Towards a Holistic Evaluation of LLMs on Factual Knowledge Recall

Jiaqing Yuan[1][*]**, Lin Pan**[2][†]**, Chung-Wei Hang**[2]**, Jiang Guo**[2]**, Jiarong Jiang**[2]**, Bonan Min**[2]**,**
Patrick Ng[2]**, Zhiguo Wang**[2]

[1]North Carolina State University   [2]AWS AI Labs

jyuan23@ncsu.edu

{linpan,cwhang,gujiang,jiarongj,bonanmin,patricng,zhiguow}@amazon.com

Abstract

Large language models (LLMs) have shown remarkable performance on a variety of NLP tasks, and are being rapidly adopted in a wide range of use cases. It is therefore of vital importance to holistically evaluate the factuality of their generated outputs, as hallucinations remain a challenging issue.

In this work, we focus on assessing LLMs' ability to recall factual knowledge learned from pretraining, and the factors that affect this ability. To that end, we construct FACT-BENCH, a representative benchmark covering 20 domains, 134 property types, 3 answer types, and different knowledge popularity levels. We benchmark 31 models from 10 model families and provide a holistic assessment of their strengths and weaknesses. We observe that instruction-tuning hurts knowledge recall, as pretraining-only models consistently outperform their instruction-tuned counterparts, and positive effects of model scaling, as larger models outperform smaller ones for all model families. However, the best performance from GPT-4 still represents a large gap with the upper-bound. We additionally study the role of in-context exemplars using counterfactual demonstrations, which 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-7B in different settings of known and unknown knowledge. In particular, fine-tuning on a model's known knowledge is beneficial, and consistently outperforms fine-tuning on unknown and 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 on NLP tasks, 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.

---

[*]Work conducted during an internship at Amazon
[†]Corresponding author

[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.

We benchmark 31 models across 10 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 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, consistent with the findings in (Mallen et al., 2023) and (Sun et al., 2023).

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-7B 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 empirically verifies the hypothesis in Schulman (2023) 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 31 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., 2023b), 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 performance across various models, we take steps to make sure questions in FACT-BENCH are *answerable* from their pretraining corpora. Although the exact pretraining corpora are not disclosed for some LLMs, it is reasonable to assume that they have all been pretrained on Wikipedia articles. Therefore, we normalize the main content of the Wikipedia page[4] for the subjects, and only select triplets whose objects also appear in the same Wikipedia page.

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 C. (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 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 5). 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 manually 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 re-

sulting 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 fine-tuning. 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 but does not necessarily mean that the prediction is wrong. Therefore, we introduce an additional metric **Contains**, which simply checks if any of the ground-truth answers appear in the prediction.

2.3 Dataset Validation

We provide a solid estimation of the upper-bound through human 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. Empirically, the upper-bound is 90% for the 15k set and 100% for the 2k subset, which we denote as PREMIUM2K.

3 Benchmarking LLMs

3.1 Experimental Setup

We consider LLMs with different model architectures, sizes, pretraining-only/instruction-tuning, and conduct zero-shot and few-shot ICL experiments. Specifically, we benchmark GPT-4, GPT-3.5-turbo[6], BLOOM/BLOOMZ (7B) (Scao et al., 2023), LLaMA (7B, 13B, 33B, 65B) (Touvron et al., 2023), Vicuna (7B, 13B, 33B) (Chiang et al., 2023), OpenLLaMA (7B, 13B) (Geng & Liu, 2023), FLAN-T5-XXL (11B) (Chung et al.,

