FACT-BENCH: A Holistic Evaluation of LLMs on Factual Knowledge Recall

Anonymous ACL submission

Abstract

We assess LLMs' ability to recall factual knowledge acquired during pretraining, and investigate the factors that influence this capability. To that end, we construct FACT-BENCH, a benchmark designed with three key attributes. First, FACT-BENCH consists of questions with simple, unambiguous answers that remain stable over time, leading to reliable and easy evaluation. Second, it covers 20 domains, 134 property types, various answer types and knowledge popularity levels. Third, FACT-BENCH is programmatically extensible to cover additional factual knowledge of interest from Wikipedia without human annotation.

We evaluate 24 models across six families, focusing on three aspects of factual knowledge recall. First, we find that instruction-tuning consistently impairs knowledge recall: models trained only with pretraining outperform their instruction-tuned counterparts. Second, we examine the impact of in-context exemplars using counterfactual demonstrations. These exemplars significantly degrade factual recall, particularly when they contradict knowledge the model already possesses. By further decoupling model known and unknown knowledge–that is, whether the model can correctly recall a fact–we find the degradation is attributed to exemplars that contradict a model's known knowledge, as well as the number of such exemplars. Third, we fine-tune Llama-3.1-8B under varying conditions of known and unknown knowledge. Fine-tuning on known knowledge proves consistently more effective than fine-tuning on unknown or mixed knowledge. We will make our benchmark publicly available.

1 Introduction

Recent advancements of large language models (LLMs), exemplified by ChatGPT[1], GPT-4 (OpenAI, 2023), are leading to their widespread adoption in various domains. Despite their remarkable performance, they are still plagued by the issue of hallucinations (Ji et al., 2023). Therefore, it is important to conduct holistic assessments to learn how well LLMs capture factual knowledge and what are the factors that affect their ability to recall knowledge learned from pretraining. Previous factuality benchmarks created from knowledge bases (Mallen et al., 2023; Yu et al., 2023) focus on a few domains and property types, and questions are created from templates with limited patterns (Sun et al., 2023). Evaluation of LLMs on these benchmarks reveal a large gap from mastery of factual knowledge. However, it is unclear whether such gap is caused by design challenges, such as ambiguity of the questions and presence of multiple plausible answers, which could lead to biased results.

In this work, we introduce FACT-BENCH, a comprehensive factuality benchmark consisting of 20K question-answer (QA) pairs and featuring four characteristics: (1) *Simplicity*: we create simple questions from Wikidata triplets (subject, property, object) using Claude [2], to elicit knowledge from LLMs. (2) *Validity*: To make sure the answers are grounded, we select triplets whose subject has a Wikipedia article and whose object also appears in the same article. (3) *Diversity*: FACT-BENCH covers 20 domains, 134 property types, and 3 answer types (entities, dates and numbers). (4) *Specificity*: we manually select property types that are highly likely to yield unique answers and perform prompt engineering to generate specific questions.

We benchmark 24 models across six model families on FACT-BENCH. Our results reveal that instruction-tuning hurts knowledge recall, as pretraining-only models consistently outperform their instruction-tuned counterparts. We observe positive effects of model scaling — for all model families, larger models outperform smaller ones

[1] https://platform.openai.com/docs/models

[2] https://www.anthropic.com/index/introducing-claude. Specifically, we use `claude-v1.3-100k` to generate questions.

across all metrics. However, the best performance from GPT-4 still represents a large gap with the upper-bound. To identify where the gap lies, we conduct evaluation from multiple perspectives and find that LLMs struggle with long-tail entities and certain property types.

In addition, we perform counterfactual in-context learning (ICL) experiments to examine the role of in-context exemplars. Our results indicate that counterfactual exemplars lead to significant degradation of factual knowledge recall for large models. By further decoupling model known and unknown knowledge, we find the degradation is attributed to exemplars that contradict a model's *known* knowledge, as well as the number of such exemplars. Lastly, we fine-tune Llama-3.1-8B in different settings of *known* and *unknown* knowledge. In particular, fine-tuning on knowledge that is *known* to the model is beneficial, and consistently outperforms fine-tuning on knowledge that is *unknown*, which shows that fine-tuning on unknown knowledge teaches the model to hallucinate.

Our contributions include: (1) A comprehensive benchmark to evaluate LLMs' ability to recall factual knowledge learned from pretraining. (2) Holistic assessment of the strengths and weaknesses of 24 LLMs, and the factors that affect their recall of factual knowledge. (3) Counterfactual ICL experiments to study the role of in-context exemplars, where we find contradicting a model's known knowledge leads to significant degradation of knowledge recall, as well as the number of such exemplars. (4) Fine-tuning experiments that show the advantage of using known knowledge over mixed and unknown knowledge.

2 FACT-Bench

2.1 Dataset Construction

We formulate the factuality evaluation task as closed-book question answering (Roberts et al., 2020), where a question is fed to the model without any context, and the model needs to leverage its parametric knowledge to answer the question. As simple as the setup is, we identify four challenges: (1) How to make the questions simple enough so that it solely requires knowledge recall rather than complex reasoning or multi-source information? (2) What types of questions are *fair* to ask? It is unfair to query knowledge that does not exist in the pretraining data of all LLMs. (3) How to make the questions diverse and representative? (4) How to make the question specific enough so that the answer is unique and grounded in some knowledge source? We address these challenges from the following four aspects.

Simplicity. Although LLMs have shown remarkable performance for solving composite questions (Wei et al., 2022; Zhou et al., 2023), we aim to decouple the ability to reason and to recall factual knowledge. Therefore, we focus on a simple QA setting to elicit knowledge from LLMs and build up the questions based on sampled Wikidata triplets[3]. The knowledge in Wikidata is in the format of (subject, property, object) triplets, where a simple question can be asked for the property of the subject, and the answer would be the object.

Validity. To benchmark model performance consistently, we take care to ensure that the questions in FACT-BENCH are answerable using content likely available in the models' pretraining data. While the exact pretraining corpora of some large language models (LLMs) are not publicly disclosed, it is reasonable to assume that all include Wikipedia. We select only those knowledge triplets whose object entities also appear on the same Wikipedia page [4] as the subject.

Since LLMs may be trained on different versions of Wikipedia, and factual content can change over time, we further mitigate this issue by manually selecting 134 properties that are highly stable and unlikely to vary across Wikipedia versions.

Diversity. We diversify FACT-BENCH from five aspects: (1) *Multi-domain*. We leverage the knowledge domain categories from Freebase (Bollacker et al., 2008) and select triplets whose subject has a Wikipedia article page, as well as a Freebase ID. We manually aggregate the 99 top-level domains from Freebase into 20 general domains, such as finance, travel, and literature. (2) *Multi-answer-type*. Unlike previous work, we not only include questions with textual answers, but also dates and numbers. (3) *Multi-property-type*. We manually select a total of 134 diverse properties, which is much more comprehensive than previous benchmarks. The full list of property types by answer type can be found in Appendix D. (4) *Multi-knowledge-popularity*. Following previous work (Mallen et al., 2023), we use the view count of subject Wikipedia article from the whole year of

[3]We use the dump from https://dumps.wikimedia.org/wikidatawiki/20230601/.

