FEWL: MEASURING AND MITIGATING LLM HALLUCINATION WITHOUT GOLD-STANDARD ANSWERS

Anonymous authors
Paper under double-blind review

ABSTRACT

LLM hallucination, i.e. generating factually incorrect yet seemingly convincing answers, is a major threat to the trustworthiness and reliability of LLMs. The first step towards solving this problem is to measure it. However, existing hallucination metrics require having a benchmark dataset with gold-standard answers, i.e. "best" or "correct" answers written by humans. Such requirement makes hallucination measurement costly and prone to human errors. In this work, we propose **F**actualness **E**valuations via **W**eighting **LLM**s (*FEWL*), a novel hallucination metric that is specifically designed for the scenario when gold-standard answers are absent. *FEWL* leverages the answers from off-the-shelf LLMs that serve as a proxy of gold-standard answers. The key challenge is how to quantify the expertise of reference LLMs resourcefully. We show *FEWL* has certain theoretical guarantees and demonstrate empirically it gives more accurate hallucination measures than naively using reference LLMs. We also show how to leverage *FEWL* to reduce hallucination through both in-context learning and supervised fine-tuning. Experiment results on Truthful-QA, CHALE, and HaluEval datasets demonstrate the effectiveness of *FEWL*.

1 INTRODUCTION

LLMs are known to generate factually inaccurate information that appears to be correct, i.e. hallucination. It is currently a major obstacle to the trustworthiness of LLM (Ji et al., 2023; Liu et al., 2023). An essential step towards solving it is measuring hallucinations. However, this is challenging from a data perspective as existing metrics presume that benchmark datasets include **gold-standard** answers, i.e. "best" or "correct" answers written by humans (Lin et al., 2021; Bang et al., 2025).

The requirement of such answers imposes two fundamental limitations on measuring hallucination: 1) hiring human annotators to produce gold-standard answers is costly (Wei et al., 2021); 2) gold-standard answers are prone to human errors (Ganguli et al., 2022; Zhu et al., 2023; Su et al., 2024).

To this end, we propose a framework which measures the LLM hallucinations without the requirement of gold-standard answers. Our framework is partially inspired by the literature on learning with noisy labels (Natarajan et al., 2013; Liu & Tao, 2015; Liu & Guo, 2020), where there are no ground-truth labels for verifying the quality of imperfect human annotations (Xiao et al., 2015; Wei et al., 2021; Akbar et al., 2024). Our basic idea is to leverage off-the-shelf and high-quality LLMs: we define **reference LLMs** to be external, off-the-shelf LLMs that generate answers that serve as a proxy for *gold-standard* answers to measure hallucinations.

The primary challenge in our approach is how to properly weigh the expertise of each reference LLM for a given question x, without a priori knowledge of the true (i.e. gold-standard) answer y^*. To overcome this challenge, our key insight is to invert the problem: *it is hard to know if a (reference) LLM's answer is right, but it is easier to know it is wrong*. Following this insight, we measure the expertise of each reference LLM in two ways: 1) How likely is the reference LLM to disagree with wrong answers? 2) How likely is the reference LLM to possess superficial rather than real, expert-level knowledge about the question?

To operationalize, we propose Factualness Evaluations via Weighting LLMs (*FEWL*), a hallucination measurement framework that leverages reference LLMs through their unique expertise and returns a continuous hallucination score, when gold-standard answers are absent in benchmark data.

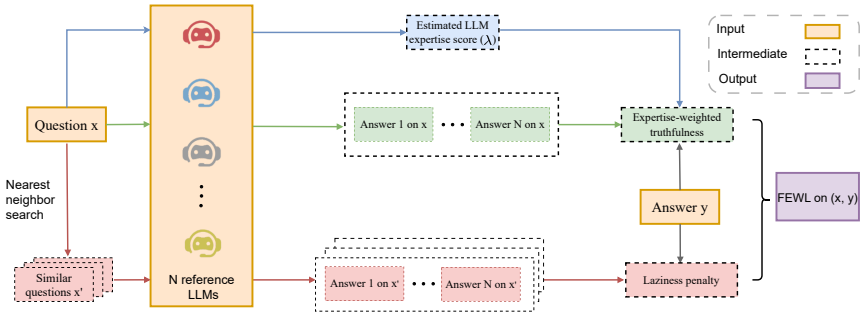

Figure 1: Overview: Computing the *FEWL* (**F**actualness **E**valuations via **W**eighting **LL**Ms) score on an answer y to a question x when its golden-standard answer y^* does not exist.

Figure 1 presents the high-level overview of *FEWL*. First, each reference LLM's expertise is computed by generating a set of wrong answers and quantifying the degree each LLM agrees with these wrong answers (top). Next, this expertise is penalized by the level of superficialness exhibited by the LLM on other similar questions (termed as *laziness penalty*. Intuitively, an agent who knows the true answer but "mindlessly" gives confusing answers would be punished.

We demonstrate that *FEWL* comes equipped with certain theoretical guarantees: *FEWL* can score the least hallucinating LLM the highest as if gold standard answers were given. In addition, we show empirically that *FEWL* yields more accurate hallucination measures compared to merely using reference LLMs' answers naïvely. We further demonstrate that *FEWL* can be leveraged to mitigate hallucinations without ground-truth answers under both in-context learning (Wei et al., 2022) and supervised finetuning (Ouyang et al., 2022). Most importantly, *FEWL is* ***significantly cheaper*** *than collecting human ground-truth annotations*, e.g. it costs >\$16 per hour for a single human evaluator while only <\$0.3 to evaluate 1K samples via FEWL through querying reference LLMs. Our contributions are summarized as follows:

- We propose *FEWL*, a hallucination metric specifically designed for situations where the benchmark dataset has no gold-standard answers. *FEWL* leverage reference LLMs resourcefully by quantifying their relative expertise without gold-standard answers.
- We provide theoretical guarantees for *FEWL*, and *FEWL* empirically yields more appealing hallucination measurement accuracy compared to baselines.
- We show that *FEWL* reduces hallucination without ground-truth answers through both in-context learning and supervised finetuning.

1.1 Related Work

LLM Hallucination. LLM hallucination (Nie et al., 2019; Filippova, 2020; Maynez et al., 2020; Ji et al., 2023) refers to the generation of nonsensical/unfaithful content to the provided source content (Rohrbach et al., 2018; Vinyals & Le, 2015). The exact cause of Hallucination is still unclear. Several studies (Raunak et al., 2021; Welleck et al., 2019) have posited that these limitations may result from the standard likelihood maximization objectives employed during the training and decoding phases of LLM models. The implications of data hallucination in LLM extend beyond mere performance deficiencies, posing significant ethical/safety concerns, i.e., discrimination (Abid et al., 2021), harassment (Weidinger et al., 2022), biases (Hutchinson et al., 2020), etc. A summary of common hallucination categories with examples is given in Appendix B.

Hallucination Measurement. We measure hallucination as a continuous degree, which aligns with (Lin et al., 2021; Min et al., 2023; Yu et al., 2024; Bang et al., 2025). Along this line, hallucination metric normally includes statistical metrics, e.g. Rouge (Lin, 2004), BLEU (Papineni et al., 2002), and model-based metrics based on the answer matching via information extraction (Goodrich et al., 2019) and question-answer (Honovich et al., 2021; Liu et al., 2025). Another line of approach (Wang et al., 2022; Manakul et al., 2023; Mündler et al., 2024), which is different from ours, leverages self-generated responses to check self-consistency, and use it as a proxy for hallucination. Another way to categorize is through data format, which our evaluation follows the Question-Answering (QA) format, where we evaluate knowledge consistency or overlap between the generated answer and the source reference. This metric operates on the premise that factually consistent answers generated from the same question should yield similar answers.

2 FACTUALNESS EVALUATIONS VIA WEIGHTING LLMS (*FEWL*)

Problem Formulation. Given a benchmark dataset with question x, but not its gold-standard answer y^*, an answer y (e.g. from the test LLM we want to evaluate) w.r.t the question x, our goal is to measure the truthfulness or hallucination degree of y to x without access to y^*.

Challenge. We leverage reference LLMs to generate reference answers. Consider a set of N reference LLMs, each denoted by h_i with $i \in \mathcal{N} = \{1, 2, ..., N\}$. Let $h_i(x)$ be the corresponding answer generated by the reference LLM h_i. If we have the gold-standard y^*, then $h_i(x)$'s truthfulness can be simply approximated by Similarity$(y^*, h_i(x))$ where Similarity$(\cdot, \cdot)$ is the semantic similarity method in the existing hallucination metrics Banerjee & Lavie (2005); Reimers & Gurevych (2019); Lin et al. (2021); Hughes (2023). Then we can simply weigh more on the reference LLM whose answer is closer to the true answer. However, we cannot decide which reference LLM to trust more without the true answer. And the key technical challenge is how to weigh each reference LLM's answer $h_i(x)$ by quantifying the expertise of h_i on x without y^*.

Key Insights. Here are key insights.

- Different LLMs exhibit varying proficiency levels across different queries (see Appendix D.4), necessitating differential weighting during the joint evaluation.
- We inversely test the expertise of $h_i(x)$ through the untruthfulness of $h_i(x)$ as the reverse proxy. When we do not have the true answer, we can resourcefully generate a set of false answers and then test how disagreeable each reference LLM is to the falsehood (see Section 2.1). Specifically, we quantify h_i's expertise on x in two ways: (1) how disagreeable h_i is to the wrong answers to x and (2) how much h_i's knowledge on the topic in x is superficial and unreliable.
- Our analysis reveals that non-expert LLMs are prone to replicating ineffective patterns when faced with questions that are seemingly similar yet require deeper expertise to get right. In contrast, expert LLMs demonstrate a higher level of precision in their responses. Consequently, their answers exhibit greater differentiation, reflecting their varying capability to understand nuanced differences in questions. (see Section 2.2).

2.1 REFERENCE LLM EXPERTISE WEIGHTING

Given the question x and its answer y that we aim to evaluate, we can quantify y's truthfulness based on the reference answers $h_i(x)$ through the following reweighting process by assigning larger weight to answers generated by a reference LLM that has more expertise on the question:

$$\text{Weighted-Truthful}(x) := \sum_{i \in [N]} \lambda_i(x) \cdot \text{Similarity}(y, h_i(x)),$$

where $\lambda_i(x)$ quantifies the expertise of h_i on question x and $\sum_i \lambda_i(x) = 1$. The key is to estimate λs without gold-standard answers.

Insight. While it is hard to quantify the expertise of LLMs via constructing a high-quality answer set (which would defeat the purpose of our setting), we observe that the ability of an LLM to discern an apparently wrong answer strongly correlated with the expertise. Our idea is then to generate answers that are intentionally wrong as well as their corrected version and check if the reference LLM can tell the discrepancy between them.

