InFact: Informativeness Alignment for Improved LLM Factuality

Roi Cohen
Hasso Plattner Institute
University of Potsdam
Germany
roi.cohen@hpi.de

Russa Biswas
Dept. of Computer Science
Aalborg University
Copenhagen, Denmark
rubi@cs.aau.dk

Gerard de Melo
Hasso Plattner Institute
University of Potsdam
Germany
gerard.demelo@hpi.de

Abstract

Factual completeness is a general term that captures how detailed and informative a factually correct text is. For instance, the factual sentence *"Barack Obama was born in the United States"* is factually correct, though less informative than the factual sentence *"Barack Obama was born in Honolulu, Hawaii, United States"*. While LLMs possess a tendency to hallucinate and generate factually incorrect text, they may also generate text that is indeed factually correct and yet less informative than other, more informative choices. In this work, we tackle this problem by proposing an informativeness alignment mechanism. This mechanism takes advantage of recent factual benchmarks and introduces an informativeness alignment objective. This objective prioritizes answers that are both correct and informative. A key finding of our work is that when training a model to maximize this objective or optimize its preference, we can improve not just informativeness but also factuality.[1]

1 Introduction

Large language models (LLMs) are known to capture and store extensive amounts of factual knowledge (Petroni et al., 2019; Brown et al., 2020; Roberts et al., 2020; Cohen et al., 2023a; Pan et al., 2023), as they are trained on vast quantities of text, which includes a significant body of factual knowledge. However, they often hallucinate or generate factually incorrect text (Maynez et al., 2020; Devaraj et al., 2022; Tam et al., 2023; Kaddour et al., 2023; Huang et al., 2024; Cohen et al., 2025).

Although the way LLMs represent their knowledge remains unclear (Rai et al., 2024), it can be effectively accessed via prompting (Veseli et al., 2023). For example, modern LLMs are likely to correctly complete the input prompt *"Barack Obama was born in"* with multiple different answers – *United States*, *Hawaii* and *Kapiolani Medical Center for Women & Children*. While these answers are all factually correct, they differ greatly in specificity and significantly in the level of their informativeness. This highlights an important gap: even when LLMs do not hallucinate, they often fail to provide the most informative answer. Therefore, in this work, our focus extends beyond the factual correctness of the LLMs by examining the informativeness of correct responses among an LLM's outputs. An answer is considered highly informative if it includes either the most specific response or all correct responses.[2]

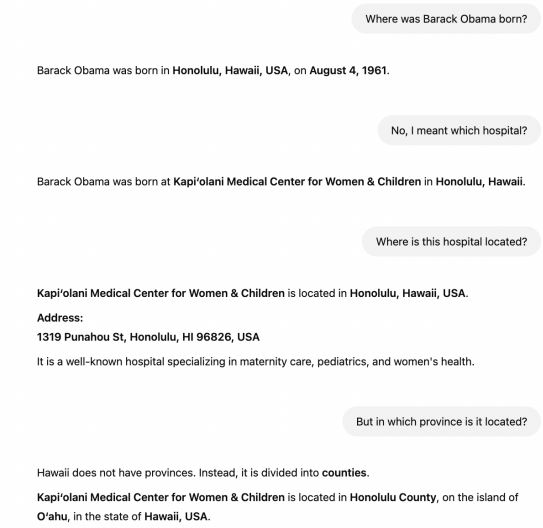

Figure 1: Illustration that an LLM knows a more informative answer than its initial answer.

Our main assumption and motivation for this work is that an LLM might indeed have correct and informative parametric knowledge about a given query, yet generate a less informative answer – as

[1] We release our code and data at https://github.com/roi-hpi/informativeness-alignment.

[2] In this study, we assume that, unless stated otherwise, the most informative answer that is also fully correct is preferred. While other preferences may be chosen for different users or applications, we argue that to produce increasingly reliable, trustworthy, and knowledgeable models it is advantageous to possess both maximal informativeness and factuality.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 17876–17888
November 4-9, 2025

illustrated in Figure 1. This might occur due to statistical or even spurious correlations between the input text (prompts or questions) and the correct yet less informative answer. For example, *"Barack Obama"* and *"The United States"* is more likely to substantially co-occur than *"Barack Obama"* and *"Kapiolani Medical Center for Women & Children"* in the training data, leading to the LLM preferring the less specific but more frequent answer.

While the distinction between factuality and informativeness may appear subtle, its practical implications are substantial. Underspecified answers, though factually correct, can introduce ambiguity or reduce the usefulness of responses in real-world applications. For instance, in the GRANOLA QA dataset, the question *"Where is Knott Rigg located?"* could be answered with *"Lake District"*, which is correct but ambiguous given multiple "Lake Districts" worldwide. A more informative answer such as *"Lake District, North West England, United Kingdom"* resolves this ambiguity and enhances utility. Similar considerations arise in high-stakes settings: in healthcare, an answer like *"cancer drugs"* to the question *"What drugs target the EGFR gene?"* is factually true but insufficiently specific, whereas *"Tyrosine kinase inhibitors; Gefitinib, Erlotinib, Afatinib, Osimertinib"* provides actionable information. Likewise, in domains such as law, education, and tutoring, informativeness ensures clarity and reliability. This motivates our focus on aligning LLMs not only to be factually correct but also maximally informative.

Therefore, we aim to align the model to prefer the most *informative* or specific answer the LLM knows. Additionally, we also address the well-known existing factual precision problem of the LLMs (Augenstein et al., 2023) by incorporating this into our alignment mechanism.

To this end, we begin by formulating the *informativeness evaluation task* and propose a novel framework to create a general informativeness dataset. In this dataset, each question or input text is paired with a set of answers associated with a hierarchical set of labels, representing different levels of informativeness. This dataset structure, which includes the informativeness metadata for each label is incorporated directly into the training.

Building on this dataset, we propose the novel training framework *InFACT*, to align a pre-trained or instruction-tuned LLM model to generate more informative and complete facts. This procedure consists of two stages – *Structure Tuning* and *Informativeness Alignment*. The *Structure Tuning* phase aims to teach the model to consider the informativeness problem setup as well as to abstain from generating misinformation, whereas the *Informativeness Alignment* seeks to train the model to answer with the most informative and complete factual answers to the questions.