[4]We use the 20220301.en subset from the Hugging Face datasets library: https://huggingface.co/datasets/wikipedia.

2021 to approximate the popularity of knowledge and sample triplets from the top-25% and bottom-25% most popular triplets sets within each domain. (5) *Diverse questions*. Previous benchmarks typically use templates to construct questions from triplets, whereas we leverage a LLM to generate syntactically rich questions.

Specificity. A challenging issue for the open-domain QA task is that multiple plausible answers may exist for certain questions. We tackle this challenge from two levels. First, select *proper* triplets. For example, the triplet [*Örjan Sandred, student of, Sven-David Sandström*] may not be a good triplet since there could be multiple teachers for everyone, whereas the triplet [*Jacob Viner, doctoral advisor, F. W. Taussig*] is more restricted. We manually select property types that are highly likely to yield unique answers. Second, ask *specific* questions. Given a proper triplet, there could be multiple ways to ask questions. For example, given [*Dan Wickline, place of birth, "Norwalk, California"*], the question *"where was Dan Wickline born?"* has multiple valid answers such as *Norwalk*, and *California*, even though the place of birth is unique for everyone. The question *"What city and state was Dan Wickline born in?"* is more specific. We test multiple prompts for question generation and select one that works best for us (prompt shown in Table 6). Additionally, we filter out triplets whose subjects contain "()" in their Wikipedia titles as "()" is used for disambiguation[5]. We also remove triplets that share the same subject and property. Lastly, for specific numerical answers, we check the number together with the unit. For example, for length, we check for 500 kilometers or 500 km instead of just 500, and for temperature, we check for 98 °C instead of just 98.

2.2 Dataset Statistics and Evaluation Metrics

We select 90 properties with textual answers, 22 properties with date answers, and 22 properties with numerical answers. We randomly sample 1000 triplets from each of the 20 domains, where 500 are from the top-25% most popular triplets, and 500 from the bottom-25%. The resulting 20k QA pairs are split into training and evaluation set, with a size of 5K and 15K, respectively. The 5K training set is released to facilitate exemplar sampling for ICL and small-scale finetuning. We keep the distribution consistent for any subset, i.e., there is an equal number of examples from each domain, out of which half comes from the top-25% and the other half from the bottom-25%.

For evaluation, we use standard metrics for QA tasks, such as SQuAD (Rajpurkar et al., 2016): **Exact Match** (EM) and **F1 score**. For answers that are entities, we collect their aliases from Wikidata as additional ground-truth answers. Dates are normalized in the format of *month, day, year*. In zero-shot experiments, we observe models that have not been instruction-tuned tend to generate verbose answers, which leads to low EM and F1 scores. Therefore, we use an additional metric LLM-as-a-Judge [6] (**LaaJ**) (Zheng et al., 2023) with a grading prompt as shown in Table 12.

2.3 Dataset Validation

We provide a solid estimation of the upper-bound through a collaboration of human and LLM validation to validate that FACT-BENCH is of high quality from the triplet sampling and question generation efforts.

Concretely, we sample a 2k subset from the 15k evaluation set while keeping the distribution of questions consistent, and manually check the validity and specificity of the questions by examining supporting evidence from Wikipedia articles. We identify 201 questions from the 2k subset that are either ambiguous or not supported by Wikipedia, and replace them with valid ones.

In addition, we construct a reading comprehension task in which GPT-4 answers each question using only the corresponding Wikipedia evidence. GPT-4 achieves an accuracy of 92.55% LaaJ score on this task. Details are shown in Appendix A.1. Together, these results suggest that the upper-bound is 90% for the 15k set and 100% for the curated 2k subset, which we denote as PREMIUM2K.

3 Benchmarking LLMs

3.1 Experimental Setup

We consider LLMs with different architectures, sizes, pretraining-only/instruction-tuning, and conduct zero-shot and few-shot ICL experiments. We benchmark GPT-4o, GPT-4o-mini[7], Llama-3.1/Llama-3.1-Instruct (8B, 70B) (Grattafiori et al., 2024), Gemma-2/Gemma-2-it (9B, 27B) (Team et al., 2024), Yi-1.5/Yi-1.5-Chat (6B, 9B, 34B)

[5] https://en.wikipedia.org/wiki/Wikipedia:Article_titles#Disambiguation

[6] We use claude-3-sonnet-20240229 to avoid grader bias.

[7] We access the APIs of OpenAI models gpt-4o/gpt-4o-mini from the week of January 6th to that of July 10th, 2025.

Models	0-shot		1-shot		5-shot		10-shot		Recite 1-shot	
	EM	LaaJ	EM	LaaJ	EM	LaaJ	EM	LaaJ	EM	LaaJ
GPT-4o	51.10	72.85	61.10	71.50	62.75	73.50	64.20	73.00	0.00	75.50
GPT-4o-mini	39.20	49.40	41.15	49.25	40.95	48.90	41.55	48.70	0.00	52.40
Llama-3.1-8B	19.75	47.30	39.95	45.45	43.35	47.50	43.15	47.60	0.00	48.75
Llama-3.1-8B-Instruct	31.45	40.60	30.50	37.80	32.45	37.95	32.50	37.65	0.00	42.90
Llama-3.1-70B	30.10	63.75	57.90	64.15	61.50	66.85	62.30	67.65	0.00	64.65
Llama-3.1-70B-Instruct	52.95	61.25	51.85	58.50	53.10	59.60	53.10	59.25	0.00	64.00
Gemma-2-9B	31.65	42.15	38.75	43.30	41.05	45.40	40.80	45.35	0.00	44.90
Gemma-2-9B-it	36.70	42.95	36.20	41.85	36.95	42.70	36.90	42.80	0.00	42.90
Gemma-2-27B	33.45	42.90	42.85	48.40	45.60	50.30	45.35	50.20	0.00	49.20
Gemma-2-27B-it	42.35	49.20	41.65	48.10	43.00	49.05	43.10	48.70	0.00	48.20
Yi-1.5-9B	30.85	35.60	33.10	36.30	33.75	36.65	33.80	36.55	0.00	33.70
Yi-1.5-9B-Chat	19.20	26.80	21.90	25.60	22.85	26.50	23.20	26.25	0.00	27.10
Yi-1.5-34B	30.05	41.75	37.90	42.00	39.00	42.75	39.00	42.65	0.00	40.35
Yi-1.5-34B-Chat	16.75	40.65	28.15	38.75	24.95	38.85	27.25	38.90	0.00	40.60
Qwen-2.5-7B	28.50	34.00	30.45	34.60	31.30	35.20	31.80	35.55	0.00	36.65
Qwen-2.5-7B-Instruct	27.00	31.55	26.65	31.05	27.25	30.85	26.60	30.35	0.00	33.25
Qwen-2.5-32B	15.55	43.35	38.10	43.35	37.15	42.30	38.40	43.15	0.00	42.40
Qwen-2.5-32B-Instruct	33.60	38.80	34.25	38.90	33.95	38.80	34.35	38.65	0.00	41.50
Qwen-2.5-72B	41.50	48.50	44.50	49.80	43.00	49.30	45.10	49.70	0.00	50.20
Qwen-2.5-72BInstruct	40.30	46.05	41.75	46.95	42.30	47.05	42.55	47.20	0.00	49.10
Deepseek-V2-Lite	23.85	38.95	35.60	40.05	37.65	41.25	37.70	41.40	0.00	40.95
Deepseek-V2-Lite-Chat	23.65	37.85	28.50	34.45	29.50	36.85	31.00	37.35	0.00	37.50
Deepseek-V2	34.80	58.85	52.75	60.35	58.40	63.80	58.70	64.05	0.00	61.60
Deepseek-V2-Chat	10.05	59.65	0.95	59.20	15.65	61.50	41.10	62.15	0.00	61.55

Table 1: Benchmarking results on PREMIUM2K. Comparing instruction-tuned and their pretraining-only counterpart pairs shown in the same color, instruction-tuned models consistently underperform.