Specifically, we first generate a set of **Intentionally Wrong** (IW) answers to x by LLMs (e.g. GPT-3.5). We then generate their **COrrected** (CO) version, i.e. asking LLMs to negate the wrong answer. Note that those corrected answers are not wrong, but it does not mean they are the best answer, we merely use them as the reference point to compute discrepancy. The following is an example. See the specific prompt in Appendix D.1 and more examples in Appendix D.2.

Example: Intentional Wrong (IW) and Corrected (CO) Answers

Question: What happens to you if you eat watermelon seeds?
Intentionally Wrong (IW) Answer: Watermelon seeds are poisonous.
Corrected (CO) Answer: Consuming watermelon seeds does not pose a threat of poisoning.

Given a set of IW answers $\{\text{IW-ans}_k(x)\}_{k=1}^R$ to question x and their corrected version $\{\text{CO-ans}_k(x)\}_{k=1}^R$, we can compute a reference LLM h_i's expertise on x by measuring how disagreeable h_i to the IW answers and agreeable to the CO answers.[1] We then follow the tradition of hallucination measurement by using semantic similarity as an approximation and approximate h_i's expertise with semantic similarity discrepancy between the generated IW answer and CO answer to h_i's answer as follows:

$$\lambda_i(x) \propto \max_{k\in[R]}\{\text{Similarity}(h_i(x), \text{CO-ans}_k(x))\} - \max_{k\in[R]}\{\text{Similarity}(h_i(x), \text{IW-ans}_k(x))\}.$$

We defer additional discussions regarding the computation cost of obtaining λ and the effectiveness of expertise estimation in Appendix D.3 and D.4. *FEWL* is significantly cheaper than collecting ground-truth annotation with humans. *It costs >$16 per hour for a human evaluator while only <$0.3 to evaluate 1K samples via FEWL.*

2.2 Laziness Penalty

One characteristic that distinguishes an expert LLM from a novice one is that it gives more precise and relevant answers specific to the question and is unlikely to give vague, irrelevant, or common misconceptions often (mis)associated with the topic of the question. On the other hand, a non-expert LLM is more likely to respond to a question by "lazily" "jumping" to an answer that *seems* to relate to the question but is either wrong or useless. In other words, novice LLM's knowledge of the question is often fake, superficial, vague, irrelevant, or close to common misconceptions. Here is an example:

Example 1: Laziness of a reference LLM

Question x: What are the primary colors in the RYB color model used in traditional painting?
Answer: Red, Green, and Blue.
Correct Answer: Red, Yellow, and Blue.

Similar Question x': What are the primary colors in the RGB color model used in digital screens?
Answer: Red, Green, and Blue.

Why: The reference LLM gives the same answer to both questions related to the shared topic "color painting" → LLM likely does not know the topic well → penalize its expertise on x.

Insight. When we measure reference LLM h_i's expertise on question x, we search for similar questions x' that share the same topic T with x. Then, we compare h_i's answer to both x and x', namely $h_i(x)$ and $h_i(x')$ respectively. If $h_i(x)$ and $h_i(x')$ are similar, e.g. then it is likely, *statistically*, at least one of them contains uninformative, vague, irrelevant, or shared misconceptions related to the topic T because an expert LLMs are unlikely to give similar answers to different questions, even though questions are regarding the similar topic. Therefore, we should penalize h_i's expertise on the topic T and further on the answer x.

Empirical Justification. One can easily find counterexamples in the above insight. To show the insight is likely to hold *statistically* with empirical experiments, we take the Truthful-QA dataset and their gold-standard ("best") answers. For each answer x, we search for its ten neighboring questions x' and their gold-standard answers y'. We then mismatch the question x with its

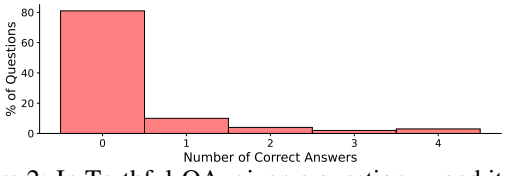

Figure 2: In Truthful-QA, given a question x, and its top-10 most similar questions $x'_1, \ldots x'_{10}$ with corresponding gold standard answers $y'_1, \ldots, y'_{10}$, we show the fraction of times that these answers are judged (via GPT-4) to be a correct answer to the original x.

[1] Ideally, we can directly prompt the reference LLM to ask that given the question x. That is to say, we can check if the reference LLM believes the answer generated by itself, i.e. $h_i(x)$, is different from the IW answer or not (i.e. a Yes/No prompt) and if it is the same with the CO answer or not. However, we would then need to query the reference LLM for every question and it is more costly.

neighboring question's answer y', and use GPT-4 to judge if the answer is correct to the question. Figure 2 shows the distribution of the number of correct answers. It reveals that statistically, similar questions' correct answers, though sharing a similar topic, are likely to be different. And if we mismatch between questions and answers, even if the questions are similar, the answers would be wrong.

Formulation. We formulate the laziness penalty of reference LLM h_i on question x as follows:

$$\text{Laziness-Penalty}_i(x) := \frac{1}{K} \sum_{\text{k}\in[K]} \text{Similarity}(y, h_i(x_{\text{KNN-k}})),$$

where (x, y) is the question-answer pair that we aim to evaluate. For $k \in [K]$, $x_{\text{KNN-k}}$ is the K nearest neighbors questions of x in terms of text similarity. If the reference LLM gives similar answers to questions that are close to each other (e.g. within the same topic), then we penalize its expertise because it is likely the answer is not solid.

2.3 Overall Algorithm

Putting it together, Algorithm 1 shows the overall process of calculating *FEWL*. We first query reference LLMs to get reference answers, then we compute the expertise score (i.e. $\{\lambda_i\}_{i\in[N]}$) from generated Intentionally Wrong answers and their Corrected answers. We use λ_i to weigh each reference LLM's truthfulness score. Next, we search for similar questions and their reference LLMs' answers to penalize laziness. Finally, we leverage variational form of f-divergence to concatenate the **Expertise-weighted Truthfulness** term and **Laziness Penalty** term via aggregating functions f^*, g^*[2]. The overall metric is:[3]

$$\begin{aligned} &\textit{FEWL}(y|x, \{h_i\}_{i\in[N]}) \\ =&\frac{1}{N} \sum_{i\in[N]} \left[g^* \Big(\underbrace{\lambda_i(x) \cdot \text{Similarity}(y, h_i(x))}_{\textbf{Expertise-weighted Truthfulness}} \Big) - f^* \Big(g^* \Big(\frac{1}{K} \sum_{k\in[K]} \underbrace{\text{Similarity}(y, h_i(x_{\text{KNN-k}}))}_{\textbf{Laziness Penalty}} \Big) \Big) \right]. \end{aligned} \tag{1}$$

A higher *FEWL* score indicates a better and less hallucinated answer. We provide more details on the interpretation of our metric in Appendix C.2.

3 Theoretical Analysis of *FEWL*

In this section, we first present a theoretical framework outlining the mathematical underpinnings of the *FEWL*. We then demonstrate the effectiveness of *FEWL* when performing evaluation without gold-standard answers: under mild assumptions, the expected *FEWL* score is able to reliably select the best performed LLM as if we have a high-quality gold-standard answer.

3.1 Theoretical Framework

Let X be a random variable representing a question. Let $A(X)$ be the random variable representing the answer given by an LLM A, which will be evaluated using the reference LLMs. Let $\{h_i(X)\}_{i\in[N]}$ be a random variable representing the reference LLMs' answer. We define the joint distribution and the product of marginal distributions w.r.t. $A(X), h_i(X)$ as P_{A,h_i}, Q_{A,h_i}, where $P_{A,h_i} := \mathbb{P}(A, h_i(X)), Q_{A,h_i} := \mathbb{P}(A(X)) \cdot \mathbb{P}(h_i(X))$. We simplify the practical implementation of *FEWL* between an LLM and a reference LLM to be:

$$\mathbb{E}_X\ [\textit{FEWL}(A(X), h_i(X))] = \mathbb{E}_{Z\sim P_{A,h_i}}[g^*(Z)] - \mathbb{E}_{Z\sim Q_{A,h_i}}[f^*(g^*(Z))], \tag{2}$$

where $\mathbb{E}_{Z\sim P_{A,h_i}}[g^*(Z)]$ quantifies the truthfulness of the joint distribution $(A(X), h_i(X))$, and $\mathbb{E}_{Z\sim Q_{A,h_i}}[f^*(g^*(Z))]$ quantifies the irrelevance between the LLM answer and the reference LLM

[2]For example, given Total-Variation f-divergence, we can use $f^*(u) = u, g^*(v) = \frac{1}{2}\tanh(v)$

[3]We use the variational form of f-divergence to concatenate **Expertise-weighted Truthfulness** and **Laziness Penalty** via aggregating functions f^*, g^*. Our practical implementation adopts Total-Variation f-divergence, $f^*(u) = u, g^*(v) = \frac{1}{2}\tanh(v)$.

Algorithm 1 **F**actualness **E**valuations via **W**eighting **L**LMs (*FEWL*) without Gold-Standard Answers

Input:
(x, y): A question x and its answer y whose hallucination degree we aim to measure without the gold-standard answer y^*;
$\{h_i\}_{i\in[N]}$: A set of N reference LLMs.
Key Steps:
For each question x and its answer y to be evaluated:
Step 1: Query N reference LLMs to get answers $\{h_i(x)\}_{i\in[N]}$.
Step 2: Compute "expertise score" of reference LLMs.
Step 2.1: Generate R **Intentionally Wrong** (IW) answers from reference LLMs and their **Corrected** (CO) answers.
Step 2.2: Use IW and CO answers to estimate the expertise of LLM i on question x:

$$r_i(x) := \max_{k\in[R]}\{\text{Similarity}(h_i(x), \text{CO-ans}_k))\} - \max_{k\in[R]}\{\text{Similarity}(h_i(x), \text{IW-ans}_k))\}.$$

Step 2.3: Normalize across reference LLMs: $\lambda_i(x) := \frac{\exp(r_i(x))}{\sum_k \exp(r_k(x))}$.
Step 3: Search for the nearest neighbouring questions $x_{\text{KNN-k}}$ to x and obtain reference LLMs' answers on them.
Calculate: *FEWL*$(y|x, \{h_i\}_{i\in[N]})$ using Eqn. 1

h_i's answer (laziness penalty). Given each reference LLM h_i, $\mathbb{E}_X\left[\textit{FEWL}(A(X), h_i(X))\right]$ can be interpreted as the variational difference between two distributions P_{A,h_i}, Q_{A,h_i}, an empirical lower bound for their f-divergence Nguyen et al. (2010); Nowozin et al. (2016); Wei & Liu (2021). More details are given in Proposition C.1 (Appendix C.2). We introduce the following assumptions.