---

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

[6]We access the APIs of OpenAI models from the week of July 3rd to that of July 17th, 2023
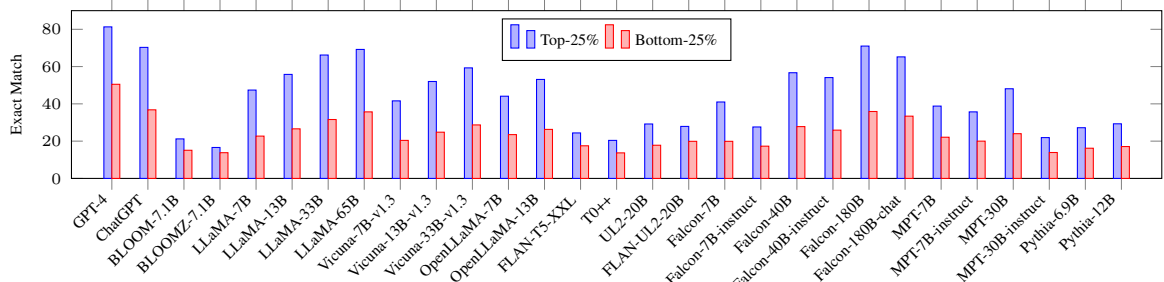
Figure 1: 10-shot EM by knowledge popularity. Knowledge popularity is a strong predictor of knowledge recall. LLMs struggle with long-tail entities (Bottom-25%) as shown by the large gap with popular entities (Top-25%).
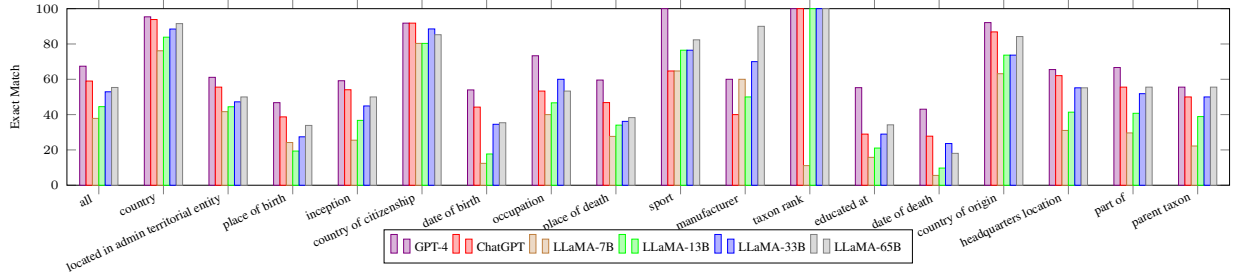

Figure 2: 10-shot EM by property type. LLMs do well on certain property types, such as country-related properties, while struggle on other property types, such as date-related properties. Due to space, we show results for GPT and LLaMA models, and the most common property types from the full set of 134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

Models	zero-shot		1-shot		6-shot		10-shot	
	EM	Contains	EM	Contains	EM	Contains	EM	Contains
GPT-4	58.60	64.65	59.85	63.20	63.35	66.45	65.90	69.15
GPT-3.5-turbo	49.75	52.60	51.25	53.70	52.65	55.80	53.55	56.40
BLOOM-7.1B	03.20	20.30	18.95	19.95	17.85	19.90	18.15	19.75
BLOOMZ-7.1B	18.00	19.45	14.05	15.35	14.40	17.05	15.20	17.70
LLaMA-7B	14.65	35.20	33.25	34.15	35.55	37.15	35.05	36.75
LLaMA-13B	21.35	39.95	36.45	37.30	41.15	42.75	41.20	42.95
LLaMA-33B	27.25	46.55	45.25	46.70	48.30	50.30	48.90	51.10
LLaMA-65B	35.25	49.20	47.15	48.45	52.15	53.80	52.45	54.10
Vicuna-7B-v1.3	24.65	33.25	31.15	33.80	30.10	35.05	31.00	34.65
Vicuna-13B-v1.3	32.95	35.15	36.45	37.60	38.00	41.20	38.40	41.15
Vicuna-33B-v1.3	34.30	44.15	41.39	44.75	44.10	48.10	44.00	48.05
OpenLLaMA-7B	14.05	32.30	31.75	32.80	32.55	34.70	33.80	35.95
OpenLLaMA-13B	25.70	37.35	37.05	38.40	38.75	40.70	39.70	41.55
FLAN-T5-XXL (11B)	20.60	21.60	20.45	21.45	21.05	22.15	20.95	22.00
T0++ (11B)	16.05	21.25	16.75	19.95	16.80	20.00	17.05	19.85
UL2 (20B)	03.40	23.55	23.50	24.40	24.15	25.75	23.50	25.00
FLAN-UL2 (20B)	24.05	25.20	24.10	25.25	24.10	25.30	23.90	24.95
Falcon-7B	23.60	30.05	30.25	31.90	30.70	32.60	30.45	32.25
Falcon-7B-instruct	10.85	25.10	21.75	24.60	22.45	25.45	22.45	25.20
Falcon-40B	26.55	30.90	39.10	40.50	42.05	43.60	42.25	43.80
Falcon-40B-instruct	21.95	40.25	38.85	40.75	40.40	42.20	40.00	41.85
Falcon-180B	44.90	47.45	49.25	50.60	52.15	55.05	53.45	55.00
Falcon-180B-chat	39.95	47.10	47.00	49.30	49.05	51.50	49.30	51.60
MPT-7B	03.45	30.35	28.85	29.85	29.75	31.15	30.45	31.55
MPT-7B-instruct	03.55	30.40	21.55	29.25	26.35	29.30	27.85	29.60
MPT-30B	25.30	35.00	34.35	35.55	35.80	37.55	36.05	37.75
MPT-30B-instruct	19.05	33.50	28.80	31.20	31.00	33.50	31.50	33.85
Pythia-6.9B	11.00	13.15	21.20	22.45	21.85	23.05	21.70	23.25
Pythia-12B	15.25	22.00	22.75	23.70	22.95	24.35	23.20	24.65
Mistral-7B	28.45	29.25	38.90	39.80	40.45	41.85	40.75	42.60
Mistral-7B-instruct	26.00	29.30	26.80	30.05	26.80	30.35	27.20	30.75