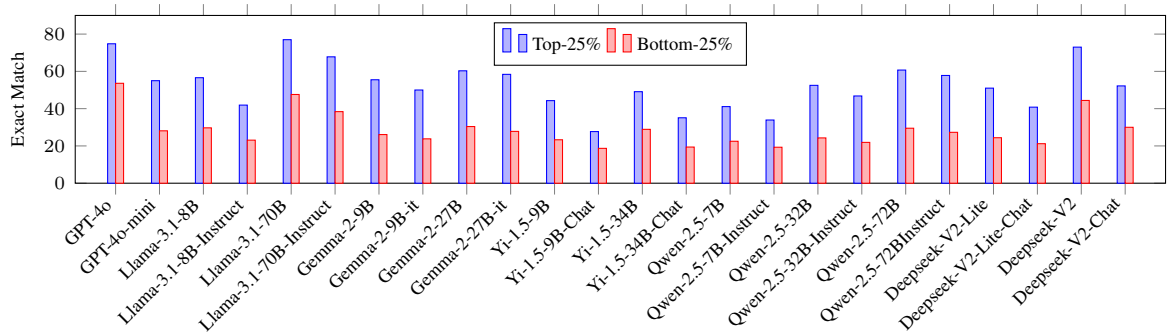

Figure 1: 10-shot EM by knowledge popularity. Knowledge popularity is a strong predictor of knowledge recall. LLMs struggle with long-tail entities (Bottom-25%) as shown by the large gap with popular entities (Top-25%).

(AI et al., 2025), Qwen2.5/Qwen2.5-Instruct (7B, 32B, 72B) (Qwen et al., 2025), and Deepseek-V2-Lite/Deepseek-V2 (DeepSeek-AI et al., 2024). For all LLMs, we use the same prompts shown in Table 8 and 9. The exemplars in the few-shot experiments are shared across models and are randomly sampled from the training set, considering coverage for all 3 answer types (entities, dates and numbers). All our experiments are conducted on the PREMIUM2K subset to reduce the cost of running LLMs with temperature setting up to 0.

3.2 Results

Benchmarking results are presented in Table 1[8].

[8]Full results including F1 scores can be found in Appendix C.3.

Large gap with upper-bound. GPT-4o outperforms the other models. However, its performance of 64.20% EM and 73% LaaJ accuracy in the 10-shot setting still represent a large gap with the upper-bound, which shows the challenge of mastering factuality, as well as the potential risks of using LLMs in certain tasks.

Positive effect of model scaling. We observe positive effects of model scaling. For all families (i.e., GPT, Llama, Gemma, Yi, Qwen and Deepseek), larger model sizes translate to better performances across settings. Closed-source GPT models significantly outperform other open-source models except for Llama-3.1-70B, which has smaller gap with GPT-4o in the 10-shot setting.

Negative impact of instruction-tuning. Com-

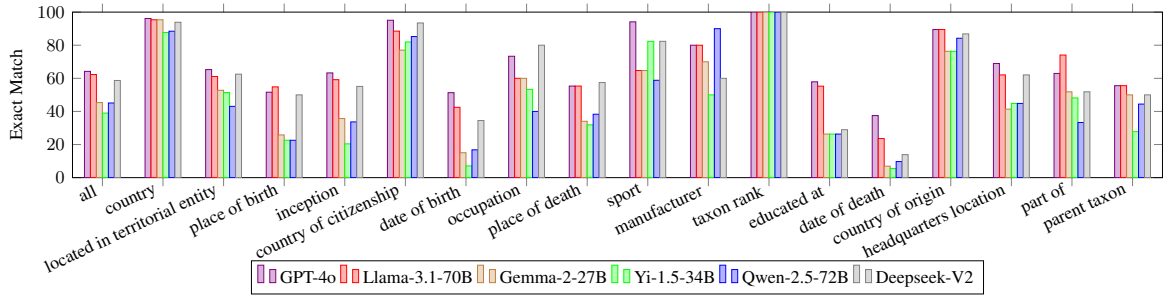

Figure 2: 10-shot EM by property type. LLMs do well on certain property types, such as country-related properties, while struggle on other property types, such as date-related properties. Due to space, we show results for GPT and LLaMA models, and the most common property types from the full set of 134 property types.

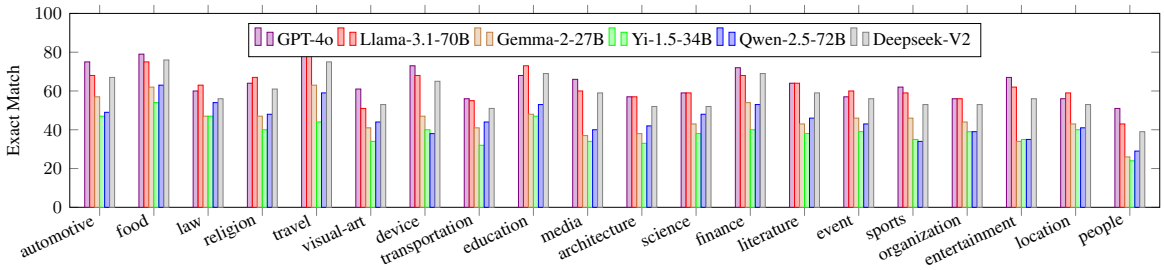

Figure 3: 10-shot EM by domain. Compared to knowledge popularity and property type, domain is less predictive of knowledge recall as model performances across different domains are more flat.

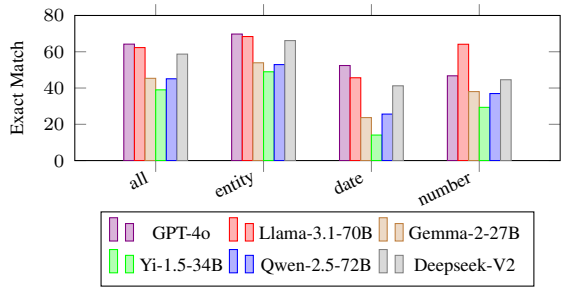

Figure 4: 10-shot EM by answer type. LLMs are less capable on date and numerical knowledge.

paring models in their pretraining-only form and their instruction-tuned counterparts in the few-shot setting, all instruction-tuned models display inferior performance for all metrics. In the zero-shot setting, pretraining-only models tend to generate verbose answers, leading to low EM scores. This empirically verifies the hypothesis that most LLM knowledge is learned during pretraining, and alignment only helps with output style and format. We hypothesize that the *alignment tax* (Ouyang et al., 2022) from instruction-tuning leads to the performance drop. Overall, the best performance for each model family is achieved by few-shot ICL with the pretraining-only version of the model. One way to mitigate such tax is to have the model recite relevant knowledge before giving the answer. Compared to simple 1-shot, recite 1-shot greatly reduce the performance gap for instruction-tuning models. Table 10 shows the recite 1-shot prompt.