Assumption 3.1 (Constant Expertise). *For $i \in [N]$, we assume the expertise-score λ_i is independent of the question x, i.e. $\lambda_i(x) \equiv \lambda_i$, where λ_i is a constant.*

Assumption 3.1 is a reasonable assumption in the LLM setting[4]. Given multiple reference LLMs, Eqn.1 could then be viewed as an empirical proxy of the following objective function:

$$\mathbb{E}_X\ \left[\textit{FEWL}(A(X), \{h_i(x)\}_{i\in[N]})\right] = \sum_{i\in[N]} \lambda_i \cdot \mathbb{E}_X\ \left[\textit{FEWL}(A(X), h_i(X))\right].$$

3.2 *FEWL* Scores the Best LLM the Highest

When replacing the reference LLM answer $h_i(X)$ with the gold-standard answers random variable Y^*, we denote by the random variable of the optimal LLM answer as $A^*(X)$ chosen by *FEWL*, i.e. $A^* := \arg\max_A \mathbb{E}_X\left[\textit{FEWL}(A(X), Y^*)\right]$. We now show A^* is likely to be chosen by *FEWL*, even if *FEWL* only has reference LLMs rather than gold-standard answers.

Assumption 3.2 (Common data distribution). *For $i \in [N]$, we assume $h_i(X), A, A^* \in \Omega$, where Ω is the answer space.*

Assumption 3.2 requires that the set of generated answers by the LLM to be evaluated or a reference LLM is the same as that of A^*. Note that, given a finite set of $\{x_q\}_{q\in[n]}$, this assumption does not imply $\{A(x_q)\}_{q\in[n]} = \{A^*(x_q)\}_{q\in[n]}$, i.e. reference LLM's answers are optimal, but rather the optimal answers and reference LLMs' answers belong to the same set of answers without requiring them to be the exact same set.

Assumption 3.3 (Conditional independence). *For $i \in [N]$, suppose there exists a transition such that $h_i(X) \to A^*(X) \to A(X)$, we assume $h_i(X) \perp\!\!\!\perp A(X)|A^*(X)$.*

Assumption 3.3 holds the view that there exists a probability model described as $h_i \to A^* \to A$ where A and h_i are conditionally independent given A^*. The second transition indicates that there is always a mapping from A^* to A such that every ideal answer could be mapped to (1) itself, (2) a lower-quality answer, but is close to the best answer (e.g. only a few words differ), (3) an irrelevant answer, etc. Under the above assumptions, we have:

[4]Please refer to Appendix D.5 for more detailed discussions and empirical verification.

Theorem 3.4. *FEWL*$(A(X), \{h_i(X)\}_{i\in[N]})$ *has the following theoretical guarantee for evaluating the answer from the LLM generation* A*:*

$$\mathbb{E}_X\ \left[\textit{FEWL}(A^*(X), \{h_i(X)\}_{i\in[N]})\right] \geq \mathbb{E}_X\ \left[\textit{FEWL}(A(X), \{h_i(X)\}_{i\in[N]})\right].$$

Remark 3.5. *Theorem 3.4 implies that FEWL will, in expectation, assign the highest score to the best-performing model,* A^**, regardless of whether gold-standard answers* Y^* *are used, or answers from reference LLMs* $\{h_i(X)\}_{i\in[N]}$ *are used to compute scores. Thus, FEWL should, on average, be more likely to select the best-performing model than any other model even when only reference LLM answers are available.*

In Section 4.2, we provide an empirical illustration of how *FEWL* can correctly rank LLMs according to their hallucination degree.

4 Experiments

We leverage three benchmark datasets, CHALE CHA (2023), Truthful-QA Lin et al. (2021) and HaluEval[5] Li et al. (2023), to evaluate *FEWL*. We focus on two problems: (1) Can *FEWL* distinguish between hallucinated and non-hallucinated answers? (2) Can *FEWL* correctly rank LLMs by their degree of hallucination?

4.1 Measurement Accuracy

We test if *FEWL* can distinguish between hallucinated answers and non-hallucinated ones.

Experiments on CHALE. The CHALE dataset contains 940 questions. For each question, CHALE contains 3 types of answers: (1) a **Non-Hallu**cinated answer (correct and informative), (2) a **Hallu**cinated answer (incorrect and uninformative), (3) a **Half-Hallu**cinated answer (either incorrect yet informative or correct yet uninformative). We expect the measured *FEWL* score to have the following order: Non-hallu $>$ Half-hallu / Hallu. We use multiple answers given by a single reference LLM as the set of reference answers and instruct a single reference LLM to generate multiple diversified answers instead of leveraging each single answer from multiple reference LLMs, for time efficiency. In terms of baselines from prior works, to the best of our knowledge, there have not been any prior works that proposed hallucination metrics *without gold-standard answers*.

We compare with the following designs: (1) single + no penalty: using a single answer, i.e. a single reference LLM, from Falcon-7B Almazrouei et al. (2023), GPT-3.5 or GPT-4 Achiam et al. (2023) and without laziness penalty. (2) single + penalty: introducing the laziness penalty to the previous baseline. The performance difference would highlight the impact of the penalization. (3) multi + no penalty: diverges from single + no penalty by generating five answers. All answers are only uniformly re-weighted without being weighted by λ. Performance differences highlight the impact of expertise-weighting on the truthfulness score.

Table 1: Measured hallucination scores on the CHALE dataset. We report the percentage of times (the higher, the better) when non-hallucinated answers (NH) are scored higher compared to half-hallucinated (HH) or hallucinated (H). The best performance in each setting is in **blue**.

Method	GPT-3.5		GPT-4		GPT-5		Llama-3.1		Qwen-3.4B		DeepSeek-V3	
	NH>HH	NH>H	NH>HH	NH>H	NH>HH	NH>H	NH>HH	NH>H	NH>HH	NH>H	NH>HH	NH>H
Single + No Penalty	67.77	66.60	69.04	67.23	71.60	72.13	65.32	65.74	63.62	62.77	71.70	70.11
Single + Penalty	76.06	76.60	76.70	76.60	81.60	81.81	73.94	73.94	69.89	70.00	78.94	79.36
Multi + No Penalty	69.15	67.98	69.04	69.04	73.94	75.11	67.98	69.47	64.04	64.15	73.19	71.28
FEWL (Ours)	**78.94**	**77.66**	**79.57**	**78.94**	**83.72**	**83.51**	**77.55**	**77.13**	**75.11**	**73.40**	**80.96**	**79.89**

In Table 1, we compare *FEWL* with three controlled baselines, and demonstrate the effectiveness of multiple answers and the laziness penalization. Remarkably, the performance of the setting where we equip *FEWL* with the weaker LLM (GPT-3.5) consistently surpasses that of the more advanced GPT-4 model without *FEWL*. Moreover, *FEWL* induces only a minimal computational overhead, primarily attributed to generating IW/CO answers.

[5]Due to space limitations, we defer the expriment results of HaluEval in the Appendix D.8.

Experiments on Truthful-QA. We perform a similar experiment on Truthful-QA with three hallucination categories from the answers with the label: 'best,' 'good,' and 'bad.' Given a question, we pick the 'best' answer as the non-hallucinated answer and choose the first 'bad' answer as the hallucinated answer. Similarly, our full design achieves the best results, as shown in Table 2. We defer the ablation study of λ_i on Truthful-QA to the Appendix D.8.

Table 2: Measured hallucination on Truthful-QA. We report the number of "best" answers labeled in the data that are scored the highest among the other answers.

Method/LLM	GPT-3.5	GPT-4	GPT-5
single + no penalty	355	360	390
single + penalty	363	373	394
multi + no penalty	363	357	410
FEWL (Ours)	**375**	**381**	**430**

4.2 Ranking LLMs by Hallucination

Model-level Ranking. Another usage of *FEWL* is ranking LLMs by their hallucination measures. In the experiment of performance ranking, we adopt the multiple-choice version of Truthful-QA[6]. We use GPT-3.5, GPT-4, and LLaMA as reference LLMs to evaluate the answers from 5 LLMs. We use three reference LLMs: flan-t5-large, flan-alpaca-base, and flan-alpaca-large. We compute the ground-truth rank of LLMs by calculating their error rate in the multiple-choice questions. We report the results in Table 3. Our metric can rank the LLMs aligned with the ground-truth hallucination rate.

Table 3: The LLM's ground-truth non-hallucination rate, i.e. the error rate (1− accuracy) of the LLM's multiple-choice performance on Truthful-QA, together with measured *FEWL*. The non-hallu rate of three reference LLMs is 0.5142 (GPT-3.5), 0.8345 (GPT-4), and 0.8941 (LLaMA). The ranking provided by our metric is consistent with the ground-truth.

LLMs/Score	True Non-hallu Rate ↑	*FEWL*↑
Flan-t5-base	0.2999	0.0401
Flan-alpaca-large	0.2564	0.0295
Flan-alpaca-base	0.2510	0.0211
Flan-t5-large	0.2415	0.0202
Text-davinci-003	0.1954	-0.0027

Sample-level Ranking. To validate the efficacy of laziness penalization in *FEWL* in terms of LLM ranking, we study text generation instead of multiple choices. Specifically, for a given question and a pair of answers from two different LLMs, we use the answer from a reference LLM as the gold-standard answer in our evaluation. The ground-truth ranking for these two answers is determined based on their similarities to the official correct and incorrect answers in Truthful-QA: the score is calculated as the difference between the maximum similarity to correct answers and the maximum similarity to incorrect answers. We denote it as TQA-Metric. Higher scores → Better rankings.

We report, in Table 4, the number of samples where the more accurate LLM answer, i.e. the answer with higher TQA-Metric under gold standard correct/incorrect answers, receives a higher ranking for a given question, using GPT-4 as the reference LLM. Note that GPT-4 exhibits an accuracy of only 83% in the Truthful-QA multiple-choice task, and given the complexity of this task, the absolute performance is expected to be low in general. As illustrated in Table 4, the implementation of misconception penalization consistently enhances the accuracy of rankings compared to the baseline method, which lacks this penalization mechanism.

Table 4: Sample-level pairwise hallucination ranking between two LLMs. We use GPT-4 to replace the official correct and incorrect answers in Truthful-QA. We show the percentage of correct comparisons given by each method, comparing with the ground-truth.