Table 1: Benchmarking results on PREMIUM2K.

2022), T0++ (11B) (Sanh et al., 2021), UL2/FLAN-UL2 (20B) (Tay et al., 2023), Falcon/Falcon-instruct (7B, 40B, 180B) (Almazrouei et al., 2023), MPT/MPT-instruct (7B, 30B) (MosaicML NLP Team, 2023), Pythia (6.9B, 12B) (Biderman et al.,

2023), and Mistral/Mistral-instruct (7B) (Jiang et al., 2023). For all LLMs, we use the same prompts shown in Table 6 and 7. 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[7].

3.2 Results

Benchmarking results are presented in Table 1[8].

Large gap with upper-bound. GPT-4 outperforms all the other models we consider on our benchmark. However, its performance of 65.9% EM in the 10-shot setting still represents a large gap with the estimated 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. Overall, we observe positive effects of model scaling. For all model families (i.e., LLaMA, Falcon, and MPT), larger model sizes translate to better performances across settings. Closed-source GPT models significantly outperform open-source models with the

---

[7]For the full 15k evaluation set, we provide zero-shot and 10-shot results in Appendix B.4 for reference.

[8]Full results including F1 scores can be found in Appendix B.3.
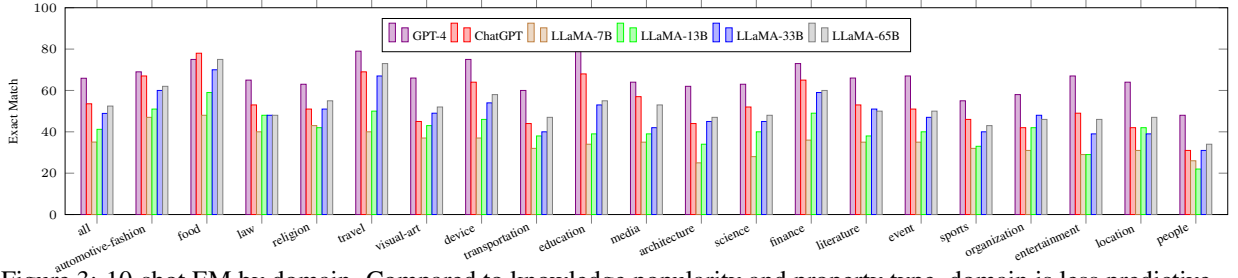
Figure 3: 10-shot EM by domain. Compared to knowledge popularity and property type, domain is less predictive of knowledge recall as model performances across different domains are more flat. Due to space, we show results for GPT and LLaMA models.
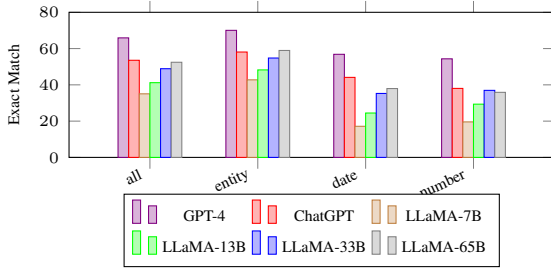

Figure 4: 10-shot EM by answer type. LLMs are less capable on date and numerical knowledge. Due to space, we show results for GPT and LLaMA models.

notable exception of LLaMA-65B, which is competitive with GPT-3.5-turbo in the 10-shot setting.