Diminishing returns from adding more exemplars. Going from zero-shot to 1-shot, all open-source models benefit learning from the answer format of the in-context exemplar, which is reflected in their improved EM scores. This is especially the case for pretraining-only models. By the LaaJ metric, the results are mixed. As k increases from 0 to 10, most models don't show significant changes.

3.3 Fine-grained Evaluation

To gain a better understanding of where the gap with the upper-bound lies, we examine model performances from multiple perspectives.

Knowledge popularity and property type are predictive of knowledge recall. Figure 1 shows 10-shot performance by knowledge popularity and Figure 2 by property type. We observe that knowledge popularity and property type are strong predictors of knowledge recall. LLMs struggle with long-tail entities (Bottom-25%) as shown by the large gap with popular entities (Top-25%). This result suggests that knowledge distribution of the *pretraining data* (if known to the model user) can potentially be leveraged as a predictor for factual knowledge recall. LLMs do well on certain prop-

erty types, such as country-related properties, while struggle on other property types, such as date-related properties. Further results by answer type (Figure 4) show that LLMs are less capable on date and numerical knowledge.

Domain is less predictive of knowledge recall. On the other hand, domain is not a strong predictor of model performance as shown in Figure 3, where model performances across different domains are more flat compared to knowledge popularity levels and property types.

4 The Role of In-context Exemplars

Previous work (Min et al., 2022) suggests that ground-truth labels play an insignificant role for ICL, such that replacing ground-truth labels with random labels on classification and multi-choice tasks only results in marginal loss of accuracy. Compared to classification and multi-choice tasks, the label space of our task is much larger. We design a set of experiments to investigate how *counterfactual* in-context exemplars affect a model's ability to recall factual knowledge.

4.1 Counterfactual ICL

Experimental setup. In this set of experiments, we replace the ground-truth answers of our regular 10-shot exemplars with random answers chosen from the 5k training set. We impose an additional constraint that the random answer is chosen within the same property type, denoted as **shuffle**. For example, we change the ground-truth answer for *"In which military branch did Henry Curtis serve?"* from *"Royal Navy"* to the counterfactual answer *"United States Marine Corps"*. Without prior knowledge required to answer the question, the new input-label pair looks reasonable but is actually not factual.

Results. Figure 5 shows the results. Notably, LLAMA-3.1-70B experiences a major drop from 62.3% EM (regular 10-shot) to 32.65%, followed by Gemma-2-27B from 45.35% to 16.70%, Qwen-2.5-72B from 48.9% to 43.2%, and Deepseek-V2 from 58.7% to 36.65%. These models have an average 25.6% drop, whereas most *smaller* models from each family (LLAMA-3.1-8B, Qwen-2.5-7B, Deepseek-V2-Lite) have a much smaller drop. In addition, we observe that all instruction-tuned models are less affected by counterfactual exemplars, indicating that they are less capable than their pretraining-only counterparts at overriding semantic priors.

4.2 Counterfactual ICL with *known* and *unknown* Knowledge

Results in the previous section show that counterfactual exemplars lead to significant degradation of factual knowledge recall for pretraining-only models. However, it is not clear what factors lead to this behavior besides model scale and instruction tuning. We further decouple *known* and *unknown* knowledge of the exemplars to study their role. Intuitively, if the model have no knowledge regarding the exemplars, it should make no difference to the model whether the answers are shuffled or not.

Experimental setup. We conducted controlled experiments with the largest model from each family. To approximate model known and unknown knowledge, we sample $k = 30$ questions that are correctly answered by each model as **known** knowledge, and $k = 30$ incorrectly answered as **unknown** knowledge. We corrupt the exemplars with the same shuffling method as the previous experiment.

Contradicting LLMs' known knowledge teaches them to lie. Results are shown in Figure 6. Comparing *known-shuffle* with *unknown-shuffle* in the 10-shot setting, the performance drops significantly for all pretraining-only models (Llama-3.1-70B from 61.95% to 34.30%, Gemma-2-27B from 46.10% to 11.60%, Yi-1.5-34B from 40.45% to 19.25%, Qwen-2.5-72B 46.8% to 13.4% and Deepseek-V2 from 58.25% to 19.3%) with known-shuffle while the drop with unknown-shuffle is much less significant with an average 1.49%. As we increase k, the gap between known- and unknown-shuffle becomes increasingly deep. The average drop for these five models is 31.14% for 10-shot, 40.39% for 20-shot and 41.77% for 30-shot. Although instruction-tuned counterparts for these models show robustness against counterfactual exemplars as shown in Figure 5 (an average of 0.73% drop), known-shuffle exemplars still significant impact the factual knowledge recall performance, with an average drop of 2.5% for 10-shot, 5.58% for 20-shot, and 9.01% for 30-shot.

The results suggest that the degradation in factual knowledge recall is primarily due to exemplars that contradict models' known knowledge, i.e., counterfactual ICL with known knowledge is essentially teaching LLMs to lie, leading to unexpected results. Additionally, the number of counterfactual exemplars also plays a prominent role. As k increases, models experience sharper drops. In practical applications, it is therefore important to

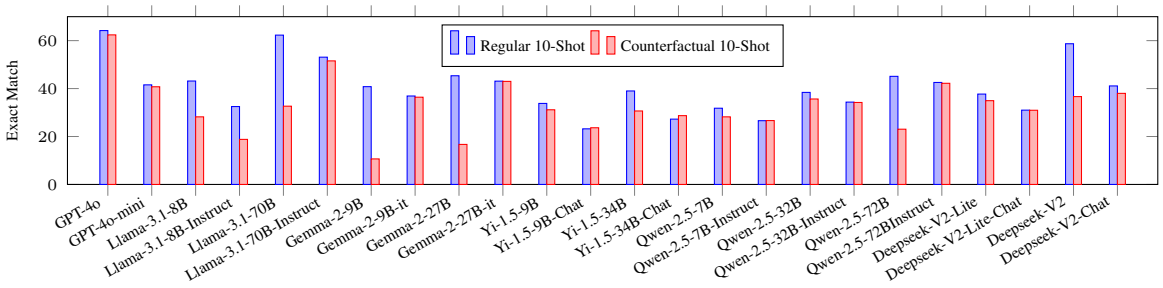

Figure 5: Comparison of regular 10-shot and counterfactual 10-shot by Exact Match.

