.40 (↑3.04)	75.20 (↑4.52)
	Literature ∪ HypoRefine					
	Original prompt	72.60	71.00	74.00	90.88	73.20
	Prompt Variation 1	64.80 (↓7.80)	73.00 (↑2.00)	80.00 (↑6.00)	89.80 (↓1.08)	76.20 (↑3.00)
	Prompt Variation 2	75.20 (↑2.60)	77.00 (↑6.00)	74.67 (↑0.67)	91.40 (↑0.52)	74.00 (↑0.80)
	Prompt Variation 3	67.20 (↓5.40)	79.00 (↑8.00)	70.33 (↓3.67)	92.40 (↑1.52)	77.60 (↑3.40)

Table 12: Accuracy numbers on IND datasets with 4 different sets of prompts.

Model	Methods	Deceptive reviews	LlamaGC	GPTGC	Persuasive Pairs	Dreaddit
GPT-4 mini	**HypoRefine**					
	$H_{\max} = 10$	78.75 (↑0.97)	52.00 (↓3.33)	48.67 (↓14.66)	88.80 (↓0.24)	78.20 (↑0.16)
	$H_{\max} = 20$	77.78	55.33	63.33	89.04	78.04
	$H_{\max} = 30$	79.69 (↑1.91)	48.67 (↓6.66)	66.67 (↑3.34)	90.40 (↑1.36)	76.40 (↓1.64)
	Literature ∪ HypoGenic					
	$H_{\max} = 10$	73.00 (↓0.59)	68.00 (↓15.00)	60.33 (↓8.89)	90.60 (↑0.72)	79.80 (↑1.60)
	$H_{\max} = 20$	72.41	83.00	69.22	89.88	78.20
	$H_{\max} = 30$	74.60 (↑2.19)	87.67 (↑4.67)	82.33 (↑13.11)	90.80 (↑0.92)	75.40 (↓2.80)
	Literature ∪ HypoRefine					
	$H_{\max} = 10$	73.80 (↓3.39)	51.33 (↓4.00)	49.00 (↓14.00)	89.40 (↓0.12)	75.80 (↓3.44)
	$H_{\max} = 20$	77.19	55.33	63.00	89.52	79.24
	$H_{\max} = 30$	76.40 (↓0.79)	49.33 (↓6.00)	67.67 (↑4.67)	90.80 (↑1.28)	74.20 (↓5.04)
Llama 70B-I	**HypoRefine**					
	$H_{\max} = 10$	73.59 (↑1.43)	71.33 (↑4.33)	77.67 (↑10.00)	84.00 (↓3.52)	79.00 (↑0.08)
	$H_{\max} = 20$	72.16	67.00	66.67	87.52	78.92
	$H_{\max} = 30$	71.09 (↓1.07)	72.33 (↑5.33)	78.33 (↑11.66)	90.00 (↑2.48)	72.80 (↓6.12)
	Literature ∪ HypoGenic					
	$H_{\max} = 10$	70.20 (↓3.52)	78.33 (↓3.00)	81.00 (↑2.33)	86.80 (↑0.08)	69.60 (↓2.96)
	$H_{\max} = 20$	73.72	81.33	78.67	86.72	72.56
	$H_{\max} = 30$	66.00 (↓7.72)	86.67 (↑5.34)	81.00 (↑2.33)	89.20 (↑2.48)	75.80 (↑3.24)
	Literature ∪ HypoRefine					
	$H_{\max} = 10$	69.20 (↓2.55)	71.33 (↑4.66)	77.67 (↑12.00)	87.20 (↓1.56)	77.80 (↑4.00)
	$H_{\max} = 20$	71.75	66.67	65.67	88.76	74.80
	$H_{\max} = 30$	69.60 (↓2.15)	72.33 (↑5.66)	78.33 (↑12.66)	85.20 (↓3.56)	71.80 (↓3.00)

Table 13: Accuracy numbers on OOD datasets with different limits on the hypothesis bank size $H_{\max}$.