Our contributions can be summarized as follows:

1. We formulate a factual informativeness evaluation task, which is integrated into both the training and evaluation pipeline. Empirical results show that our method leads to more informative answers, validating the effectiveness of this evaluation formulation.

2. We propose a novel two-stage training framework called *InFACT*, enhancing the model to generate more informative answers.

3. We evaluate our proposed framework with different models and demonstrate that it is effective in both improving information and factual accuracy. Our findings show significant improvement in factual precision, suggesting that the model learned to refrain from answering questions it would have made mistakes on, while overall high factual recall is preserved despite minor drops.

4. Finally, we conduct in-depth analyses of both techniques individually and examine potential spurious correlations. Our results confirm that the proposed framework outperforms several strong baselines focused on factual accuracy alone, enhancing both informativeness and factual reliability simultaneously.

2 Informativeness Problem Setup

2.1 Background

We first rigorously define a general setup of the *informativeness evaluation task*, which is the foundation of our alignment model. In factual question answering, each question q_i, e.g., (1)*What is the location of the capital of Australia?"* has a correct answer t_i (e.g., *the city of Canberra"*), which may appear in different string representations but conventionally refers to the same real-world entity. However, for a given factual question q_i, there may be several different correct answers. We distinguish the following three scenarios (see Fig. 2 for details):

1. **Multiple-Answer Questions.** A question with multiple correct answers – such as the question *"What are the awards received by Barack Obama?"*, which has several correct answers referring to different entities.

2. **Descriptive QA**. Responses may differ in their verbosity, yet point to the same real-world entity. For example – the text *"The city is located inland, about 150 kms from the coast, and is surrounded by picturesque mountains ..."* is a correct answer to (1).

3. **Granularity-based QA**. Answers to certain questions may have a varying level of completeness yet be factually correct; for instance, the valid answers to the question *"Where was Barack Obama born?"* would be *"Kapiolani Medical Center for Women & Children"*, or *"Honolulu"*, where the hospital is located, or *"United States"*, where the city is located, or all of them, depending on the level of granularity.

Considering these types of QA, we formulate the dataset and the evaluation.

2.2 Dataset Formulation

An informativeness evaluation question-answering dataset is composed of a collection of questions, where we define a hierarchy of factually correct answers with varying levels of informativeness for each question. For a given question q_i, Level 1 includes the most *informative* answers, while L_i has the least informative ones; L_i being the number of levels in the hierarchy of answers. More formally, we define a general informativeness-evaluation dataset as $\mathcal{D} = \{(q_i, H_i)\}_{i=1}^{N}$, where N is the number of questions in the dataset, q_i is the i-th question in the dataset, and $H_i = (A_1, \ldots, A_{L_i})$ is the corresponding hierarchy of answers. A given $A_j = \{a_1, ..., a_{b_j}\}$ represents the j-th level in such a hierarchy, which is a set of all the possible answers at this level. For example, in Figure 1, the answer *Kapiolani Medical Center for Women & Children, Honolulu, Hawaii, United States* is considered to be in set A_1, *Honolulu, Hawaii, United States* in A_2, and *United States* in A_3, reflecting the descending order of informativeness.

3 Informativeness Alignment

In this section, we present a detailed overview of our proposed method, *InFACT*, which leverages the informativeness evaluation dataset introduced in the last section to improve both the informativeness and factual correctness of LLM outputs.

3.1 Baseline Mechanism

In the baseline model, we select the most informative set of answers, for instance A_1, as the gold label for each question, and fine-tune the LLM using a randomly chosen answer from the selected set A_1. The goal is to train the model to learn the structure, preferring the most informative answer, such that it follows this structure during testing.

3.2 Abstention Detection

Prior to our Informativeness Training (see Section 3.4), we introduce an automated LLM agnostic abstention detection mechanism to detect when an LLM abstains from answering a question. This is needed because when the LLM abstains from answering instead of answering incorrectly, we provide a higher reward for the model, thereby enhancing factual reliability. We define a function that accepts natural language text and determines whether the LLM output is any form of abstention from answering. In this work, we leveraged *GPT-4*, a powerful LLM in an in-context learning setup. A wide range of in-context data for abstention detection is collected from the LLM outputs of our experiments. For a given text x, we consider $x \in$ ABSTAIN, if and only if our abstention detection mechanism classifies it as a text that is in abstaining form. This mechanism encourages the model to prefer to err on the side of caution.

3.3 Structure Tuning

For efficient training, we first tune the model for the following:

1. Optimizing the model to consistently generate the most factual and informative answers;

2. Learning to abstain more accurately, which most of the existing models lack without any additional training, often generating misinformation instead.

Therefore, prior to the informativeness-tuning phase, in structure tuning, we conduct the following training: For every question q_i in our training set, we first let the model generate its answer. If the answer is correct, namely, it appears in one of the levels of H_i, then we train the model with a random answer from the level above it, i.e., the model

is encouraged to learn a more informative answer. Otherwise, if the answer is in the highest level, i.e., it is the most informative answer according to our dataset, we take no further action, as the model has achieved the desired outcome. If the answer is incorrect, we teach the model to abstain. More formally – for a model M and a question q_i, the gold label, $L_{\text{G}}(M, q_i)$, is defined by:

$$L_{\text{G}}(M, q_i) = \begin{cases} \textsc{Random}(A_{j-1}) & \text{if } \exists j \geq 2 : \hat{y} \in A_j \\ \textsc{Skip} & \text{if } \hat{y} \in A_1 \\ \text{IDK} & \text{otherwise.} \end{cases} \quad (1)$$

Here, IDK denotes a response such as "I don't know the answer".