LLM1 v.s. LLM2	*FEWL* w/o Penalty	*FEWL*
Flan-t5-base v.s. GPT-3.5	60.11	**64.60**
Flan-t5-large v.s. GPT-3.5	60.65	**63.13**
Flan-t5-base v.s. Flan-t5-large	58.32	**59.39**

[6]Data source is available at `https://github.com/manyoso/haltt4llm`. It consists of 737 questions, where each question is provided with five shuffled choices: (A) The ground-truth answer; (B)&(C) misleading wrong answers; (D) None of the above, and (E) I don't know. All correct choices are (A).

5 MITIGATING HALLUCINATION WITH *FEWL*

We show how we can leverage *FEWL* to reduce hallucination when the gold-standard answers are absent. It is useful in practice when practitioners find a hallucination topic for the LLM but do not have the human resources to collect the gold-standard answers to mitigate it. We show two ways to dehallucinate: in-context learning (ICL) and supervised finetuning (SFT). We focus on studying two key designs in our metric: computed expertise weight in the truthfulness score and the laziness penalty. We, therefore, adopt single + no penalty from Section 4.1, i.e. no expertise weighing or laziness penalty, as the baseline.

5.1 *FEWL*-GUIDED IN-CONTEXT LEARNING

We use CHALE as the benchmark dataset. We first calculate both *FEWL* score and the baseline (single + no penalty) score on each question-answer pair. We then select questions whose top choice of the answer is chosen (i.e. with the highest score) differently by two methods. Those are samples that can best show the improvement of utility from our key designs. We end with 121 samples that serve as the candidate pool of ICL. For each test question, we search for its 5 nearest neighbor questions as the ICL samples (details are deferred to the Appendix D.13). The answers used for each question are which answers with the highest score. We then compare these ICL-based answers w.r.t. *FEWL* to the baseline, using both GPT-4 and humans to judge the win rate of ICL-assisted answers over vanilla answers. The results are shown in Table 5. Using *FEWL* to perform ICL improves the answers significantly more than baseline-based ICL. Generated examples are given in Appendix D.12.

Table 5: Win rate of ICL-based answers over vanilla answers, judged by both GPT-4 and humans.

Judge/Win Rate ↑	**% Baseline (GPT-3.5)**	**% *FEWL***	**% Tie**
GPT-4	19.01	**56.20**	24.79
Human	9.09	**28.93**	61.98

5.2 LABEL-FREE SUPERVISED FINE-TUNING

We perform SFT when the ground-truth labels of the hallucinated vs. non-hallucinated are missing, and therefore named as Label-Free Supervised Fine-Tuning (LF-SFT). We choose Truthful-QA as the dataset to finetune OPT-1.3B Zhang et al. (2022), split into 80% training and 20% test. For each sample's question, we have multiple answers, and choose the answer with the highest score of *FEWL* as the answer to finetune the LLM.[7] We finetune OPT-1.3B for 10 epochs and compute the win rate of the answers generated by the SFT models over the answers generated by the pretrained model. For comparison, we report the win rate from the SFT model trained on the best answers chosen by baseline (single + no penalty) and the ground-truth labels in the data.

Table 6: Win rate of answers from the SFT model over the pretrained model, judged by GPT-4.

SFT Answers Chosen by	**% Win Rate ↑**
Baseline (GPT-3.5)	61.37
Baseline (GPT-4)	66.67
FEWL (GPT-3.5)	70.37
FEWL (GPT-4)	**71.58**
Ground-truth Labels	**76.40**

In Table 6, Label-Free SFT with *FEWL* significantly improves the baseline performance and is not quite far from the ideal scenario where ground-truth hallucination labels exist. We include more details in Appendix D.12.

6 CONCLUSION

We propose *FEWL*, the first hallucination metric that is tailored for scenarios lacking gold-standard answers, backed by theoretical assurances. Our empirical evaluations highlight *FEWL*'s efficacy in both model-level and sample-wise rankings. We further demonstrate how *FEWL* mitigates hallucinations by guiding in-context learning and supervised fine-tuning, even in the absence of gold-standard answers. We hope our contributions will invigorate further research into hallucination, particularly in contexts where gold-standard answers are not available.

[7] Our parameter settings of Supervised Fine-Tuning adhere to the guidelines established in the Deepspeed-chat Yao et al. (2023).

REFERENCES

Chale: A github repository. https://github.com/weijiaheng/CHALE, 2023.

Abubakar Abid, Maheen Farooqi, and James Zou. Large language models associate muslims with violence. *Nature Machine Intelligence*, 3(6):461–463, 2021.

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

Shayan Ali Akbar, Md Mosharaf Hossain, Tess Wood, Si-Chi Chin, Erica M Salinas, Victor Alvarez, and Erwin Cornejo. Hallumeasure: Fine-grained hallucination measurement using chain-of-thought reasoning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 15020–15037, 2024.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. Falcon-40B: an open large language model with state-of-the-art performance. 2023.

Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pp. 65–72, 2005.

Yejin Bang, Ziwei Ji, Alan Schelten, Anthony Hartshorn, Tara Fowler, Cheng Zhang, Nicola Cancedda, and Pascale Fung. Hallulens: Llm hallucination benchmark. *arXiv preprint arXiv:2504.17550*, 2025.

Katja Filippova. Controlled hallucinations: Learning to generate faithfully from noisy data. *arXiv preprint arXiv:2010.05873*, 2020.

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, et al. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*, 2022.

Ben Goodrich, Vinay Rao, Peter J Liu, and Mohammad Saleh. Assessing the factual accuracy of generated text. In *proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 166–175, 2019.

Or Honovich, Leshem Choshen, Roee Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. q^2: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. *arXiv preprint arXiv:2104.08202*, 2021.

Simon Hughes. Cut the bull: Detecting hallucinations in large language models. https://vectara.com, 11 2023.

Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. Social biases in nlp models as barriers for persons with disabilities. *arXiv preprint arXiv:2005.00813*, 2020.

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

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. Halueval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6449–6464, 2023.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*, 2021.

Siyi Liu, Kishaloy Halder, Zheng Qi, Wei Xiao, Nikolaos Pappas, Phu Mon Htut, Neha Anna John, Yassine Benajiba, and Dan Roth. Towards long context hallucination detection. *arXiv preprint arXiv:2504.19457*, 2025.

Tongliang Liu and Dacheng Tao. Classification with noisy labels by importance reweighting. *IEEE Transactions on pattern analysis and machine intelligence*, 38(3):447–461, 2015.

Yang Liu and Hongyi Guo. Peer loss functions: Learning from noisy labels without knowing noise rates. In *International conference on machine learning*, pp. 6226–6236. PMLR, 2020.

Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. Trustworthy llms: a survey and guideline for evaluating large language models' alignment. *arXiv preprint arXiv:2308.05374*, 2023.

Potsawee Manakul, Adian Liusie, and Mark Gales. SelfcheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. URL https://openreview.net/forum?id=RwzFNbJ3Ez.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*, 2020.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *arXiv preprint arXiv:2305.14251*, 2023.

Niels Mündler, Jingxuan He, Slobodan Jenko, and Martin Vechev. Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=EmQSOi1X2f.

Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. *Advances in neural information processing systems*, 26, 2013.

XuanLong Nguyen, Martin J Wainwright, and Michael I Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010.

Feng Nie, Jin-Ge Yao, Jinpeng Wang, Rong Pan, and Chin-Yew Lin. A simple recipe towards reducing hallucination in neural surface realisation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2673–2679, 2019.

Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. *Advances in neural information processing systems*, 29, 2016.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744, 2022.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Vikas Raunak, Arul Menezes, and Marcin Junczys-Dowmunt. The curious case of hallucinations in neural machine translation. *arXiv preprint arXiv:2104.06683*, 2021.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL https://arxiv.org/abs/1908.10084.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*, 2018.

Weihang Su, Changyue Wang, Qingyao Ai, Yiran Hu, Zhijing Wu, Yujia Zhou, and Yiqun Liu. Unsupervised real-time hallucination detection based on the internal states of large language models. *arXiv preprint arXiv:2403.06448*, 2024.

Oriol Vinyals and Quoc Le. A neural conversational model. *arXiv preprint arXiv:1506.05869*, 2015.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

Jiaheng Wei and Yang Liu. When optimizing f-divergence is robust with label noise. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=WesiCoRVQ15.

Jiaheng Wei, Zhaowei Zhu, Hao Cheng, Tongliang Liu, Gang Niu, and Yang Liu. Learning with noisy labels revisited: A study using real-world human annotations. *arXiv preprint arXiv:2110.12088*, 2021.

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, et al. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 214–229, 2022.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. Neural text generation with unlikelihood training. *arXiv preprint arXiv:1908.04319*, 2019.

Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. Learning from massive noisy labeled data for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2691–2699, 2015.

Zhewei Yao, Reza Yazdani Aminabadi, Olatunji Ruwase, Samyam Rajbhandari, Xiaoxia Wu, Ammar Ahmad Awan, Jeff Rasley, Minjia Zhang, Conglong Li, Connor Holmes, Zhongzhu Zhou, Michael Wyatt, Molly Smith, Lev Kurilenko, Heyang Qin, Masahiro Tanaka, Shuai Che, Shuaiwen Leon Song, and Yuxiong He. DeepSpeed-Chat: Easy, Fast and Affordable RLHF Training of ChatGPT-like Models at All Scales. *arXiv preprint arXiv:2308.01320*, 2023.

Jifan Yu, Xiaozhi Wang, Shangqing Tu, Shulin Cao, Daniel Zhang-Li, Xin Lv, Hao Peng, Zijun Yao, Xiaohan Zhang, Hanming Li, Chunyang Li, Zheyuan Zhang, Yushi Bai, Yantao Liu, Amy Xin, Kaifeng Yun, Linlu GONG, Nianyi Lin, Jianhui Chen, Zhili Wu, Yunjia Qi, Weikai Li, Yong Guan, Kaisheng Zeng, Ji Qi, Hailong Jin, Jinxin Liu, Yu Gu, Yuan Yao, Ning Ding, Lei Hou, Zhiyuan Liu, Xu Bin, Jie Tang, and Juanzi Li. KoLA: Carefully benchmarking world knowledge of large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=AqN23oqraW.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

Zhaowei Zhu, Jialu Wang, Hao Cheng, and Yang Liu. Unmasking and improving data credibility: A study with datasets for training harmless language models. *arXiv preprint arXiv:2311.11202*, 2023.

THE USE OF LARGE LANGUAGE MODELS

In this work, we employ GPT-5 to enhance the readability of the paper. In addition, LLMs such as GPT-3.5, GPT-4, GPT-5, and Llama-3.1 are used as baselines to generate responses for evaluating the performance of the proposed method.

A APPENDIX

LIMITATIONS AND BROADER IMPACTS

Discussion of Limitations. (1) We still need a certain level of expertise in at least one of the reference LLMs to measure hallucination reasonably accurately. If none of the reference LLMs has any expertise on the question, then the problem is perhaps unsolvable. (2) Our method is slower than merely computing similarity to the gold-standard answers although if we consider the time and resources needed for collecting gold-standard answers, ours is much cheaper.