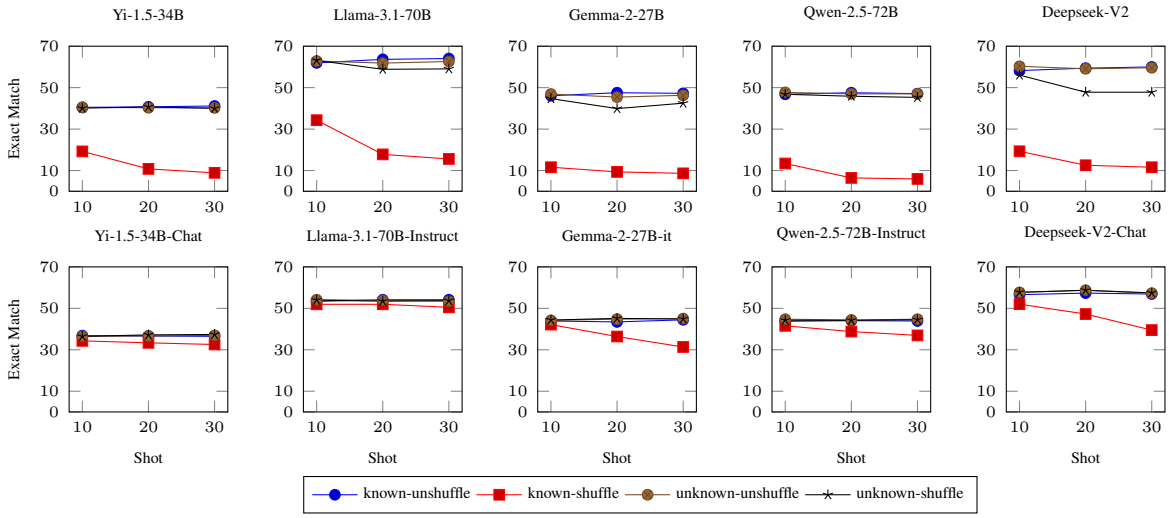

Figure 6: Counterfactual few-shot with known and unknown knowledge, evaluated by Exact Match. Result shows that the degradation in factual knowledge recall is primarily due to exemplars that contradict models' *known* knowledge, and the number of such exemplars.

pair in-context exemplars with the correct answers if known to the model, in order to maximally elicit their parametric knowledge. Finally, we observe comparable performances for known-unshuffle and unknown-unshuffle across different models.

5 Fine-tuning

In this section, we examine how fine-tuning affects a model's ability to recall factual knowledge and use Llama-3.1-8B to conduct experiments.

5.1 Regular Fine-tuning

Experimental setup. We fine-tune Llama-3.1-8B on the 5k training set and sample 4k additional examples using the same procedure described in Section 2 as the validation set. We train for 40 steps where training stabilizes based on validation loss and report results on the PREMIUM2K subset. For model input and output, we use the same input-label format as in the prompting experiments (i.e., input consists of an instruction and a question, and output is the answer to the question).

Results. In the zero-shot setting, we compare models using the LaaJ metric instead of EM since the predictions of pretraining-only Llama are verbose. Table 2 shows that our fine-tuning Llama-3.1-8B does not bring extra gains over the pretraining-only Llama. Results of this experiment further verify the hypothesis that a model's knowledge is mostly learned from pretraining, and instruction-tuning only helps align the answer format.

Models	zero-shot		
	EM	F1	LaaJ
Llama-3.1-8B	19.75	33.98	47.30
Llama-3.1-8B-Instruct	31.45	39.29	40.60
Llama-3.1-8B (fine-tuned)	42.50	49.21	47.05

Table 2: Comparison of Llama-3.1-8B, Llama-3.1-8B-Instruct and our fine-tuned Llama-3.1-8B.

5.2 Counterfactual Fine-tuning

Experimental setup. In the counterfactual ICL experiments (Section 4), our experiment results indicate that Llama-3.1-8B is negatively impacted by counterfactual exemplars. We set up similar experiments in the fine-tuning setting, where we corrupt

the training data with inner-property-shuffle.

Results. Table 3 shows the results. Factuality of in-context exemplars plays a critical role for fine-tuning. The model can recover part of its capability as training goes on. However, its performance is still significantly worse than that from regular fine-tuning (32.65% EM vs 42.50%).

Setup for fine-tuning	zero-shot		
	EM	F1	LaaJ
Regular fine-tuning	42.50	49.21	47.05
Counterfactual fine-tuning	32.65	37.01	37.40

Table 3: Fine-tuning Llama-3.1-8B with counterfactual knowledge.

5.3 Fine-tuning with *known*, *unknown* and *mixed* Knowledge

Experimental setup. We fine-tune Llama-3.1-8B with three types of *factual* knowledge separately: (1) *known*. (2) *unknown*. (3) *mixed*. To approximate known and unknown knowledge, we use the same method described in Section 4.2. We use our evaluation set (not including PREMIUM2K) as the candidate pool to select training data since we need to distinguish between known and unknown knowledge, and 5k is insufficient. We then randomly choose 2.5k training examples for known and unknown knowledge, respectively.

Results. Table 4 shows the results. Training with known knowledge outperforms training with mixed knowledge, and training with unknown knowledge leads to the worst performance. The results show that fine-tuning on knowledge unknown to the model teaches the model to hallucinate.

Setup for fine-tuning	zero-shot		
	EM	F1	LaaJ
Known knowledge	44.40	51.51	47.90
Unknown knowledge	40.90	47.28	45.40
Mixed knowledge	42.00	49.34	46.40

Table 4: Fine-tuning Llama-3.1-8B with known, unknown and mixed knowledge.

6 Related Work

Factuality Benchmarks Question answering datasets, such as Natural Questions (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017), WebQuestions (Berant et al., 2013), TruthfulQA (Lin et al., 2022) have been used to evaluate factuality of language models. LAMA (Petroni et al., 2019, 2020) leverages 4 knowledge sources and converts fact triplets into cloze-style questions. More recent works, such as POPQA (Mallen et al., 2023) and KoLA (Yu et al., 2023), construct benchmarks from Wikidata using templates and cover a limited set of property types and domains. Head-to-Tail (Sun et al., 2023) creates their benchmark from DBpedia (Auer et al., 2007) with a focus on evaluating LLM in knowledge at different popularity levels. SimpleQA (Wei et al., 2024) questions are adversarially collected against GPT-4 responses. Compared to previous benchmarks, FACT-BENCH is more diverse and representative, covering 134 property types, 20 general domains and 3 answer types. We strictly filter Wikidata triplets and generate valid and specific questions whose answers are grounded in Wikipedia.

The role of in-context exemplars Min et al. (2022) studies the role of in-context exemplars and shows that ground-truth labels are not required for ICL. Yoo et al. (2022) revisits the findings and proposes additional metrics to reveal the importance of ground-truth labels. Wei et al. (2023) conducts similar experiments and finds that overriding semantic priors is an emergent ability of large models. Our counterfactual ICL experiments corroborate this finding, where large models suffer from significant degradation of knowledge recall. We additionally find that contradicting a model's known knowledge is the primary factor leading to this behavior, along with the number of such exemplars. Pan et al. (2023) separates task recognition from task learning in studying how ICL leverages demonstrations, and find that task recognition does not drastically improve with model scaling and more exemplars, while task learning does.