3.4 Informativeness Alignment

To fully leverage the hierarchical structure of our dataset, we introduce a preference-based training mechanism to reward the model for a more informative and correct answer. Additionally, we enhance abstention for potentially incorrect or misleading LLM outputs. As mentioned earlier, our dataset consists of a hierarchy of answers for each question, which is exploited to design a reward function to train an LLM as a policy to maximize its reward. Recall that for each question q_i in our dataset, there is a corresponding hierarchy of answers H_i as defined earlier. Let M be our LLM and $M(q_i) = \hat{y}$ be the model's answer to the question q_i. We define the following reward function:

$$R(M, \hat{y}) = \begin{cases} \frac{1}{\sqrt{j}} & \text{if } \exists j : \hat{y} \in A_j \\ 0 & \text{if } \hat{y} \in \textsc{Abstain} \\ -1 & \text{otherwise.} \end{cases} \quad (2)$$

Observe that as long as there exists j such that $\hat{y} \in A_j$, this means that $\hat{y}$ is correct, and thus we provide the model with a positive reward. The magnitude of the reward depends on the level in the hierarchy in which this correct answer is located; i.e., the more informative the answer, the larger the reward. In cases of the model output being wrong, meaning, its answer $\hat{y}$ is not at any level of the hierarchy, the model receives a negative reward, i.e., it is penalized to prevent hallucinations and factual errors.

Having defined our reward function, we use RL techniques – specifically PPO – to train an LLM as a policy to maximize the reward. We also consider the reward function as a preference score and use a preference-training algorithm, specifically DPO in our experiments. This approach effectively balances completeness and correctness, resulting in more informative and factually reliable outputs.

Statistic	Value
Total number of examples	3,000
Average number of levels per example	8.9
Average number of answers per level	4.6

Table 1: Summary statistics of the dataset.

3.5 Overall Framework

To summarize, our overall training mechanism of *InFACT* is as follows: (1) **Initialization**: We begin with any foundation model, which can be either a pre-trained or an instruction-tuned LLM. (2) We deploy our **Structure Tuning** mechanism (Section 3.3) to extract relevant information and lay the ground work for the Informativeness Alignment training. (3) Finally, we introduce the second training phase, the **Informativeness Alignment** technique (Section 3.4) to optimize the model to generate more informative and factually correct outputs.

4 Experimental Setup

In this section, we describe our experimental setup, including the investigated models, our baselines, and experimental details of our training procedure.

Models. In our experiments, for both the baseline and the full completeness training, we use the following models: `Llama-3.2-1B`, `Llama-3.2-3B`, `Llama-3.1-8B`, and `Llama-3.1-8B-Instruct` (Touvron et al., 2023; Dubey et al., 2024), `Mistral-7B-v0.1` (Jiang et al., 2023) and `Qwen2.5-7B` (Bai et al., 2023; Yang et al., 2024).

Training Data. In order to create our training dataset, we follow the setup in Section 2, as well as the formulation in Section 2.2. To construct our training set, we use the following datasets: GRANOLA QA (Yona et al., 2024), QAMPARI (Amouyal et al., 2023), and RoMEQA (Zhong et al., 2022). Specifically, we randomly sample 1k examples from each. The examples from GRANOLA QA are already organized in the way we defined in Section 2.2, i.e., as a hierarchy of answers consisting of a sequence of answer levels. For the examples from QAMPARI and RoMEQA, we adjust the structure in the following way: In

each of these two datasets, each question has a list of answers. We construct our hierarchy with all single answers as the gold answers in the lowest level, followed by all possible pairs, triplets, and so on – culminating in the full list of answers with all permutations as the most informative response. Table 1 shows the statistics about the dataset. The main takeaway is the average number of levels per example, which is 8.9, as demonstrating that each of the questions is associated with a broad range of different informative answers. Table 8 in Appendix A.2 presents two sample examples from the dataset, along with the corresponding training details provided in the same section.

Evaluation Data. For informativeness evaluation, we consider the test split of the benchmarks we used to build our training data from (GRANOLA QA, QAMPARI and RoMEQA). For factual accuracy, we consider several other QA datasets: TriviaQA (Joshi et al., 2017), PopQA (Mallen et al., 2022), TruthfulQA (Lin et al., 2021), Natural Questions (Kwiatkowski et al., 2019), and PIQA (Bisk et al., 2019). These cover a wide range of questions, for example, general knowledge trivia questions (TriviaQA), subject-relation-object facts phrased as questions (PopQA), real-world user queries (Natural Questions), questions about human falsehoods (TruthfulQA), and physical commonsense reasoning questions (PIQA). We consider the closed-book open-ended setting, where we do not provide any context or answer choices to the model.

Baselines. For informativeness evaluation, we compare the `informativeness`-tuned model with its original base model without any further training. For factuality evaluation, we proceed similarly, but consider three different baselines:

1. **Confidence Threshold** baseline: We use the predicted probability of the first generated token from the LM's modeling head as a confidence score, following Yoshikawa and Okazaki (2023). If this confidence score is greater than a fixed threshold, we consider it as a valid generation; otherwise, we consider this as an uncertainty expression (analogous to abstention in our model). To create a strong baseline, we find the best threshold via hyperparameter tuning on the development set.

2. **Prompting** baseline: We adopt a zero-shot approach where we instruct the model to be more informative in its answers but also to abstain in cases it does not know the answer. We use the following prompt: *"Please answer the following question. Please answer with the most informative answer you can. Please refrain form answering if you don't know the correct answer. The question is: "*.

3. **ICL** baseline: We adopt a few-shot approach by implicitly instructing the model to be more informative in its answers, using in-context demonstrations. Specifically, we sample 8 examples from our training set, and use the most informative answers as gold answers.

4. **P(True)** baseline (Kadavath et al., 2022): Given an input sentence to complete, say I, we use the original model to generate the completion, A. We then concatenate I and A and ask the model: *"Please answer either with 'true' or 'false' only. Is it true that: IA"*. If the model answer is not 'true', we consider this specific example as unknown for the model – similar to our model's abstention.

5. **Semantic Entropy** baseline (Kuhn et al., 2023; Aichberger et al., 2024): We sample K text generations from the model, encoding them using a state-of-the-art semantic encoder and cluster their encodings. If the largest cluster size is larger than $\frac{K}{2}$, we take a random generation from this cluster as the model's answer; otherwise, consider it as an unknown.

6. **FT Baseline**: This is described in Section 3.1. This baseline evaluation distinguishes the effects of our proposed informativeness alignment method and the created training set.