Model	Methods	Deceptive reviews	LlamaGC	GPTGC	Persuasive Pairs	Dreaddit
	HypoRefine					
GPT-4 mini	$H_{\max}$ = 10	65.60 (↓2.94)	53.67 (↓8.66)	49.67 (↓7.66)	89.00 (↓0.76)	69.80 (↓0.96)
	$H_{\max}$ = 20	68.56	62.33	57.33	89.76	70.76
	$H_{\max}$ = 30	67.40 (↓1.16)	59.33 (↓3.00)	66.00 (↑8.67)	92.20 (↑2.44)	70.80 (↑0.04)
	Literature ∪ HypoGenic					
	$H_{\max}$ = 10	69.60 (↑0.84)	68.00 (↓5.00)	74.33 (↑6.00)	90.20 (↑0.36)	69.80 (↓1.08)
	$H_{\max}$ = 20	68.76	73.00	68.33	89.84	70.88
	$H_{\max}$ = 30	68.00 (↓0.76)	77.00 (↑5.00)	86.67 (↑18.34)	90.20 (↑0.36)	66.00 (↓4.88)
	Literature ∪ HypoRefine					
	$H_{\max}$ = 10	67.40 (↓3.36)	53.67 (↓6.66)	49.33 (↓5.67)	88.40 (↓2.12)	68.00 (↓1.88)
	$H_{\max}$ = 20	70.76	60.33	55.00	90.52	69.88
	$H_{\max}$ = 30	65.40 (↓5.36)	57.33 (↓3.00)	67.33 (↑12.33)	90.40 (↓0.12)	68.00 (↓1.88)
	HypoRefine					
Llama 70B-I	$H_{\max}$ = 10	66.80 (↓4.88)	72.33 (↑1.33)	79.00 (↑3.67)	87.80 (↓1.12)	79.20 (↑0.52)
	$H_{\max}$ = 20	71.68	71.00	75.33	88.92	78.68
	$H_{\max}$ = 30	68.40 (↓3.28)	82.33 (↑11.33)	77.33 (↑2.00)	91.40 (↑2.48)	71.20 (↓7.48)
	Literature ∪ HypoGenic					
	$H_{\max}$ = 10	63.00 (↓7.28)	78.67 (↑0.34)	85.33 (↑5.33)	88.60 (↓0.76)	66.00 (↓4.68)
	$H_{\max}$ = 20	70.28	78.33	80.00	89.36	70.68
	$H_{\max}$ = 30	62.60 (↓7.68)	86.33 (↑8.00)	80.33 (↑0.33)	90.00 (↑0.64)	74.60 (↑3.92)
	Literature ∪ HypoRefine					
	$H_{\max}$ = 10	65.40 (↓8.20)	72.33 (↑1.33)	79.00 (↑5.00)	89.80 (↓1.08)	72.20 (↓1.00)
	$H_{\max}$ = 20	72.60	71.00	74.00	90.88	73.20
	$H_{\max}$ = 30	68.40 (↓4.20)	72.33 (↑1.33)	77.33 (↑3.33)	88.20 (↓2.68)	70.80 (↓2.40)

Table 14: Accuracy numbers on IND datasets with different limits on the hypothesis bank size $H_{\max}$.

Method	**Example Hypotheses and Findings**
SciMON (Wang et al., 2024)	Exploiting Social Media for Irish Language Learning: An Analysis of Twitter Data. In this context, we use social media data, particularly from Twitter, as a method for Irish language learning, because it provides a rich source of authentic and diverse language examples that can be used to enhance learning opportunities for L2 learners in a minority language setting.
HypoRefine	Reviews that provide specific accounts of the checkin and checkout processes, including exact times, the names of staff members involved, and descriptions of any unique features or services utilized (e.g., "I used the self-check-in kiosk at 3 PM"), are more likely to be truthful. Conversely, reviews that mention issues like long wait times or check-in problems without contextual details or specific examples (e.g., "the check-in took too long") are more likely to be deceptive.
Li et al. (2014)	Deceptive reviews often contain a higher frequency of first-person singular pronouns, while truthful reviews may use these pronouns less frequently.

Table 15: Examples of generated hypotheses from SciMON, HypoRefine, and findings from (Li et al., 2014). Note that the SciMON idea is about creating a new method, where our hypothesis is about a new explanation for deception detection. We also show an existing finding from Li et al. (2014) on deception detection, demonstrating that our generated hypothesis is highly relevant to the field of interest.