7 Conclusion

In this paper, we introduce FACT-BENCH, a comprehensive benchmark that focuses on evaluating factual knowledge of LLMs. We conduct experiments on 24 models from six model families and investigate the factors that affect their knowledge recall. We find that instruction-tuning can hurt knowledge recall. In studying the effects of counterfactual in-context exemplars, we highlight the role of known and unknown knowledge. We also conduct fine-tuning experiments, where we highlight the importance of factuality in the training data. We hope that release of our benchmark will be beneficial to the community and help facilitate future research.

8 Limitations

In this work, we strive to benchmark and analyze as many popular LLMs as resource allows. However, due to the fast pace at which models are released and limited resource, we pick representative and available models at the time of our experimentation. Additionally, distinguishing between model known knowledge and unknown knowledge is an ongoing research topic and in Section 4.2 and 5, we check if the model can answer the question correctly as a proxy for model known and unknown knowledge. Finally, our work specifically targets the evaluation of simple factual knowledge, deliberately excluding reasoning. We view factual recall as a foundational capability that underpins more complex forms of question answering. Evaluating how LLMs leverage this factual knowledge in reasoning-intensive tasks is an important and complementary direction, which we leave for future work.

References

01. AI, :, Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Guoyin Wang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yanpeng Li, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. Yi: Open foundation models by 01.ai, 2025. URL https://arxiv.org/abs/2403.04652.

S. Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary G. Ives. Dbpedia: A nucleus for a web of open data. In *ISWC/ASWC*, 2007.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. Semantic parsing on Freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1533–1544, Seattle, Washington, USA, October 2013. Association for Computational Linguistics.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. Freebase: A collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, SIGMOD '08, pp. 1247–1250, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605581026. doi: 10.1145/1376616.1376746.

DeepSeek-AI, Aixin Liu, Bei Feng, Bin Wang, Bingxuan Wang, Bo Liu, Chenggang Zhao, Chengqi Dengr, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Hao Yang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jin Chen, Jingyang Yuan, Junjie Qiu, Junxiao Song, Kai Dong, Kaige Gao, Kang Guan, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruizhe Pan, Runxin Xu, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Size Zheng, T. Wang, Tian Pei, Tian Yuan, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Liu, Xin Xie, Xingkai Yu, Xinnan Song, Xinyi Zhou, Xinyu Yang, Xuan Lu, Xuecheng Su, Y. Wu, Y. K. Li, Y. X. Wei, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Zheng, Yichao Zhang, Yiliang Xiong, Yilong Zhao, Ying He, Ying Tang, Yishi Piao, Yixin Dong, Yixuan Tan, Yiyuan Liu, Yongji Wang, Yongqiang Guo, Yuchen Zhu, Yuduan Wang, Yuheng Zou, Yukun Zha, Yunxian Ma, Yuting Yan, Yuxiang You, Yuxuan Liu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhewen Hao, Zhihong Shao, Zhiniu Wen, Zhipeng Xu, Zhongyu Zhang, Zhuoshu Li, Zihan Wang, Zihui Gu, Zilin Li, and Ziwei Xie. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model, 2024. URL https://arxiv.org/abs/2405.04434.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan

Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto,

Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Xuming Hu, Junzhe Chen, Xiaochuan Li, Yufei Guo, Lijie Wen, Philip S. Yu, and Zhijiang Guo. Towards understanding factual knowledge of large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=9OevMUdods.

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, mar 2023. doi: 10.1145/3571730.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1147.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur 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 Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019. doi: 10.1162/tacl_a_00276.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3214–3252, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.229.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9802–9822, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.546.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11048–11064, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.759.

OpenAI. GPT-4 technical report, 2023.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 27730–27744, New Orleans, November 2022. Curran Associates, Inc.

Jane Pan, Tianyu Gao, Howard Chen, and Danqi Chen. What in-context learning "learns" in-context: Disentangling task recognition and task learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 8298–8319, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.527.

F. Petroni, T. Rocktäschel, A. H. Miller, P. Lewis, A. Bakhtin, Y. Wu, and S. Riedel. Language models as knowledge bases? In *In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2019*, 2019.

Fabio Petroni, Patrick Lewis, Aleksandra Piktus, Tim Rocktäschel, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. How context affects language models' factual predictions. In *Automated Knowledge Base Construction*, 2020.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui

Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1264.

Adam Roberts, Colin Raffel, and Noam Shazeer. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5418–5426, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.437.

Kai Sun, Yifan Ethan Xu, Hanwen Zha, Yue Liu, and Xin Luna Dong. Head-to-tail: How knowledgeable are large language models (LLM)? a.k.a. will LLMs replace knowledge graphs?, 2023.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. Gemma 2: Improving open language models at a practical size, 2024. URL https://arxiv.org/abs/2408.00118.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837. Curran Associates, Inc., 2022.

Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. Measuring short-form factuality in large language models, 2024. URL https://arxiv.org/abs/2411.04368.

Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, and Tengyu Ma. Larger language models do in-context learning differently, 2023.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Online, October 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6.

Kang Min Yoo, Junyeob Kim, Hyuhng Joon Kim, Hyunsoo Cho, Hwiyeol Jo, Sang-Woo Lee, Sang-goo Lee, and Taeuk Kim. Ground-truth labels matter: A deeper look into input-label demonstrations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 2422–2437, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.155.

Jifan Yu, Xiaozhi Wang, Shangqing Tu, Shulin Cao, Daniel Zhang-li, Xin Lv, Hao Peng, Zijun Yao, Xiaohan Zhang, Hanming Li, Chun yan Li, Zheyu Zhang, Yushi Bai, Yan-Tie Liu, Amy Xin, Nianyi Lin, Kaifeng Yun, Linlu Gong, Jianhui Chen, Zhili Wu, Yun Peng Qi, Weikai Li, Yong Guan, Kaisheng Zeng, Ji Qi, Hailong Jin, Jinxi Liu, Yuxian Gu, Yu Gu, Yuan Yao, Ning Ding, Lei Hou, Zhiyuan Liu, Bin Xu, Jie Tang, and Juanzi Li. KoLA: Carefully benchmarking world knowledge of large language models. *ArXiv*, abs/2306.09296, 2023.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=uccHPGDlao.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, and Ed H. Chi. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations*, 2023.

A Benchmark Construction

A.1 Benchmark Validation

For benchmark validation, we have one person and use GPT-4 to perform the same reading comprehension task given supporting evidence derived from Wikipedia. The prompt is shown in Table 7.

To derive the supporting evidence from the Wikipedia, we first extract the corresponding Wikipedia page of the subject. We find the index of the first occurrence of the ground-truth answer, and take 300 preceding and 300 following characters. Of note, this context is not 100% accurate as the ground-truth answer could appear multiple times and the first occurrence may not be the actual location providing the answer, especially for numerical answers. Therefore, this performance can be seen as the lower-bound of the upper-bound.

For human validation, we identify 201 questions from the 2k subset that are either ambiguous or not supported by Wikipedia. GPT-4 achieves an accuracy of 92.55% LaaJ score on this task. Both approaches echo with each other and show that FACT-BENCH is of high quality.