Evaluation. We assess the effect of our proposed model by measuring its *informativeness, factuality, and knowledge retention*. For informativeness, we adopt the metrics that have been used in the original benchmarks. For example, GRANOLA QA provides a specific evaluation metric for informativeness that takes into account the location of the model's answer in the hierarchy of answers. QAMPARI and RoMEQA measure the precision and recall of the list of answers given by the model. For factuality and knowledge recall, we use the following metrics: (1) **Precision**: the portion of factually correct answers out of all the questions that have a non-abstaining answer (as determined

	GRANOLA QA		QAMPARI			RoMEQA		
	Accuracy	**Informativeness**	**P**	**R**	**F1**	**P**	**R**	**F1**
`Llama-3.2-1B`	45.1	37.5	20.4	3.1	5.4	24.4	1.5	2.8
`Llama-3.2-1B` + prompting	50.6	39.1	17.7	6.4	9.4	19.0	2.9	5.0
`Llama-3.2-1B` + FT	51.1	45.2	20.6	7.1	10.6	23.5	4.9	8.1
`Llama-3.2-1B`+ informativeness-alignment	**56.7**	51.9	23.9	**10.8**	**14.9**	**25.9**	10.4	14.8
+ PPO	52.1	**52.4**	**24.7**	10.7	**14.9**	24.1	**11.2**	**15.3**
`Llama-3.2-3B`	51.3	42.9	**23.1**	4.3	7.2	23.6	2.3	4.2
`Llama-3.2-3B` + prompting	55.9	46.0	19.6	7.7	11.0	21.8	6.7	10.2
`Llama-3.2-3B` + FT	57.2	50.3	23.0	7.5	11.3	23.7	7.1	10.9
`Llama-3.2-3B`+ informativeness-alignment	**62.6**	**56.8**	22.4	**11.5**	15.2	**23.8**	**11.8**	**15.8**
`Llama-3.1-8B`	61.1	53.4	22.4	8.9	12.3	30.4	4.9	15.8
`Llama-3.1-8B` + prompting	71.2	56.8	21.5	9.5	13.2	**32.5**	6.6	10.2
`Llama-3.1-8B` + ICL	63.1	58.5	22.7	13.9	17.2	32.4	8.6	13.6
`Llama-3.1-8B` + FT	71.5	58.0	22.5	12.9	16.4	32.5	13.1	18.7
`Llama-3.1-8B`+ informativeness-alignment	**74.8**	**64.2**	**22.7**	**17.4**	**19.7**	**32.5**	**13.1**	**18.7**
- Structure-Tuning	64.5	61.1	21.3	11.8	15.2	26.7	8.3	12.7
- Informativeness-Alignment	64.8	57.4	21.5	9.7	13.4	29.4	6.4	10.5
`Mistral-7B-v0.1`	59.9	53.9	21.0	8.6	12.2	30.1	5.5	9.3
`Mistral-7B-v0.1` + ICL	65.1	60.0	22.8	10.4	14.3	30.3	7.9	12.5
`Mistral-7B-v0.1` + FT	62.5	61.3	21.7	11.9	15.4	29.6	10.4	15.4
`Mistral-7B-v0.1`+ informativeness-alignment	**72.5**	**64.1**	22.9	**13.6**	17.1	**30.8**	**12.9**	**18.2**
`Qwen2.5-7B`	55.5	51.7	22.4	5.2	8.4	28.6	4.9	8.4
`Qwen2.5-7B` + ICL	56.5	53.9	22.9	5.8	9.3	28.5	7.1	11.4
`Qwen2.5-7B` + FT	56.4	55.5	22.7	6.9	10.5	28.4	9.9	14.7
`Qwen2.5-7B` + informativeness-alignment	**68.8**	**60.7**	22.8	**11.5**	15.3	**28.6**	**10.9**	**15.8**

Table 2: Evaluation scores of `Llama-3.2-1B`, `Llama-3.2-3B`, `Llama-3.1-8B`, `Mistral-7B-v0.1` and `Qwen2.5-7B` on our informativeness evaluation QA benchmarks.

by our abstention detection model, see Section 3.2), i.e., the questions that the model was certain about, and tried to be factually complete. (2) **Recall**: the portion of factually correct answers out of all the questions in the dataset, namely, the portion of knowledge retention the model has, out of the entire test set. (3) **F1**: the harmonic mean of precision and recall. In the case of base models without additional calibration methods, the precision, recall, and F1-scores all correspond to their accuracy.

5 Results

We present the main results obtained in our experiments, focusing on testing the influence of our method on the factual accuracy and informativeness of the answers generated by the models.

5.1 Informativeness.

To measure the informativeness dynamics post-training, we evaluate the model using the test set of the three different datasets used to build our training data: GRANOLA QA, QAMPARI, and RoMEQA, using the original metrics proposed for each dataset. Table 2 summarizes the results. Evaluating Granola QA, the informativeness measurement clearly goes up, as does the recall of the answers on QAMPARI and RoMEQA. We thus can conclude that overall, our models demonstrate a substantial improvement at each of these benchmarks, suggesting that the aligned model has indeed captured the concept of informativeness and has improved in it. Moreover, comparing to our FT baseline, we observe substantial improvements in results demonstrating the significance of our alignment method in obtaining these improved results.

5.2 Factual Accuracy.

Tables 3, 4, 5, 6, and 9 present the performance of our models, as well as the relevant baselines, on our QA benchmarks. As can be seen, across all model sizes and all benchmarks, the overall F1 performance of our models is the highest. These gains are mostly a result of the increase in precision, that is, the model has learned to generate significantly fewer incorrect answers and to refrain where appropriate. In addition, recall performance tends to not show any major drop, which implies that with