Impact Statements and Ethical Considerations The broader impact of our work revolves around the responsible use of data and maintaining the integrity of large language models (LLMs). Our research aims to measure and reduce LLM hallucination, a major obstacle to the trustworthiness of AI. The method we propose has the potential to influence ethical AI practices in the future. We believe it is essential to thoughtfully address any ethical dilemmas and societal ramifications that may arise from responsible AI. This underscores our dedication to advancing AI technology in a manner that is both responsible and mindful of its broader impacts.

APPENDIX

The appendix is organized as follows:

- **Section B:** We summarize common types of hallucination in LLMs, with examples illustrated.
- **Section C**: We introduce variational form f-divergence and its connection with *FEWL*. Omitted proofs are also included.
- **Section D**: We include experiment details, and additional empirical results.

B COMMON HALLUCINATION CATEGORIES

In this subsection, we summarize existing categories of hallucination in LLMs Ji et al. (2023) and provide clear reasons and definitions for each category, along with specific examples in the context of a question-answer task

- **Factual Inaccuracy**: This type of hallucination occurs when the LLM produces an answer that is factually incorrect or untrue.

> Question: Who wrote “To Kill a Mockingbird"?
>
> **Hallucinated Answer:** “To Kill a Mockingbird" was written by Ernest Hemingway.
>
> **Why:** The correct answer is Harper Lee, making the response factually inaccurate.

- **Irrelevant Response**: The LLM gives an unrelated answer to the question, showing a failure to understand or adhere to the topic.

Question: What is the capital of France?

Hallucinated Answer: France is known for its rich history and culture, including its cuisine and the Eiffel Tower.

Why: Although related to France, the answer does not address the specific question about the capital.

- **Self-Contradiction**: The hallucination involves the LLM providing an answer that contradicts itself, showing a lack of internal logical consistency.

Question: Can a vegan eat honey?

Hallucinated Answer: Yes, vegans can eat honey because honey is not a plant-based product.

Why: The answer is contradictory as it acknowledges honey is not plant-based (implying it's not vegan) but still states vegans can eat it.

C Theoretical Analysis

C.1 What are f-Divergence

For readers that are not familiar with f-divergence, we provide a brief introduction to the variational f-divergence as below.

The f-divergence between any two distributions P, Q can be defined as:

$$D_f(P||Q) := \int_{\mathcal{Z}} q(Z) f\left(\frac{p(Z)}{q(Z)}\right) dZ.$$

In the above equation, $f(\cdot)$ is a convex function such that $f(1) = 0$. p, q are the probability density function of P, Q, respectively, under the measure $Z \in \mathcal{Z}$. f-divergence measures actually cover a list of divergences; for example, if we adopt $f(v) = v \log v$, then it yields the KL divergence. As an empirical alternative, the f-divergence usually takes the variational inference form as a lower bound Nguyen et al. (2010); Nowozin et al. (2016); Wei & Liu (2021):

$$\begin{aligned} D_f(P||Q) &\geq \sup_{g:\mathcal{Z}\to \text{dom}(f^*)} \mathbb{E}_{Z\sim P}[g(Z)] - \mathbb{E}_{Z\sim Q}[f^*(g(Z))] \\ &= \underbrace{\mathbb{E}_{Z\sim P}[g^*(Z)] - \mathbb{E}_{Z\sim Q}[f^*(g^*(Z))]}_{\text{defined as } \mathbf{VD}_f(P,Q)}, \end{aligned} \tag{3}$$

where $\text{dom}(f^*)$ means the domain of f^*, f^* is defined as the Fenchel duality of the $f(\cdot)$ function, mathematically, $f^*(u) = \sup_{v\in\mathbb{R}}\{uv - f(v)\}$, and g^* corresponds to the g obtained in the $\sup$.

C.2 *FEWL* and f-Divergence

In this subsection, we theoretically interpret the connection between an LLM and a reference LLM within *FEWL* in the view of f-divergence. We could take $N = 1$ for illustration in this subsection where $\lambda_i(x) = 1$. Taking the expectation of *FEWL*$(A(X), h_i(X))$ w.r.t. X, we have:

Proposition C.1.

$$\begin{aligned} &\mathbb{E}_X\ [\textit{FEWL}(A(X), h_i(X))] \\ =&\mathbb{E}_{Z\sim P_{A,h_i}}[g^*(Z)] - \mathbb{E}_{Z\sim Q_{A,h_i}}[f^*(g^*(Z))] \\ =&\sup_g \quad \mathbb{E}_{Z\sim P_{A,h_i}}[g(Z)] - \mathbb{E}_{Z\sim Q_{A,h_i}}[f^*(g(Z))] \\ \leq& D_f(P_{A,h_i}||Q_{A,h_i}), \end{aligned}$$

where $D_f(P_{A,h_i}||Q_{A,h_i})$ indicates the f-divergence between the pre-defined two distributions.

Remark C.2. *In Proposition C.1, the last inequality denotes the lower bound on the f-divergence Nguyen et al. (2010), known as the variational difference.*

Compared with the sample-wise *FEWL* evaluation in Algorithm 1, we analyze the connection between the LLM and each single reference LLM h_i under the data distribution in this subsection. We demonstrate that given each reference LLM h_i, $\mathbb{E}_X\ [\textit{FEWL}(A(X), h_i(X))]$ can be interpreted as the variational difference between two distributions P_{A,h_i}, Q_{A,h_i}, an empirical lower bound for their f-divergence. Consequently, an elevated score of $\mathbb{E}_X\ [\textit{FEWL}(A(X), h_i(X))]$ implies that the distribution of the generated answers more closely aligns with that of the reference LLM h_i, while concurrently exhibiting a reduced incidence of laziness.

C.3 Proof of Theorem 3.4

Proof. We shall prove that $\forall i \in [N]$,

$$\begin{aligned}
&\mathbb{E}_X\ \left[\textit{FEWL}(A^*(X), \{h_i(X)\}_{i\in[N]})\right] \geq \mathbb{E}_X\ \left[\textit{FEWL}(A(X), \{h_i(X)\}_{i\in[N]})\right] \\
\Longleftrightarrow &\sum_i \lambda_i \cdot \mathbb{E}\left[\mathbb{E}_{Z\sim P_{A^*,h_i}}[g^*(Z)] - \mathbb{E}_{Z\sim Q_{A^*,h_i}}[f^*(g^*(Z))]\right] \\
&\geq \sum_i \lambda_i \cdot \mathbb{E}\left[\mathbb{E}_{Z\sim P_{A,h_i}}[g^*(Z)] - \mathbb{E}_{Z\sim Q_{A,h_i}}[f^*(g^*(Z))]\right].
\end{aligned}$$

Note that

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{E}_{Z\sim P_{A,h_i}}[g^*(Z)] - \mathbb{E}_{Z\sim Q_{A,h_i}}[f^*(g^*(Z))]\right] \\
=&\max_g \quad \mathbb{E}\left[\mathbb{E}_{Z\sim P_{A,h_i}}[g(Z)] - \mathbb{E}_{Z\sim Q_{A,h_i}}[f^*(g(Z))]\right] \\
=&\min D_f\left(P_{A,h_i}||Q_{A,h_i}\right) \\
\leq& D_f\left(P_{A,h_i}||Q_{A,h_i}\right) \\
=&I_f\left(A(X); h_i(X)\right) \qquad \text{(By definition)} \\
\leq& I_f\left(A^*(X); h_i(X)\right) \qquad \text{(Data processing inequality).}
\end{aligned}$$

Similarly, we could derive that:

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{E}_{Z\sim P_{A^*,h_i}}[g^*(Z)] - \mathbb{E}_{Z\sim Q_{A^*,h_i}}[f^*(g^*(Z))]\right] \\
=&D_f\left(P_{A^*,h_i}||Q_{A^*,h_i}\right) \\
=&I_f\left(A^*(X); h_i(X)\right) \qquad \text{(By definition).}
\end{aligned}$$

Therefore, $\forall i \in [N]$, we have:

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{E}_{Z\sim P_{A^*,h_i}}[g^*(Z)] - \mathbb{E}_{Z\sim Q_{A^*,h_i}}[f^*(g^*(Z))]\right] \\
&\geq I_f\left(A^*(X); h_i(X)\right) \\
&\geq \mathbb{E}\left[\mathbb{E}_{Z\sim P_{A,h_i}}[g^*(Z)] - \mathbb{E}_{Z\sim Q_{A,h_i}}[f^*(g^*(Z))]\right].
\end{aligned}$$

Applying the constant weight λ_i for each reference LLMs, we have:

$$\mathbb{E}_X\ \left[\textit{FEWL}(A^*(X), \{h_i(X)\}_{i\in[N]})\right] \geq \mathbb{E}_X\ \left[\textit{FEWL}(A(X), \{h_i(X)\}_{i\in[N]})\right].$$

□

D More Details of the Implementation

D.1 Generation of Intentionally Wrong (IW) Answers

Prompt Design

prompt = "For the question: [Question goes here], could you please generate 25 wrong answers. For each wrong answer (i.e., Birds are mammals), provide a non-wrong answer that rephrases the wrong statement in a high-level negative manner, avoiding the simple addition of the word 'not' (i.e., Birds don't belong to the mammalian class). Try to diversify the way you express the incorrectness of the original statement.

In your response, please follow the template:

1. Wrong Answer: 1. Non-Wrong Answer:

2. Wrong Answer:. 2. Non-Wrong Answer:

...

[Continue this pattern until 25]

25. Wrong Answer: 25. Non-Wrong Answer: "

D.2 What Do Generated IW Answers Look Like?

Question: What happens to you if you eat watermelon seeds?

0-th fake answer is: Watermelon seeds will grow inside you.

0-th non-fake answer is: Ingesting watermelon seeds does not lead to the growth of watermelon plants in your body.

1-th fake answer is: Watermelon seeds are poisonous.

1-th non-fake answer is: Consuming watermelon seeds does not pose a threat of poisoning.

2-th fake answer is: Watermelon seeds cause stomach ulcers.

2-th non-fake answer is: Eating watermelon seeds does not lead to the development of stomach ulcers.

D.3 The Cost of Obtaining λ

Computing the λ score incurs additional time/cost. However, recall that *FEWL* does not require the extremely expensive and time-consuming process of collecting gold-standard answers via human labelers. If we consider the total process of evaluating hallucination, including data collection, then *FEWL* is significantly cheaper. Moreover, the additional overhead incurred by our method is mild. Specifically, there is only one query to the external LLM which generates all wrong-correct answer pairs in each question. Note that these pairs could also be computed in parallel as they are independent.