Dataset	Generated Hypothesis	Literature Source/Novel
DECEPTIVE REVIEWS	Deceptive reviews often contain a higher frequency of first-person singular pronouns, while truthful reviews may use these pronouns less frequently.	Li et al. (2014)
	The use of repetitive phrasing across multiple reviews is a strong indicator of deception, while truthful reviews are more likely to exhibit unique language and perspectives.	Maurya et al. (2022)
	Reviews that provide specific accounts of the check-in and check-out processes, including exact times, the names of staff members involved, and descriptions of any unique features or services utilized (e.g., "I used the self-check-in kiosk at 3 PM"), are more likely to be truthful. Conversely, reviews that mention issues like long wait times or check-in problems without contextual details or specific examples (e.g., "the check-in took too long") are more likely to be deceptive.	Novel (from data)
GPTGC and LLAMAGC	AI-generated content may struggle with maintaining coherence over longer passages, while human writing typically maintains clarity and focus.	Tang et al. (2023)
	AI-generated texts are more likely to follow conventional narrative structures, while human-written texts may experiment with form and structure.	Novel (from data)
DREADDIT	Posts that show erratic posting behavior or changes in tone (e.g., from positive to negative) are more likely to indicate stress, while consistent posting patterns with a stable tone are more likely to indicate no stress.	Wan and Tian (2024)
	Posts that exhibit avoidance behaviors (e.g., avoiding social situations or responsibilities) are more likely to indicate stress, while posts that demonstrate proactive engagement with challenges are more likely to indicate no stress.	Doan et al. (2017)
	Posts that reflect on personal struggles with mental health or addiction (e.g., "I was a severe addict") are more likely to indicate that the poster has stress, while posts that discuss academic or professional experiences without emotional turmoil (e.g., "I've explained the aforementioned to people") are more likely to indicate that the poster does not have stress.	Novel (from data)
PERSUASIVE PAIRS	Persuasive texts that incorporate rhetorical devices, such as rhetorical questions and direct appeals, are more likely to engage the reader and compel them to consider the writer's viewpoint.	Wagemans (2023)
	Texts that utilize strong, action-oriented verbs are generally more persuasive, as they convey confidence and urgency, compelling the audience to take action.	Novel (from data)
	Arguments that include a clear and compelling call to action are more persuasive, as they provide the audience with a specific next step to take, reinforcing the urgency and importance of the message.	Novel (from data)

Table 16: Examples of generated hypotheses using our method accompanied by labels indicating their sources. For hypotheses from literature, we include the specific paper, while for hypotheses that are not explicitly suggested by our literature base, we set the label to "Novel (from data)".

Method	**Invalid or Irrelevant Hypothesis**
NOTEBOOKLM	**Truthful reviews are more likely to be written in a style and tone that aligns with the reviewer's demographic information available on the platform, if any.** Conversely, deceptive reviews might exhibit inconsistencies between the writing style and the reviewer's claimed demographic, signaling a potential fabrication. **Truthful reviews are more likely to be posted at various times and days, reflecting the organic behavior of genuine guests.** Conversely, deceptive reviews, particularly those orchestrated by paid posters, might be posted in clusters or at unusual times, indicating a coordinated effort. **Truthful reviews are more likely to be written in a way that aligns with the overall sentiment expressed in the review's star rating.** Conversely, deceptive reviews might show inconsistency between the sentiment expressed in the written content and the assigned star rating, indicating a potential attempt to manipulate perception.
HYPERWRITE	**Relevant Images:** Truthful reviews are more likely to include relevant images. Deceptive reviews less likely to include images. **First-Person Pronouns:** Truthful reviews use first-person pronouns (I, my). Deceptive reviews use third-person (one). **Overly Formal Language:** Deceptive reviews use overly formal language. Truthful reviews use conversational tone.

Table 17: Examples of generated hypotheses using NOTEBOOKLM and HYPERWRITE on DECEPTIVE REVIEWS that are invalid or irrelevant, leading to degraded inference performance for these methods.