	TriviaQA			PopQA			TruthfulQA			Natural Questions			PIQA		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
`Llama-3.2-1B`	48.1	48.1	48.1	35.0	35.0	35.0	29.5	29.5	29.5	20.5	20.5	20.5	71.9	71.9	71.9
`Llama-3.2-1B` + Confidence Threshold	60.4	38.3	46.9	42.9	25.1	31.6	45.8	14.9	22.5	33.0	11.4	16.9	78.0	55.8	65.0
`Llama-3.2-1B` + prompting	51.8	40.8	45.6	34.1	22.8	27.3	39.4	21.0	27.4	35.5	14.9	20.1	76.9	59.6	67.1
`Llama-3.2-1B` + P(True)	54.4	40.5	46.4	39.6	22.0	28.3	44.4	20.7	30.0	36.2	15.5	21.7	79.3	60.1	68.4
`Llama-3.2-1B` + Semantic Entropy	57.8	41.0	48.0	41.2	21.9	28.6	44.9	21.1	28.7	35.3	18.4	24.2	78.5	65.1	71.2
`Llama-3.2-1B` + FT	49.1	46.5	47.8	32.4	37.0	34.5	40.1	24.9	30.7	20.4	23.8	22.0	73.5	69.8	71.6
`Llama-3.2-1B`+ informativeness-alignment	**63.8**	40.6	**49.6**	**47.6**	31.0	37.5	49.8	24.5	32.8	**35.9**	20.0	**25.7**	**78.1**	70.5	**74.1**
+ PPO	61.5	43.0	50.6	47.2	31.5	**37.8**	**50.1**	24.8	**33.2**	34.1	20.4	25.5	76.6	71.2	73.8
- Structure-Tuning	54.2	41.2	46.8	40.5	30.7	34.9	43.2	23.8	30.7	32.9	20.2	25.0	73.2	70.8	72.0
- Informativeness-Alignment	54.2	31.0	39.4	49.8	24.5	32.8	35.9	20.0	25.7	31.9	18.1	23.1	74.6	67.8	71.0

Table 3: Precision (P), Recall (R), and F1-scores for `Llama-3.2-1B`. Our informativeness-aligned model achieves the best precision with minor decreases in recall, outperforming previous work.

.

	TriviaQA			PopQA			TruthfulQA			Natural Questions			PIQA		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
`Llama-3.2-3B`	50.2	50.2	50.2	36.1	36.1	36.1	32.4	32.4	32.4	26.0	26.0	26.0	75.4	75.4	75.4
`Llama-3.2-3B` + Confidence Threshold	58.9	40.1	47.7	37.2	29.2	32.7	34.8	25.8	29.6	34.6	20.0	25.3	78.5	66.7	72.1
`Llama-3.2-3B` + prompting	52.3	43.9	47.7	36.2	33.4	34.7	32.1	32.6	32.3	29.7	25.4	27.4	79.0	68.2	73.2
`Llama-3.2-3B` + P(True)	58.2	42.5	49.1	36.7	31.0	33.6	35.5	26.1	30.1	34.4	24.9	28.9	76.6	69.7	73.0
`Llama-3.2-3B` + Semantic Entropy	56.5	44.1	49.5	39.0	29.7	33.7	38.9	25.7	30.1	34.0	25.3	29.0	76.6	71.4	73.7
`Llama-3.2-3B` + FT	47.2	47.2	47.2	39.5	36.7	38.0	36.6	30.8	33.4	27.8	27.2	27.5	77.5	74.9	76.1
`Llama-3.2-3B`+ informativeness-alignment	**64.0**	42.9	**51.4**	**44.0**	33.7	**38.2**	**45.5**	29.8	**36.0**	**48.1**	25.4	**33.2**	**83.2**	72.5	**77.5**
- Structure-Tuning	55.7	43.1	48.6	36.9	34.4	35.6	37.4	28.4	32.3	33.3	25.0	28.6	75.2	73.4	74.3
- Informativeness-Alignment	56.8	40.3	47.8	37.6	30.1	33.4	37.2	24.6	29.6	26.9	24.3	29.3	76.8	67.6	71.4

Table 4: Precision (P), Recall (R), and F1-scores for `Llama-3.2-3B`. Our informativeness-aligned model achieves the best precision with minor decreases in recall, outperforming previous work.

high probability, a substantial portion of the knowledge remains preserved in the model's parameters after our training. Overall, our results are promising, as they demonstrate that the informativeness of the answers generated by the model has also increased – that is, we can improve both accuracy and informativeness without a tradeoff between the two. Notably, similar to the informativeness results, our alignment method leads to a significant improvement in factual precision compared to the fine-tuned (FT) baseline, highlighting the impact of our approach.

5.3 PPO vs. DPO

As discussed in Section 3.4, to train our models, we can use both policy-optimization methods like PPO and preference-based algorithms such as DPO. Here we compare these two in terms of results. Tables 2 and 3 include the results of our model using the PPO algorithm, with `Llama-3.2-1B`. As observed, the resulting gap between these two algorithms is not compelling enough to be statistically significant. We thus use only DPO for the larger models, due to resource considerations.

5.4 Ablation Study

As outlined earlier, our method is composed of two components: Structure Tuning (Section 3.3) and Informativeness Alignment (Section 3.4). Here we study the specific impact of each of them separately. We follow the same experimental setup, yet we apply it to two different models: one only trained via Structure Tuning and one trained via Informativeness Alignment. Tables 2, 3, 4, and 5 show these results for the different model sizes on the different evaluation datasets we have used. The pattern is evident – omitting the Structure Tuning degrades the factual precision of the resulting model. We attribute this decline to foundation models' limited ability to effectively abstain, which frequently leads to the production of misinformation.

No-Informativeness Rewarding. To further disentangle the contribution of our method's components, we also tested a variant where the Informativeness Alignment step is modified such that informativeness rewarding is removed. Specifically, all positively rewarded answers are mapped to 1, abstentions remain rewarded with 0, and hallucinations are penalized with -1 (with DPO training adapted accordingly). Table 7 presents the results