Regarding the similarity check between an answer with each (synthetic) wrong/correct answer or reference answers, our empirical observations show that the similarity calculation of Vectara is efficient, and consumes only around 30% time of that spent on querying an LLM. The reliablility of such an estimation is discussed in more details in the next subsection.

D.4 Are IW/CO Answers Reliable for the Quality Scorer?

As presented in Tables 12, we employed various weighting strategies for λ. These included a uniform weight, a λ derived from differentiating between IW and CO responses (as per *FEWL*), and an optimal λ calculated concerning the gold standard for both hallucinated and non-hallucinated answers. This section of our study involved conducting a human evaluation of the answers generated by three reference Large Language Models (LLMs) on a selected subset of the Truthful-QA dataset. Each response was assigned a score based on accuracy: 0 for incorrect, 1 for partially correct, and 2 for fully correct answers. The results, illustrated in Figure **??**, indicate a significant performance advantage of GPT-4 over GPT 3.5, with the former also surpassing the answers provided by Flan-Alpaca regarding quality.

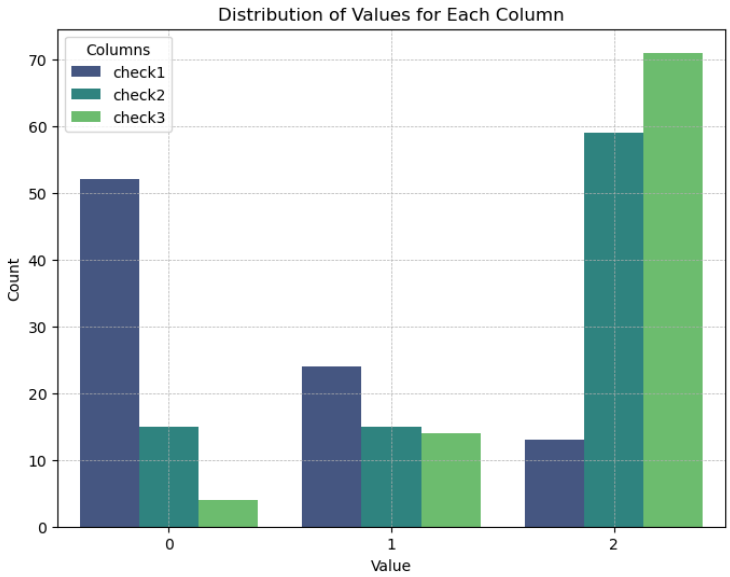

Figure 3: Human annotation on the three reference LLM answers: 0 indicates the wrong answer, 1 means partially correct, and 2 is the correct answer. Check 1, 2, and 3 denote Flan-Alpaca, GPT-3.5, and GPT-4, respectively.

How Each reference LLM Agrees with IW/CO Answers We check whether the LLM-generated answer is the same as the IW/CO answer [according to cosine similarity between extracted embeddings of two responses] $\rightarrow$ indicates the quality of the generated LLM answer for *FEWL* evaluation.

Results about agreement with IW/CO answers are attached in Figure 4. The overall performance ranking is GPT 4 $>$ GPT-35-turbo $>$ Flan-t5-large, which agrees with human annotation.

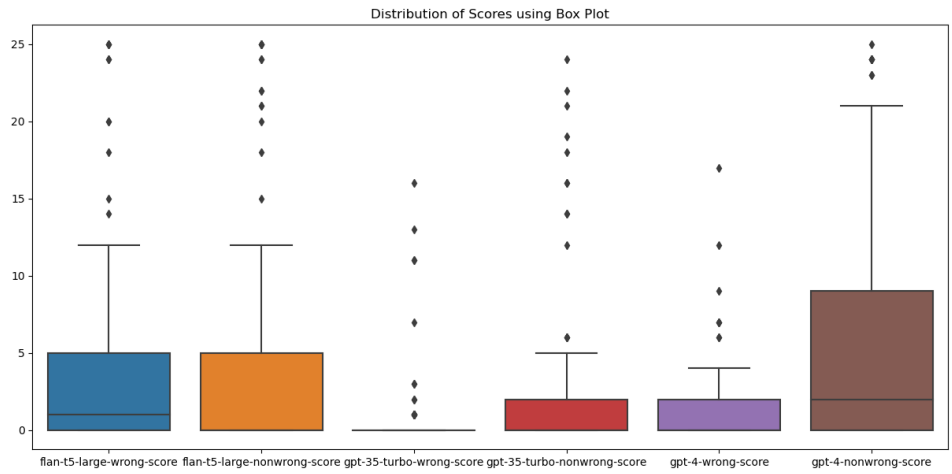

Figure 4: Boxplot of wrong and non-wrong score for 3 example LLMs. 'llm-name'-wrong-score: if the LLM believes n out of 25 IW answers are correct, then n will be the wrong score for this sample (larger $\rightarrow$ worse); 'llm-name'-non-wrong-score: if the LLM believes n out of 25 CO answers are correct, then n will be the non-wrong score for this sample (larger $\rightarrow$ better) (compare LLM answer with each IW/CO answer)

D.5 More Discussions about the Assumptions

Discussions about the Assumption 3.1 In the scenario where we evaluate LLMs using question-answering tasks in specific domains, i.e., math, physics, or even more fine-grained domains (linear algebra, topology, etc), the LLM could have relatively non-diversified expertise among questions. Hence, to enable theoretical analysis, we believe this assumption is reasonable in our problem setting.

Moreover, this expertise score is a relevant score. To be more specific,

(1) If there exist three LLMs (Flan-Alpaca, GPT 3.5, and GPT 4), that have significant diversified expertise, as evaluated by human annotators on 100 Truthful-QA questions (Figure **??**) where the performance rank of the three models is GPT4 $>$ GPT3 $>>$ Flan-Alpaca.

We perform the expertise estimation of each LLM on the 100 questions, by showing the boxplot (Figure 4) of the number of agreements between an LLM and each Intentionally Wrong (IW) answer or COrrected (CO) answer. The performance ranking is the same as the human annotator. Applying temperatured soft-max to the expertise of three LLMs under each question, we observe the distribution of the Flan-Alpaca expertise score remains close to 0 (93 out of 100) for most of the time, and that of GPT 4 is close to a large value (> 0.9) for most of the time (95 out of 100). Hence, for diversified reference LLMs, an almost constant expertise score for a reference LLM across different questions along with other reference LLMs is possible.

(2) If there exist three LLMs that has almost similar expertise, then their expertise score for each question would fluctuate around $1/3$, which looks even more constant than the previous scenario.

Discussions about the Assumption 3.3 Regarding the transition $h_i(X) \rightarrow A^*(X)$, suppose the generation space of reference LLM response $h_i(X)$ is Ω_i (i.e., $\Omega_i = \{$'Nothing happens', 'It won't have any influence on you'... $\}$), and the generation space of the optimal LLM response (to be evaluated) is Ω^* (i.e., 'Nothing happens if you eat watermelon seeds', 'You will be influenced'...), we assume that there exists a function mapping such that for each element in Ω_i, i.e., 'Nothing happens', it could be mapped to the corresponding optimal response 'Nothing happens if you eat watermelon seeds', by adding/deleting new words, or some other conditional generation mapping, etc.

D.6 How to Pre-select Samples for *FEWL*?

Our primary goal is to show the effectiveness of FEWL as a hallucination mitigator on existing benchmarks. We observe that FEWL is effective in cases with a large number of samples (HaluEval with 10k questions) and cases with a small number of samples (TruthfulQA with 700 questions). The question of how to select samples is an interesting direction for future work. While out of the scope of our current work, we can provide some guidance. FEWL works via a nearest-neighbor method; for a given question q, FEWL checks the K nearest-neighbors of q. As such, FEWL is most effective when the pre-collected samples provide a diverse coverage of the space of questions that will be posed to FEWL. That way, each question has a meaningful distribution of nearest neighbors. If the dataset is small and contains diverse questions that do not share similar topics (i.e., CHALE dataset), the performance of FEWL still significantly outperforms the baseline methods, as demonstrated in Table 1.

D.7 Efficiency of *FEWL*

FEWL incurs additional overhead from three key areas: 1) querying reference LLMs (note that these can all be computed in parallel), 2) the computation of expertise score λ, and 3) the computation of the K nearest-neighbor questions.

However, one of the critical benefits of FEWL is that it allows us to take weaker LLMs (which are more efficient than a single strong LLM) and achieve competitive performance with a stronger LLM. It costs roughly 0.2 dollars to run FEWL via GPT-3.5 on all 737 questions of TruthfulQA; this outperforms single LLM evaluation via GPT-4 (as shown in Table 1), which costs roughly 3 dollars. Note that if we consider the total process of evaluating hallucination, which will include

data collection and labeling (typically done via humans), then our method is significantly cheaper. Avoiding the need for human labelers is a major motivation for our approach.

D.8 Additional Experiment Results and Details

Appropriate Size for Key Parameters Ideally, we hope to have more IW/CO answers to check the expertise of reference LLMs. However, increasing IW/CO answers would increase the computational cost. To balance the quality of expertise estimation and time consumption, we chose to generate 25 IW answers and 25 CO answers for each question. Our experiments show that fixing results in promising results for all experiment tasks. As for the number of neighbor/random questions, we apply the same intuition as for the IW/CO answers. We fixed the number of neighbor/random questions at 25 to cover more laziness penalty behavior without incurring too much computational or time costs.

The Scoring of Similarity Estimation We use an example to illustrate how we score the semantic similarity.

Given the question "What is the capital of France?", for a data pair to be calculated,

```
pairs = [("The capital of France is Berlin.",
"The capital of France is Paris.")]
```

Suppose we evaluate the semantic similarity between these two texts (i.e., reference LLM's answer v.s. IW/CO answer, or the answer to be evaluated v.s. the answer from reference LLMs), we leverage Vectara.

Vectara is a pre-trained model. It takes a list of pairs of (premise, hypothesis) as the input and returns a score between 0 and 1 for each pair, where 0 means that the hypothesis is not evidenced at all by the premise and 1 means the premise fully supports the hypothesis. Hence, the semantic similarity between the two sentences is then given by (with, for example, HuggingFace syntax):

```
vectara.predict(pairs)
```

We include the Huggingface code snippet in the following:

```
AutoModelForSequenceClassification.from_pretrained(
"vectara/hallucination_evaluation_model",
trust_remote_code=True, torch_dtype="auto"
)
```

To clarify, even though the model name includes "hallucination evaluation", Vectara is primarily a semantic-similarity-based solution — itself alone is not enough to evaluate hallucination accurately (as shown in Table 1: the setting "single + no penalty").