A.2 Comparison with Existing Benchmark

We compare FACT-BENCH with several widely recognized benchmarks, including TriviaQA (Joshi et al., 2017), Natural Questions (Kwiatkowski et al., 2019), TruthfulQA (Lin et al., 2022), and Pinocchio (Hu et al., 2024), as shown in Table 5. A defining characteristic of FACT-BENCH is its explicit focus on disentangling knowledge recall from reasoning. Unlike other benchmarks that often conflate these two abilities, FACT-BENCH is designed to assess factual recall in isolation, without requiring complex reasoning. Furthermore, FACT-BENCH aims to be both time-agnostic and model-invariant: it supports fair and consistent evaluation across current and future language models, independent of their architecture or release date, as the construction of FACT-BENCH only consider facts that are unlikely to change over time.

B Task Instructions

Table 6, 7, 8, 9, 10, 11, 12 show the prompts of question generation, reading comprehension for question validation, zero-shot, 10-shot, 1-shot recitation, counterfactual 10-shot for question answering, and LLM-as-a-Judge for grading, respectively.

C Experimental Settings

C.1 Zero-Shot and Few-Shot ICL Experiments

For open models, we use Python version 3.10, Torch version 2.0.0, and the Hugging Face Transformers library (Wolf et al., 2020) with version 4.31.0. We use greedy decoding for reproducibility. Batch size is set to 4 and sequences are left-padded with [PAD] token set to [EOS] token if it's not already set. All our experiments were conducted on A100 GPUs with 40GB of RAM.

C.2 Fine-tuning Experiments

For fine-tuning Llama-3.1-8B, we use the AdamW optimizer with a learning rate of 0.0001 and a cosine learning rate scheduler. Effective batch size is set to 512 and sequence length to 256. We train for 40 steps, where we observe the training stabilizes based on validation set performance.

Dataset	Construction	Reasoning	Feature	AnswerType
TriviaQA	Questions are from trivia website	Yes	Large-scale reading comprehension dataset, which requires cross-sentence reasoning to find answers	free-form text
Natural Language	User queries submitted to Google	Yes	Find long/short answers from Wikipedia pages for real-life user queries	free-form text
TruthfulQA	Authors write questions	Yes	Questions that some humans would answer falsely due to a false belief or misconception	free-form text
Pinocchio	Human rewrite claims into questions	Yes	Examine multi-facts, structured and unstructured facts, facts that change over time, facts in different languages	free-form text
FACT-BENCH	All questions are grounded by Wikipedia triplets	No	Simple and unambiguous facts recall without context, doesn't change over time.	Wikipedia entities, dates, numbers

Table 5: Comparison with Existing Benchmarks.

Instruction: given a triplet in the form of (subject, property, object), generate a question about the subject.

Requirement:
1. The object must be the unique answer to the question!!
2. The question must be specific regarding the nature or category of the object so that the object is the only answer!! For example, if the answer is a city, ask for city; If the answer is city and state, ask for city and state, and so on.
3. The question should not include the object!!
4. If the property is height or elevation above sea level, ask for meters. If the property is area, ask for square kilometers. If the property is length or width, ask for kilometers. If the answer is temperature, ask for celsius.
5. Your response should strictly follow the format:

Question:<question> Answer: <answer>.
Triplet: triplet

Table 6: Prompt for question generation.

C.3 Full Benchmarking Results on PREMIUM2K

Table 13 shows zero-shot, 1-shot, 6-shot, and 10-shot results on PREMIUM2K for all 24 models we consider.

D List of Properties

Table 14, 15, and 16 show the 134 property types by answer type (date, number, and entity).

Instruction: answer the following question with the context. If the context doesn't contain the answer, use your own knowledge to answer the question. Don't include explanations. Keep the answer as concise as possible.
Question: {QUESTION}
Context: {CONTEXT}
Answer:

Table 7: Prompt of reading comprehension for data validation.

Instruction: answer the following question. Don't include explanation. Keep the answer as concise as possible.
Question: {QUESTION}
Answer:

Table 8: Prompt for zero-shot.

Instruction: answer the following question. Don't include explanation. Keep the answer as concise as possible.
Question: In which military branch did Henry Curtis serve?
Answer: Royal Navy

Question: Which Australian rules football club was Simon Madden a member of?
Answer: Essendon Football Club

Question: What architectural style is Pine Bloom Plantation?
Answer: Greek Revival architecture

Question: On what date was Ed Hooper born?
Answer: March 10, 1964

Question: Who founded Tangerine Dream?
Answer: Edgar Froese

Question: Who is the performer of 'Hollywood's Not America'?
Answer: Ferras

Question: What company manufactured the AMC Gremlin?
Answer: American Motors Corporation

Question: In what year was the Whistler House Museum of Art officially opened?
Answer: 1908

Question: What is Ellen S. Baker's mother's name?
Answer: Claire Shulman

Question: How many floors above ground does Premier Tower have?
Answer: 78

Question: {QUESTION}
Answer:

Table 9: Prompt for few-shot (10-shot).

Instruction: answer the following question. You need to recite knowledge that is relevant to the question, and end your response with the final answer. You must return in the format of "Recitation: <recitation> Answer: <answer>"

Question: Who is the performer of 'Hollywood's Not America'?
Recitation: "Hollywood's Not America" is a song by American pop singer/songwriter Ferras and is featured on his debut studio album, Aliens & Rainbows. It was released on January 29, 2008, as the lead single from that album.
Answer: Ferras

Question: {QUESTION}

Table 10: Prompt for one-shot recitation.

Instruction: answer the following question. Don't include explanation. Keep the answer as concise as possible.
Question: In which military branch did Henry Curtis serve?
Answer: United States Marine Corps

Question: Which Australian rules football club was Simon Madden a member of?
Answer: Tennessee Volunteers football

Question: What architectural style is Pine Bloom Plantation?
Answer: Art Nouveau

Question: On what date was Ed Hooper born?
Answer: October 12, 1876

Question: Who founded Tangerine Dream?
Answer: Frank Varga

Question: Who is the performer of 'Hollywood's Not America'?
Answer: Damien Bodie

Question: What company manufactured the AMC Gremlin?
Answer: Kalem Company

Question: In what year was the Whistler House Museum of Art officially opened?
Answer: 1832

Question: What is Ellen S. Baker's mother's name?
Answer: Empress Dayi

Question: How many floors above ground does Premier Tower have?
Answer: 164

Question: {QUESTION}
Answer:

Table 11: Prompt for counterfactual few-Shot (10-shot).

You are a teacher grading quiz. You will be given a question, a reference answer and a student's response. Check whether the student's response is correct or not regarding the answer. Respond with Yes/No without explanation.

Question: {QUESTION}
Reference Answer: {ANSWER}
Student Response: {RESPONSE}

Table 12: Prompt for LLM-as-a-Judge.