	TriviaQA			PopQA			TruthfulQA			Natural Questions			PIQA		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
`Llama-3.1-8B`	53.5	53.5	53.5	38.5	38.5	38.5	34.9	34.9	34.9	28.9	28.9	28.9	80.2	80.2	80.2
`Llama-3.1-8B` + Confidence Threshold	61.6	39.5	48.1	50.8	31.2	38.6	55.6	21.9	31.4	49.6	18.7	27.2	84.8	61.3	71.2
`Llama-3.1-8B` + prompting	64.2	46.1	53.7	51.4	31.9	39.4	52.5	26.8	35.5	46.4	24.1	31.7	82.5	67.4	74.2
`Llama-3.1-8B` + ICL	65.3	46.1	54.0	58.4	33.8	42.8	52.4	28.9	37.2	55.7	21.4	30.9	83.4	67.8	74.8
`Llama-3.1-8B` + P(True)	65.4	44.5	53.0	54.9	30.7	39.4	55.2	23.7	33.2	50.2	21.5	30.1	82.2	68.5	74.7
`Llama-3.1-8B` + Semantic Entropy	64.5	45.8	53.6	55.5	31.4	40.1	50.9	24.6	33.2	51.2	22.3	31.1	79.1	68.9	73.6
`Llama-3.1-8B` + FT	54.7	54.7	54.7	37.1	37.1	37.1	40.1	32.1	35.6	27.2	29.8	28.4	79.9	80.2	80.0
`Llama-3.1-8B`+ informativeness-alignment	**70.1**	47.2	**56.4**	**65.5**	35.9	**46.4**	**59.4**	26.2	**36.4**	**61.7**	21.5	**31.8**	**86.9**	75.1	**80.6**
- Structure-Tuning	59.5	49.1	53.8	58.0	35.6	44.1	51.3	26.0	32.7	50.0	21.8	30.4	81.0	68.5	74.2
- Informativeness-Alignment	57.1	49.9	53.2	49.8	35.1	41.2	46.8	26.7	34.0	43.0	22.1	29.2	84.1	64.8	73.2

Table 5: Precision (P), Recall (R), and F1-scores for `Llama-3.1-8B`. Our informativeness-aligned model achieves the best precision with minor decreases in recall, outperforming previous work.

	TriviaQA			PopQA			TruthfulQA			Natural Questions			PIQA		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
`Mistral-7B-v0.1`	53.1	53.1	53.1	35.5	35.5	35.5	33.2	33.2	33.2	27.8	27.8	27.8	77.8	77.8	77.8
`Mistral-7B-v0.1` + prompting	60.5	45.6	52.0	47.8	34.3	39.9	44.9	30.1	36.0	48.7	22.2	30.5	79.4	68.9	73.8
`Mistral-7B-v0.1` + ICL	48.5	56.1	52.0	56.9	27.6	37.2	32.5	34.9	33.7	50.5	21.5	30.2	72.6	79.5	75.9
`Mistral-7B-v0.1` + FT	57.6	50.2	53.6	42.0	34.8	38.0	33.5	31.1	32.3	32.8	22.8	26.9	77.9	76.8	77.3
`Mistral-7B-v0.1` + informativeness-alignment	**66.7**	49.5	**56.8**	**60.2**	34.6	**43.9**	**48.8**	30.9	**37.8**	**55.8**	21.8	**31.4**	**83.6**	74.4	**78.7**

Table 6: Precision (P), Recall (R), and F1-scores for `Mistral-7B-v0.1`. Our informativeness-aligned model achieves the best precision with minor decreases in recall, outperforming previous work.

Model Variant	**Precision**	**Recall**
Full model (ours)	68.7	41.2
FT baseline	47.8	46.8
Structure-Tuning only	60.0	40.2
No informativeness rewarding	56.2	39.7

Table 7: Ablation results disentangling the contribution of informativeness rewarding. We report average factuality precision and recall across all factuality evaluation datasets. Removing informativeness rewarding leads to substantial drops compared to the full model.

of `Llama-3.1-8B`. We find that this variant substantially underperforms the full model, with factuality precision dropping by 4–9 points across datasets, and recall degrading even more severely. These findings confirm that the informativeness training signal is essential not only for informativeness itself but also for achieving improved factuality.

5.5 Error Analysis

To better assess our model's generations, we conducted two experiments: We randomly sampled 200 factual mistakes made by the model, and 40 from each of the factual evaluation datasets. We then compared the responses of our model with the original model. Our findings are as follows:

1. For 84% of the questions, the original model generated a wrong answer, which suggests that most of the mistakes are very likely due to the lack of parametric knowledge.
2. For 3%, the model abstained from answering, but our automatic mechanism did not recognize it.
3. For the remaining 13%, our models' answers were significantly longer (in terms of words). We assume that the model might have learned a spurious correlation from the proposed training that longer answers are more informative.

6 Related Work

Informativeness Evaluating the informativeness of dialogue agents and LLMs is a crucial research goal because of their widespread use. Some studies focus on evaluating the informativeness of dialogue model responses (Freitas et al., 2020; Thoppilan et al., 2022; Lu et al., 2023), and some on factual sentence completions (Huang et al., 2022). More recent work proposes a benchmark for evaluating the factual granularity of model responses (Yona et al., 2024), which we use in this work for both training and evaluation. External tools are also used for better informativeness and correctness (Schick et al., 2023). In this work, we take a model alignment approach to tackle these open challenges.

Factuality Factuality has been widely studied lately from various different perspectives (Augenstein et al., 2023). One way to approach it is via

the task of factual error detection, where a binary prediction model is provided instead of a continuous probability. This is also related to the setting of selective prediction, where models can abstain from answering a query (Varshney et al., 2022; Kamath et al., 2020). Another approach is to adopt model calibration (Guo et al., 2017). The goal is to provide a measure of the probability that a prediction is incorrect alongside the actual prediction. Common techniques for calibration include performing various transformations on a model's output logits (Desai and Durrett, 2020; Jiang et al., 2021) and measuring uncertainty (e.g., see Kuhn et al., 2023). More recent work has studied the use of LMs for providing calibration by training them on statements known to be factually correct or not. This "supervised" approach has been explored via fine-tuning (Kadavath et al., 2022; Lin et al., 2022), in-context learning (Cohen et al., 2023a; Alivanistos et al., 2022), zero-shot instruction-oriented (Cohen et al., 2023b; Dhuliawala et al., 2023; Feng et al., 2024), and consistency sampling (Yoran et al., 2023) techniques. Further recent studies (Azaria and Mitchell, 2023) use the internal state of the model for classifying whether it is certain or not, use a new token for unanswerable inputs (Lu et al., 2022) or for uncertainty representation (Cohen et al., 2025), or construct a specific dataset for effectively tuning the model for answering refusal (Zhang et al., 2024). Our work addresses the factuality problem by aligning an LLM for better informativeness and correctness simultaneously.

7 Conclusion

We propose a novel method *InFACT*, which aims to improve both the correctness and informativeness of LLM's responses. This mechanism takes advantage of factual questions that can be correctly answered at various levels of informativeness and aligns the LLM to exhibit a preference for more informative, yet still correct answers.