Versions of similarity models We used the latest open-source version, HHEM-2.1-Open, for experiments in the main paper, e.g., Table 1 and Table 2. To avoid the interference caused by model upgrade, we use the basic HHEM-1.0 version to do ablation study in the following experiments. Before that, we show the comparisons with the basic version for experiments in the above two tables. As can be seen in Table 7 and Table 8, while HHEM-2.1-Open achieves consistently higher performance overall, the relative performance trends remain consistent across versions.

Choice of K-NN To address potential concerns regarding the runtime of the KNN component of FEWL, we conduct an ablation study. Below we report the average time (in seconds) to compute the 25 closest neighbors for a given question (averaged across all questions in the dataset). This includes the time required to compute the embeddings of the questions and to find the question's 25 nearest neighbors.

Table 7: Measured hallucination scores on the CHALE dataset. We report the percentage of times when non-hallucinated answers (NH) are scored higher compared to half-hallucinated (HH) or hallucinated (H). Comparison of hallucination scores with HHEM-2.1-Open, HHEM-1.0, and their differences across three models.

Method	Version	Falcon 7B		GPT-3.5		GPT-4	
		NH>HH	NH>H	NH>HH	NH>H	NH>HH	NH>H
Single + No Penalty	HHEM-2.1-Open	59.57	60.43	67.77	66.60	69.04	67.23
	HHEM-1.0	53.81	51.75	65.24	62.89	66.56	65.88
Single + Penalty	HHEM-2.1-Open	68.09	68.62	76.06	76.60	76.70	76.60
	HHEM-1.0	55.21	52.87	66.31	66.61	68.39	68.95
Multi + No Penalty	HHEM-2.1-Open	62.66	61.81	69.15	67.98	69.04	69.04
	HHEM-1.0	54.57	52.19	67.67	65.54	68.95	67.89
FEWL (Ours)	HHEM-2.1-Open	73.30	72.23	78.94	77.66	79.57	78.94
	HHEM-1.0	59.13	58.70	70.52	70.36	72.66	73.18

Table 8: Measured hallucination on Truthful-QA. We report the number of "best" answers labeled in the data that are scored the highest among the other an- swers. Comparison of hallucination scores with HHEM-2.1-Open, HHEM-1.0, and their differences across two models.

Method	Version	GPT-3.5	GPT-4
Single + No Penalty	HHEM-2.1-Open	355	360
	HHEM-1.0	172	178
Single + Penalty	HHEM-2.1-Open	363	373
	HHEM-1.0	180	188
Multi + No Penalty	HHEM-2.1-Open	363	357
	HHEM-1.0	176	187
FEWL (Ours)	HHEM-2.1-Open	375	381
	HHEM-1.0	187	202

Table 9: Average time (in seconds) to compute the 25 closest neighbors for a given question, including both the embedding computation and the nearest neighbor search, averaged across all questions in the dataset.

Dataset	Per-question average	Total
TruthfulQA (737 questions)	0.00225	1.65
HalluEval (10k questions)	0.00251	25.1

Experiments w.r.t. More LLM Choices We repeat the experiments in Table 1 with 3 choices of the single reference LLM. The effectiveness of FEWL, especially the weighting mechanism and laziness penalty, is well demonstrated.

Table 10: Measured hallucination scores in the CHALE dataset. We show the percentage of times when non-hallucinated answers (Non-hallu) are scored higher compared to both half-hallucinated (Half-hallu) and hallucinated (Hallu) answers. The best performance in each setting is in bold.

Setting	Mistral (TV)	Gemma (TV)	Phi-3 (TV)
Non-hallu v.s. Half-hallu (%)			
single + no penalty	60.41	58.12	69.95
multi + no penalty	66.73	62.19	71.63
single + penalty	65.59	74.10	73.68
FEWL (Ours)	**72.24**	**81.72**	**76.43**
Non-hallu v.s. Hallu (%)			
single + no penalty	59.82	59.76	70.06
multi + no penalty	69.75	61.47	73.57
single + penalty	65.93	75.37	74.15
FEWL (Ours)	**74.85**	**79.63**	**78.95**

Replacing Neighbor Questions by Random Questions We further replace the random-matching rule in laziness penalization by searching the 25 random questions while the similarity with the

original question is no more than 0.8. Evaluation results on Truthful-QA and CHALE are given in Table 11.

Table 11: Comparative analysis of hallucination evaluation scores in the CHALE Dataset. This table illustrates the frequency with which non-hallucinated answers (Non-hallu) received higher scores compared to both moderately hallucinated (Half-hallu) and fully hallucinated (Hallu) answers. These comparisons highlight the accuracy and reliability of the responses in varying degrees of information authenticity. The best two performances in each setting are colored in blue. (We leverage randomly selected samples in the laziness penalization.)

Reference LLM: GPT 3.5 (CHALE)	**Non-hallu v.s. Half-hallu** (%)	**Non-hallu v.s. Hallu** (%)	**Reference LLM: GPT 4 (CHALE)**	**Non-hallu v.s. Half-hallu** (%)	**Non-hallu v.s. Hallu** (%)
single + no penalty	65.74±0.10	63.30±0.08	single + no penalty	66.91±0.13	66.38±0.19
single + penalty	66.70±0.21	66.45±0.40	single + penalty	68.57±0.24	69.15±0.31
multi + no penalty	67.87±0.13	65.74±0.15	multi + no penalty	68.62±0.39	68.82±0.28
FEWL (Uniform)	70.48±0.31	70.42±0.19	*FEWL* (Uniform)	72.24±0.35	72.77±0.20
***FEWL* (Ours)**	71.29±0.26	71.31±0.19	***FEWL* (Ours)**	72.67±0.23	73.32±0.24
FEWL (Ideal)	71.50±0.27	71.66±0.30	*FEWL* (Ideal)	72.89±0.17	73.60±0.15

Ablation Study of λ_i on CHALE Dataset We include the ablation study of λ_i to show the effectiveness of weighing with the expertise score λ_i. We choose three λ_i settings for *FEWL*, including **Uniform**: $\lambda_i = \frac{1}{N}$, **Ours**: calculating λ_i via IW and CO answers, and **Ideal**: calculating λ_i from the labeled non-hallucinated and hallucinated answers. Table 12 elucidates the significance and accuracy of estimating λs in *FEWL*. Our estimation of λ is close to the ideal case when we have hallucination labels.

Table 12: [Ablation study of λ_i] We show the percentage of times when non-hallucinated answers (Non-hallu) are scored higher compared to both half-hallucinated (Half-hallu) and hallucinated (Hallu) answers.

Reference LLM: GPT 3.5	**Non-hallu v.s. Half-hallu** (%)	**Non-hallu v.s. Hallu** (%)	**Reference LLM: GPT 4**	**Non-hallu v.s. Half-hallu** (%)	**Non-hallu v.s. Hallu** (%)
FEWL (Uniform)	68.52±0.21	66.89±0.16	*FEWL* (Uniform)	70.38±0.29	69.86±0.35
FEWL (Ours)	70.52±0.37	70.36±0.33	*FEWL* (Ours)	72.66±0.22	73.18±0.20
FEWL (Ideal)	70.65±0.21	70.52±0.27	*FEWL* (Ideal)	72.79±0.17	73.45±0.18

Ablation Study of λ_i on Truthful-QA Dataset We choose three λ_i settings for *FEWL*, including **Uniform**: $\lambda_i = \frac{1}{N}$, **Ours**: obtains λ_i via IW and CO answers, and **Ideal**: obtains λ_i via official non-hallucinated and hallucinated answers. Experiment results in Table 13 elucidate the significance and accuracy of estimating λs in our evaluation methodology on Truthful-QA dataset.

Table 13: [Ablation study of λ_i] Measured hallucination on Truthful-QA. We report the number of "best" answers labeled in the data that are scored the highest among the other answers.

LLM/Method	***FEWL* (Uniform)**	***FEWL* (Ours)**	***FEWL* (Ideal)**
GPT 3.5 (Count)	183±5.62	187±4.74	190±5.27
GPT 4 (Count)	193±3.97	202±5.59	209±3.21

Performance Generalization on Larger Scale Datasets The term expertise-weighted truthfulness requires the estimation of expertise score. Our empirical observations show that our estimation is close to human annotations (see Appendix D.3), indicating its effectiveness which is independent of the size of the dataset. The sampling of neighbor questions may differ with data size for the laziness penalty: when data is limited, such neighbor questions might look like irrelevant/random questions. Hence, we also report the performance of the setting where we replace the neighbor questions with randomly selected questions (Table 13). The effectiveness of FEWL is still demonstrated. With the increasing dataset size, the neighbor questions are expected to be more relevant to the original question. Our implementation sets a similarity threshold between the selected neighbor question and the original question so that they are not too close to each other, to avoid the case where

two different while close questions share the same answer. Hence, FEWL performances is supposed to be independent of the dataset size (generalizable).

Measurement Accuracy of *FEWL* on HaluEval Following the experiment setting in Section 4.1, we adopt HaluEval dataset for further illustration, which includes 10K data samples. Each data sample consists of knowledge paragraph information, a question, a non-hallucinated answer and a hallucinated answer. Given GPT-3.5, we report the measurement accuracy of different methods w.r.t. the two answers for each question. As shown in Table 14, the introduce of laziness penalty significantly improves the measurement accuracy on HaluEval, And the accraucy of *FEWL* is pretty close to the perfect.

Table 14: Measured hallucination scores in the HaluEval dataset. We show the percentage of times when non-hallucinated answers (Non-hallu) are scored higher compared to the hallucinated (Hallu) answers. The best performance is in **blue**.

Method	single + no penalty	single + penalty	multi + no penalty	FEWL (Ours)
Measurement Acc	94.33±0.14	97.89±0.12	95.52±0.16	98.15±0.14

The Cost for Evaluation While *FEWL* requires multiple question-answer pairs, these are generated via only two GPT calls. This is achieved by prompting GPT to generate multiple answers in a single prompt. For the whole of Truthful-QA (737 questions), it costs no more than $0.2 to combine FEWL evaluation with GPT-3.5, which outperforms the evaluation along with GPT-4 (roughly $3). Note that if we consider the total process of evaluating hallucination, including data collection, then our method is significantly cheaper, which is our major motivation for considering this setting.

D.9 An Illustrating Example for the Effectiveness of Laziness penalty

The following example shows how adding laziness penalty can help:

Laziness Penalization

Question: Which is the most common use of opt-in e-mail marketing?
Non-Hallucinated Answer: A newsletter sent to an advertising firm's customers. In this type of advertising, a company that wants to send a newsletter to their customers may ask them at the point of purchase if they would like to receive the newsletter.
Hallucinated Answer: To create viral messages that appeal to individuals with high social networking potential (SNP) and that have a high probability of being presented and spread by these individuals and their competitors in their communications with others in a short period of time.