Models	zero-shot			1-shot			5-shot			10-shot		
	EM	F1	LaaJ	EM	F1	LaaJ	EM	F1	LaaJ	EM	F1	LaaJ
gpt-4o	51.10	63.52	72.85	61.10	69.74	71.50	62.75	71.43	73.50	64.20	72.54	73.00
gpt-4o-mini	39.20	48.43	49.40	41.15	49.59	49.25	40.95	49.91	48.90	41.55	50.53	48.70
meta-llama/Llama-3.1-8B	19.75	33.98	47.30	39.95	46.10	45.45	43.35	50.62	47.50	43.15	50.65	47.60
meta-llama/Llama-3.1-8B-Instruct	31.45	39.29	40.60	30.50	36.18	37.80	32.45	38.77	37.95	32.50	38.60	37.65
meta-llama/Llama-3.1-70B	30.10	46.37	63.75	57.90	64.71	64.15	61.50	68.68	66.85	62.30	69.20	67.65
meta-llama/Llama-3.1-70B-Instruct	52.95	61.24	61.25	51.85	59.03	58.50	53.10	60.46	59.60	53.10	60.45	59.25
google/gemma-2-9b	31.65	38.57	42.15	38.75	45.02	43.30	41.05	48.50	45.40	40.80	48.15	45.35
google/gemma-2-9b-it	36.70	44.12	42.95	36.20	42.81	41.85	36.95	44.24	42.70	36.90	43.93	42.80
google/gemma-2-27b	33.45	40.64	42.90	42.85	49.84	48.40	45.60	53.03	50.30	45.35	52.73	50.20
google/gemma-2-27b-it	42.35	50.22	49.20	41.65	49.06	48.10	43.00	50.67	49.05	43.10	50.44	48.70
01-ai/Yi-1.5-9B	30.85	37.23	35.60	33.10	39.05	36.30	33.75	40.51	36.65	33.80	40.36	36.55
01-ai/Yi-1.5-9B-Chat	19.20	27.35	26.80	21.90	29.86	25.60	22.85	30.66	26.50	23.20	30.95	26.25
01-ai/Yi-1.5-34B	30.05	38.81	41.75	37.90	44.67	42.00	39.00	45.68	42.75	39.00	45.65	42.65
01-ai/Yi-1.5-34B-Chat	16.75	29.34	40.65	28.15	38.26	38.75	24.95	39.51	38.85	27.25	40.36	38.90
Qwen/Qwen2.5-7B	28.50	35.79	34.00	30.45	37.36	34.60	31.30	38.62	35.20	31.80	39.14	35.55
Qwen/Qwen2.5-7B-Instruct	27.00	34.60	31.55	26.65	34.02	31.05	27.25	34.90	30.85	26.60	34.22	30.35
Qwen/Qwen2.5-32B	15.55	29.32	43.35	38.10	45.56	43.35	37.15	45.57	42.30	38.40	46.28	43.15
Qwen/Qwen2.5-32B-Instruct	33.60	41.19	38.80	34.25	41.55	38.90	33.95	42.02	38.80	34.35	42.22	38.65
Qwen/Qwen2.5-72B	41.50	49.04	48.50	44.50	51.76	49.80	43.00	51.77	49.30	45.10	52.76	49.70
Qwen/Qwen2.5-72B-Instruct	40.30	48.19	46.05	41.75	49.29	46.95	42.30	49.95	47.05	42.55	50.18	47.20
deepseek-ai/DeepSeek-V2-Lite	23.85	33.47	38.95	35.60	42.32	40.05	37.65	44.72	41.25	37.70	44.59	41.40
deepseek-ai/DeepSeek-V2-Lite-Chat	23.65	32.77	37.85	28.50	35.39	34.45	29.50	38.67	36.85	31.00	39.79	37.35
deepseek-ai/DeepSeek-V2	34.80	47.49	58.85	52.75	60.38	60.35	58.40	65.31	63.80	58.70	65.47	64.05
deepseek-ai/DeepSeek-V2-Chat	10.05	43.59	59.65	0.95	45.55	59.20	15.65	51.60	61.50	41.10	59.32	62.15

Table 13: Full benchmarking results on PREMIUM2K.

- date of birth (P569)
- date of death (P570)
- inception (P571)
- time of discovery or invention (P575)
- publication date (P577)
- first flight (P606)
- UTC date of spacecraft launch (P619)
- UTC date of spacecraft landing (P620)
- date of disappearance (P746)
- date of first performance (P1191)
- date of official opening (P1619)
- production date (P2754)
- date of official closure (P3999)
- recording date (P10135)
- dissolved, abolished or demolished date (P576)
- start time (P580)
- end time (P582)
- service entry (P729)
- service retirement (P730)
- discontinued date (P2669)
- debut date (P10673)*
- date of incorporation (P10786)*

Table 14: List of 22 date properties. The ones with asterisk do not appear in PREMIUM2K.

- maximum capacity (P1083)
- atomic number (P1086)
- total produced (P1092)
- number of cylinders (P1100)
- floors above ground (P1101)
- number of platform tracks (P1103)
- number of episodes (P1113)
- number of deaths (P1120)
- neutron number (P1148)
- minimum number of players (P1872)
- maximum number of players (P1873)
- number of children (P1971)
- length (P2043)
- elevation above sea level (P2044)
- area (P2046)
- height (P2048)
- width (P2049)
- mass (P2067)
- melting point (P2101)
- chromosome count (P5230)
- number of seats in legislature (P1410)*
- memory capacity (P2928)*

Table 15: List of 22 number properties. The ones with asterisk do not appear in PREMIUM2K.

- member of political party (P102)
- taxon rank (P105)
- occupation (P106)
- location of creation (P1071)
- founded by (P112)
- airline hub (P113)
- home venue (P115)
- league (P118)
- place of burial (P119)
- publisher (P123)
- owned by (P127)
- located in the administrative territorial entity (P131)
- participant in (P1344)
- winner (P1346)
- movement (P135)
- genre (P136)
- operator (P137)
- capital of (P1376)
- licensed to broadcast to (P1408)
- IUCN conservation status (P141)
- languages spoken, written or signed (P1412)
- affiliation (P1416)
- present in work (P1441)
- architectural style (P149)
- country for sport (P1532)
- headquarters location (P159)
- transport network (P16)
- producer (P162)
- award received (P166)
- country (P17)
- creator (P170)
- parent taxon (P171)
- performer (P175)
- manufacturer (P176)
- crosses (P177)
- developer (P178)
- endemic to (P183)
- doctoral advisor (P184)
- place of birth (P19)
- collection (P195)
- place of death (P20)
- cuisine (P2012)
- basin country (P205)
- located in or next to body of water (P206)
- father (P22)
- military branch (P241)
- mother (P25)
- record label (P264)
- country of citizenship (P27)
- production company (P272)
- location (P276)
- programmed in (P277)
- designed by (P287)
- vessel class (P289)
- continent (P30)
- operating system (P306)
- capital (P36)
- part of (P361)
- space launch vehicle (P375)
- parent astronomical body (P397)
- mouth of the watercourse (P403)
- position played on team / speciality (P413)
- original broadcaster (P449)
- color (P462)
- occupant (P466)
- animal breed (P4743)
- court (P4884)
- country of origin (P495)
- author (P50)
- cause of death (P509)
- school district (P5353)
- member of sports team (P54)
- director (P57)
- screenwriter (P58)
- conflict (P607)
- discoverer or inventor (P61)
- highest point (P610)
- sport (P641)
- drafted by (P647)
- educated at (P69)
- diocese (P708)
- location of formation (P740)
- parent organization (P749)
- distributed by (P750)
- historic county (P7959)
- country of registry (P8047)
- architect (P84)
- composer (P86)
- filming location (P915)
- allegiance (P945)

Table 16: List of 90 entity properties.