An in-depth evaluation across diverse QA benchmarks suggests that this mechanism upgrades both the factual precision of the model's answers, by more effectively abstaining rather than generating wrong facts, as well as the informativeness of its answers, by generating a larger amount of correct answers (higher recall of multiple-answer questions), including answers that are more descriptive and of more appropriate granularity.

This work has the potential to facilitate several intriguing follow-up studies. One of them is the curation of a new unified and qualitative dataset for informativeness evaluation, which may have the potential to further improve the factual consistency of the LLMs.

Limitations

We note a few limitations of our method. First, it depends on the availability of sufficient labeled data. For this data, there must be some notion of informativeness, as our method requires an informativeness-based hierarchy of labels for each input example, as discussed in Section 2.2. For certain domains, this could be hard to obtain. In general, however, such data can be procured from numerous sources. When existing datasets lack explicit answer hierarchies, they can be extended using external knowledge sources such as KGs or crowd annotations. This highlights the flexibility of our framework, enabling the transformation of flat QA datasets into more informative training resources.

Second, as discussed in Section 5.5, the model might learn undesirable spurious correlations through the proposed alignment process, such as with the answer length, as the goal of the method is to teach the model to extract the most informative answer. This can be mitigated with careful tuning.

Third, the design of our method was motivated by the assumption that one would rather obtain fully correct answers only. In this setting, it may occur that the model generates answers that are partially correct, but we then teach it to abstain instead. Thus, the method design as well as the evaluation might need to be customized to tailor to the needs in specific application setups or domains.

References

Lukas Aichberger, Kajetan Schweighofer, Mykyta Ielanskyi, and Sepp Hochreiter. 2024. Semantically diverse language generation for uncertainty estimation in language models. *arXiv preprint arXiv:2406.04306*.

Dimitrios Alivanistos, Selene Báez Santamaría, Michael Cochez, Jan-Christoph Kalo, Emile van Krieken, and Thiviyan Thanapalasingam. 2022. Prompting as probing: Using language models for knowledge base construction. *arXiv preprint arXiv:2208.11057*.

Samuel Amouyal, Tomer Wolfson, Ohad Rubin, Ori Yoran, Jonathan Herzig, and Jonathan Berant. 2023.

QAMPARI: A benchmark for open-domain questions with many answers. In *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 97–110, Singapore. Association for Computational Linguistics.

Isabelle Augenstein, Timothy Baldwin, Meeyoung Cha, Tanmoy Chakraborty, Giovanni Luca Ciampaglia, David Corney, Renee DiResta, Emilio Ferrara, Scott Hale, Alon Halevy, et al. 2023. Factuality challenges in the era of large language models. *arXiv preprint arXiv:2310.05189*.

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

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenhang Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, K. Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Yu Bowen, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xing Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *ArXiv*, abs/2309.16609.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2019. Piqa: Reasoning about physical commonsense in natural language. In *AAAI Conference on Artificial Intelligence*.

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

Roi Cohen, Konstantin Dobler, Eden Biran, and Gerard de Melo. 2025. I don't know: Explicit modeling of uncertainty with an [idk] token. *Advances in Neural Information Processing Systems*, 37:10935–10958.

Roi Cohen, Mor Geva, Jonathan Berant, and Amir Globerson. 2023a. Crawling the internal knowledge-base of language models. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1856–1869, Dubrovnik, Croatia. Association for Computational Linguistics.

Roi Cohen, May Hamri, Mor Geva, and Amir Globerson. 2023b. LM vs LM: Detecting factual errors via cross examination. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12621–12640, Singapore. Association for Computational Linguistics.

Shrey Desai and Greg Durrett. 2020. Calibration of pre-trained transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 295–302, Online. Association for Computational Linguistics.

Ashwin Devaraj, William Sheffield, Byron Wallace, and Junyi Jessy Li. 2022. Evaluating factuality in text simplification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7331–7345, Dublin, Ireland. Association for Computational Linguistics.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2023. Chain-of-verification reduces hallucination in large language models. *arXiv preprint arXiv:2309.11495*.

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

Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. 2024. Don't hallucinate, abstain: Identifying llm knowledge gaps via multi-llm collaboration. *ArXiv*, abs/2402.00367.

Daniel De Freitas, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. 2020. Towards a human-like open-domain chatbot. *ArXiv*, abs/2001.09977.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR.

Jie Huang, Kevin Chen-Chuan Chang, Jinjun Xiong, and Wen mei W. Hwu. 2022. Can language models be specific? how? In *Annual Meeting of the Association for Computational Linguistics*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2024. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*.

Albert Qiaochu Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L'elio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *ArXiv*, abs/2310.06825.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.

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

Jean Kaddour, Joshua Harris, Maximilian Mozes, Herbie Bradley, Roberta Raileanu, and Robert McHardy. 2023. Challenges and applications of large language models. *arXiv preprint arXiv:2307.10169*.

Amita Kamath, Robin Jia, and Percy Liang. 2020. Selective question answering under domain shift. *arXiv preprint arXiv:2006.09462*.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur P. Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc V. Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words. *arXiv preprint arXiv:2205.14334*.

Stephanie C. Lin, Jacob Hilton, and Owain Evans. 2021. TruthfulQA: Measuring how models mimic human falsehoods. In *Annual Meeting of the Association for Computational Linguistics*.

Hongyuan Lu, Wai Lam, Hong Cheng, and Helen Meng. 2022. On controlling fallback responses for grounded dialogue generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2591–2601, Dublin, Ireland. Association for Computational Linguistics.

Hua Lu, Siqi Bao, Huang He, Fan Wang, Hua Wu, and Haifeng Wang. 2023. Towards boosting the open-domain chatbot with human feedback. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4060–4078, Toronto, Canada. Association for Computational Linguistics.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Hannaneh Hajishirzi, and Daniel Khashabi. 2022. When not to trust language models: Investigating effectiveness and limitations of parametric and non-parametric memories. *arXiv preprint arXiv:2212.10511*.

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

Jeff Z. Pan, Simon Razniewski, Jan-Christoph Kalo, Sneha Singhania, Jiaoyan Chen, Stefan Dietze, Hajira Jabeen, Janna Omeliyanenko, Wen Zhang, Matteo Lissandrini, Russa Biswas, Gerard de Melo, Angela Bonifati, Edlira Vakaj, Mauro Dragoni, and Damien Graux. 2023. Large Language Models and Knowledge Graphs: Opportunities and Challenges. *Transactions on Graph Data and Knowledge*, 1(1):2:1–2:38.