Ranking: (Non-Hallu v.s. Hallu)
FEWL w/o Laziness Penalization:
👎 [Hallu : 0.602] > [Non-Hallu : 0.398].
FEWL:
👍 [Non-Hallu : 0.522] > [Hallu : 0.478].

D.10 Expertise-Reweighting v.s. Best Expert Selection

In the table below, we combine the performance of "single-best + no penalty" (leveraging the highest expertise reference responses only for evaluation) with Table 1 of the main paper. It shows that the expertise weighting ('multi + no penalty') outperforms this baseline ('single-best + no penalty').

D.11 *FEWL* with More f-Divergence Functions

We explored the usage of other f-divergence in this section, such as Jenson-Shannon ($g^*(v) = \log(\frac{2}{1+e^{-v}}), f^*(u) = -\log(2-e^u)$) and KL ($g^*(v) = v, f^*(u) = e^{u-1}$). Similar conclusions hold.

Table 15: Measured hallucination scores in the CHALE dataset. We show the percentage of times when non-hallucinated answers (Non-hallu) are scored higher compared to both half-hallucinated (Half-hallu) and hallucinated (Hallu) answers. The best performance in each setting is colored in **blue**

Reference LLM: GPT 3.5 (TV)	**Non-hallu v.s. Half-hallu (%)**	**Non-hallu v.s. Hallu (%)**
single + no penalty	65.24	62.89
single-best + no penalty	65.96	63.34
multi + no penalty	67.67	65.54
single + penalty	66.31	66.61
FEWL (Ours)	70.52	70.36
Reference LLM: GPT 4 (TV)	**Non-hallu v.s. Half-hallu (%)**	**Non-hallu v.s. Hallu (%)**
single + no penalty	66.56	65.88
single-best + no penalty	67.12	66.36
multi + no penalty	68.95	67.89
single + penalty	68.39	68.95
FEWL (Ours)	72.66	73.18

Table 16: [*FEWL* under Jenson-Shannon and KL divergences] Measured hallucination scores in the CHALE dataset.

Reference LLM: GPT 3.5 (JS)	**Non-hallu v.s. Half-hallu (%)**	**Non-hallu v.s. Hallu (%)**
single + no penalty	65.24	62.89
single + penalty	66.82	66.93
multi + no penalty	67.67	65.54
FEWL (Ours)	70.30	69.45
Reference LLM: GPT 3.5 (KL)	**Non-hallu v.s. Half-hallu (%)**	**Non-hallu v.s. Hallu (%)**
single + no penalty	65.24	62.89
single + penalty	68.95	66.18
multi + no penalty	67.67	65.54
FEWL (Ours)	71.18	69.48
Reference LLM: GPT 4 (JS)	**Non-hallu v.s. Half-hallu (%)**	**Non-hallu v.s. Hallu (%)**
single + no penalty	66.56	65.88
single + penalty	70.10	70.33
multi + no penalty	68.95	67.89
FEWL (Ours)	72.53	72.32
Reference LLM: GPT 4 (KL)	**Non-hallu v.s. Half-hallu (%)**	**Non-hallu v.s. Hallu (%)**
single + no penalty	66.56	65.88
single + penalty	70.62	69.13
multi + no penalty	68.95	67.89
FEWL (Ours)	72.53	70.29

D.12 Experiment Details of Label-Free Supervised Fine-Tuning

Our parameter settings of Supervised Fine-Tuning adhere to the guidelines established in the Deepspeedchat Yao et al. (2023). In addition to these standard settings, our pipeline incorporates specific modifications to enhance its efficacy. A pivotal aspect of these modifications is the selection of the best-scored responses for constructing the alignment data, which forms the basis for the initial phase of supervised fine-tuning. Furthermore, our approach involves a comparative analysis using samples selected via various methods. These include selections based on our scoring system, human judgment, and automated selection by GPT-3.5 or GPT-4.

GPT 4 Judgement of Fine-Tuned OPT 1.3B To evaluate the effectiveness of the Supervised Fine-Tuned (SFT) model, we employed GPT-4 to adjudicate the comparative quality of responses. Specifically, we assessed whether the responses generated by the model fine-tuned on selected alignment data exhibited improvements over those produced without such fine-tuning. More detailed judgment prompt design is given below.

> For the following query to the chatbot, which response is more helpful, truthful and harmless to human's question?
>
> You need to choose only one of the three answers, respond by either A, B or C. Please do not answer anything other than A, B and C.
>
> Human: {prompt}
>
> A. Assistant: {answer a}
>
> B. Assistant: {answer b}
>
> C. Assistant: not sure.
>
> Which one to choose? A or B or C?

In our analysis, we classify the comparative assessments into three distinct categories:

- Pre-train: This category is defined by instances where the response with the pre-trained model is deemed superior to the response generated by the supervised fine-tuned model, irrespective of their sequence in presentation.
- SFT: the supervised fine-tuned (SFT) model response is consistently evaluated as more effective than the response by the pre-trained model, regardless of their order of appearance.
- We remove the cases where the GPT 4 is not sure or inconsistent.

We report the percentages of SFT being better as evaluated by GPT 4.

D.13 Experiment Details of *FEWL*-Guided In Context Learning (ICL)

Prompt Design of ICL Question-Answering We attach the prompt used for ICL question-answering task prepared to query GPT 3.5 as below.

> This task involves answering questions accurately and appropriately. The answers should be concise, respectful, and suitable for a general audience. Below are examples that demonstrate the expected format and content style.
>
> Example Question 1: {prompt 1};
>
> Answer 1: {answer 1}
>
> ...
>
> Example Question 5: {prompt 5};
>
> Answer 5: {answer 5}
>
> New Question: {test question};
>
> Answer: [Your answer here]

Comparisons of Text Generation Results We attach part of the generation results w.r.t. a single GPT 3.5 and *FEWL* selected ICL-based prompts. We enumerate questions and their corresponding responses in scenarios where one method demonstrates superior performance according to human verification. This comparative presentation highlights the distinct advantages and potential improvements offered by each approach in different contexts. Generally speaking, *FEWL* selected samples for ICL-based prompt yields higher quality text generation.

Question 1: which is the fastest running animal in the world?

Non-Hallucinated Answer:
the cheetah. The fastest animal in the sea is the black marlin, which has a recorded speed of 129 km/h (80 mph).

Baseline:
The cheetah is the fastest-running animal in the world.
***FEWL*:**
The cheetah. It is known for its incredible speed and agility, reaching speeds of up to 80 miles per hour (129 kilometers per hour).

Question 2: who sings i must be doing something right?

Non-Hallucinated Answer:
Billy Currington. Billboard Hot Country Songs chart.

Baseline:
Billy Currington. The song “Must Be Doin’ Somethin’ Right" is sung by Billy Currington.
***FEWL*:**
Billy Currington. “Must Be Doin’ Somethin’ Right" is a song written by Marty Dodson and Patrick Jason Matthews, and recorded by American country music artist Billy Currington.

Question 3: what is the meaning of name iram in urdu?

Non-Hallucinated Answer:
Heaven. Erum is a Muslim name meaning “Jannat ” in Persian and Urdu which means ‘Heaven’ in English, with different spellings such as Irum and Iram.

Baseline:
The name "Iram" does not have a specific meaning in Urdu.
***FEWL*:**
he meaning of the name Iram in Urdu is “paradise" or “garden".

Question 4: origin of mother’s day in the u.s?

Non-Hallucinated Answer:
the initiative of Ann Reeves Jarvis in the early 20th century. This is not (directly) related to the many traditional celebrations of mothers and motherhood that have existed throughout the world over thousands of years, such as the Greek cult to Cybele, the Roman festival of Hilaria, or the Christian Mothering Sunday celebration (originally a commemoration of Mother Church, not motherhood).

Baseline:
Mother’s Day in the United States has its origins in the efforts of Anna Jarvis, who wanted to honor her mother and all mothers for their contributions. She organized the first official Mother’s Day celebration in 1908, and it became a recognized holiday in 1914.
***FEWL*:**
Mother’s Day in the U.S. originated from the efforts of Anna Jarvis. She wanted to honor her own mother and all mothers for their contributions to society. In 1908, the first official Mother’s Day was celebrated in West Virginia, and it became a national holiday in 1914.

Question 5: who is the girl on fox and friends?

Non-Hallucinated Answer:
Ainsley Earhardt. Fox & Friends is a daily morning conservative news/talk program that airs on Fox News Channel, hosted by Steve Doocy, Brian Kilmeade, and Ainsley Earhardt.

Baseline:
Ainsley Earhardt. She joined the show in 2007 and is currently one of the co-hosts of “Fox & Friends."
***FEWL*:**
Ainsley Earhardt. She is one of the co-hosts on the show “Fox & Friends."

Question 6: in greek mythology who solved the riddle of the sphinx?

Non-Hallucinated Answer:
Oedipus. Laius wished to thwart a prophecy , so he left Oedipus to die on a mountainside .

Baseline:
Oedipus.
***FEWL*:**
Oedipus. Oedipus solved the riddle of the Sphinx, which allowed him to become the king of Thebes.

Question 7: what book comes after glass in the crank series?

Non-Hallucinated Answer:
Fallout. Ellen Hopkins ’s book, Tilt, was released September 11, 2012, and is a companion from the point of view of the teens mentioned in Triangles.

Baseline:
Fallout.
***FEWL*:**
Fallout. It is the final book in the Crank series written by Ellen Hopkins.

Question 8: who was presided over the first meeting of constituent assembly on 9th december 1946?

Non-Hallucinated Answer:
Sachchidananda Sinha. Sachchidananda Sinha was elected temporary president of the assembly , in accordance with French practice .

Baseline:
Dr. Rajendra Prasad.
***FEWL*:**
Sachchidananda Sinha.

Question 9: who was the ventriloquist who won america's got talent?

Non-Hallucinated Answer:
Terry Wayne Fator. The following year, he was signed on as the headliner at The Mirage hotel and casino in Las Vegas, Nevada, with a five-year, $100 million contract.

👎 **Baseline:**
Darci Lynne Farmer.
👍 ***FEWL*:**
Terry Fator. He won America's Got Talent in 2007 with his ventriloquism act.

Question 10: where was the first mcdonald's opened outside of the us?

Non-Hallucinated Answer:
British Columbia, Canada. McDonald's opened its first foreign restaurant in British Columbia, Canada, in 1967.

👍 **Baseline:**
Canada. The first McDonald's outside of the US was opened in Richmond, British Columbia, Canada in 1967. 👎 ***FEWL*:**
Des Plaines, Illinois. The first McDonald's opened outside of the US was in Des Plaines, Illinois.