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

Daking Rai, Yilun Zhou, Shi Feng, Abulhair Saparov, and Ziyu Yao. 2024. A practical review of mechanistic interpretability for transformer-based language models. *ArXiv*, abs/2407.02646.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? *arXiv preprint arXiv:2002.08910*.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *arXiv preprint arXiv:2302.04761*.

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

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Neeraj Varshney, Swaroop Mishra, and Chitta Baral. 2022. Investigating selective prediction approaches across several tasks in IID, OOD, and adversarial settings. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1995–2002, Dublin, Ireland. Association for Computational Linguistics.

Blerta Veseli, Sneha Singhania, Simon Razniewski, and Gerhard Weikum. 2023. Evaluating language models for knowledge base completion. *ArXiv*, abs/2303.11082.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Gal Yona, Roee Aharoni, and Mor Geva. 2024. Narrowing the knowledge evaluation gap: Open-domain question answering with multi-granularity answers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6737–6751, Bangkok, Thailand. Association for Computational Linguistics.

Ori Yoran, Tomer Wolfson, Ben Bogin, Uri Katz, Daniel Deutch, and Jonathan Berant. 2023. Answering questions by meta-reasoning over multiple chains of thought. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5942–5966, Singapore. Association for Computational Linguistics.

Hiyori Yoshikawa and Naoaki Okazaki. 2023. Selective-LAMA: Selective prediction for confidence-aware evaluation of language models. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2017–2028, Dubrovnik, Croatia. Association for Computational Linguistics.

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

Victor Zhong, Weijia Shi, Wen tau Yih, and Luke Zettlemoyer. 2022. RoMQA: A benchmark for robust, multi-evidence, multi-answer question answering. In *Conference on Empirical Methods in Natural Language Processing*.

A Additional Details

A.1 Task and Data Details

Figure 2 clarifies the different types of question–answer scenarios discussed in our paper. Table 8 provides two examples from our training dataset.

A.2 Experimental Details

Training Details We use a maximum learning rate of 2×10^{-5} with a linear warmup for 10% of the training steps and a cosine decay down to 2×10^{-6}. We use a batch size of 256, weight decay of 0.05, gradient clipping of 1.0, and AdamW β values (0.9, 0.95). We train for 256 optimization steps. Regarding infrastructure, we use 4 NVIDIA A100 40G GPUs. We use the same parameters and constraints for both stages - structure-tuning and informativeness-alignment.

Additional Results Table 9 provides additional experimental results.

Type of Questions	Questions	ChatGPT	Knowledge Graph
Multiple Answer Questions	What are the awards received by Barack Obama?	• Nobel Prize • Presidential Medal of Freedom with Distinction • Order of the Aztec Eagle • Profile in Courage Award • Grammy Award for Best Spoken Word Album • NAACP Image Award • NAACP Image Award for Outstanding Literary Work • Time Person of the Year • John F. Kennedy Profile in Courage Award • German Media Award	• Nobel Peace Prize • Time Person of the Year • Profile in Courage Award • Order of the Rajamitrabhorn • National Medal of Science • Romy • German Media Award • Order of Sikatuna • Financial Times Person of the Year Grammy Award for Best Audio Book, Narration & Storytelling Recording • Presidential Medal of Distinction ....
Descriptive Question Answering	What is the location of the capital of Australia?	The capital of Australia is located in the southeastern part of the country, within the Australian Capital Territory (ACT). It is situated approximately 150 kms inland from the east coast, between the two major cities of Sydney and Melbourne, surrounded by mountains and bushland.	Canberra
Granularity Based Question Answering	Where was Barack Obama born?	Barack Obama was born in Honolulu, Hawaii, USA, on August 4, 1961.	Kapiolani Medical Center for Women and Children Makiki Honolulu USA

Figure 2: Different Types of Question Answers

Example (a): *Where was Luke Prokopec born?*

Level	Answer	Description
A1	Blackwood	Specific town (most fine-grained answer)
A2	Caerphilly County Borough	Local administrative district
A3	Wales	Country within the UK
A4	United Kingdom	Sovereign state (most general)

Example (b): *Where are newspapers owned at some point in time by Voice Media Group published?*

Level	Answer(s)	Description
A4	{Houston, Dallas, Palm Beach, Phoenix, Denver}	Individual cities (fine-grained locations)
A3	{"Houston and Dallas", "Phoenix and Denver", ...}	Pairs of cities (small groupings)
A2	{"Dallas, Phoenix and Denver", ...}	Medium-sized groupings of cities
A1	{"Houston, Dallas, Palm Beach, Phoenix and Denver"}	All cities grouped as a single set

Table 8: Hierarchical answer representations for two different questions, from specific (A1) to more abstract (A4/A5).

	TriviaQA			PopQA			TruthfulQA			Natural Questions			PIQA		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
`Qwen2.5-7B`	52.7	52.7	52.7	30.8	30.8	30.8	30.2	30.2	30.2	25.3	25.3	25.3	78.6	78.6	78.6
`Qwen2.5-7B` + prompting	54.8	50.5	52.5	45.5	30.4	36.4	42.1	27.9	33.5	44.5	19.7	27.3	80.4	66.7	72.9
`Qwen2.5-7B` + ICL	48.6	53.7	51.0	49.2	25.8	33.8	27.0	30.1	28.5	45.9	22.2	29.9	72.4	79.9	76.0
`Qwen2.5-7B` + FT	52.0	55.9	53.8	31.8	34.2	33.0	35.1	34.5	34.8	30.3	29.9	30.1	75.8	79.9	**77.8**
`Qwen2.5-7B` + informativeness-alignment	**60.9**	51.2	**55.6**	**54.8**	30.1	**38.9**	**49.1**	27.6	**35.3**	**48.7**	24.0	**32.2**	**80.5**	72.7	76.4

Table 9: Precision (P), Recall (R), and F1-scores for `Qwen2.5-7B`. Our informativeness-aligned model achieves the best precision with minor decreases in recall, outperforming previous work.