Negative impact of instruction-tuning. Comparing models in their pretraining-only form and their instruction-tuned counterparts, such as LLaMA/Vicuna, Falcon/Falcon-instruct, and MPT/MPT-instruct 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, which leads to low EM and F1 scores, but the *Contains* metric reveals that they outperform their instruction-tuned counterparts. This result empirically verifies the hypothesis in Zhou et al. (2023a) that most knowledge of LLMs 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.

Diminishing returns from adding more exemplars. Going from zero-shot to 1-shot, all open-source models benefit greatly learning from the answer format of the in-context exemplar, which is reflected in their improved EM scores. This is espe-

cially the case for pretraining-only models. By the Contains metric, results are mixed. As *k* increases to 6, all models, with the exception of BLOOM and T0++, show improvements over zero-shot and 1-shot. However, going from 6-shot to 10-shot, model performances mostly stay flat, except for GPT-4, improving by 2.7%. To further validate this, we run zero, 1–10 shots with LLaMA models, and results are shown in Figure 5, where the curves flatten after providing 3–5 exemplars.
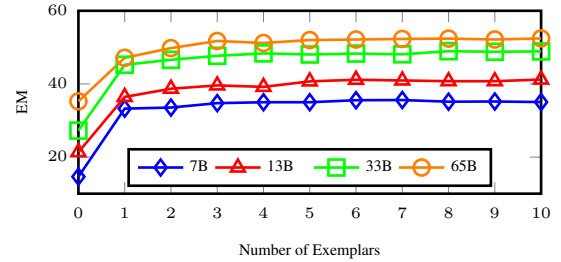

Figure 5: LLaMA zero-to-10-shot results by EM.

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 similar findings in Mallen et al. (2023) 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 property 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 6 shows the results. Notably, LLaMA-65B experiences a major drop from 52.45% EM (regular 10-shot) to 29.45%, followed by Falcon-180B from 53.45% to 37.05%, and LLaMA-33B from 48.9% to 43.2%, while the performance of *smaller* models remains flat and unaffected. In addition, we observe that instruction-tuned models are less affected by counterfactual exemplars than their pretraining-only counterparts. For example, compared to LLaMA, Falcon and MPT, the drop is less significant for Vicuna, Falcon-chat, and MPT-instruct models, respectively.

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 large models. However, it is not clear what factors lead to this behavior besides model scale. LLaMA-65B, Falcon-180B and LLaMA-33B are the three most capable open-source models on our benchmark (Table 1). Since the 10 in-context exemplars are randomly sampled, it is expected that these three models have more knowledge about the exemplars than the other models. Therefore, we further decouple *known* and *unknown* knowledge of the exemplars to study their role.

Experimental setup. We conduct controlled experiments on LLaMA models and `text-davinci-002`[9]. To approximate model known and unknown knowledge, we sample $k = 32$ questions that are correctly answered by each model as **known** knowledge, and $k = 32$ 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 7. Comparing *known-shuffle* with *unknown-shuffle* in the 10-shot setting, LLaMA-65B drops from 52.45% EM (regular 10-shot) to 26.60% with known-shuffle while the drop with unknown-shuffle is much less significant from 52.45% EM to 42.90%. For LLaMA-33B, performance drops from 48.30% to 42.20% with known-shuffle, and from 48.30% to 46.25% with unknown-shuffle. For the larger text-davinci-002 model, performances are near identical with known-shuffle and unknown-shuffle (41.60% vs 41.30%). However, as we increase k, the gap between known- and unknown-shuffle becomes increasingly deep (i.e., 34.10% vs 42.80% in 20-shot, and 29.95% vs 40.55% in 32-shot). Similar effect from increasing k is observed on LLaMA-33B and LLaMA-65B. Notably, as k increases, the smaller LLaMA-13B also starts experiencing sharp drops with known-shuffle. In the 32-shot setting, its performance drops from 41.29% (regular 10-shot) to 24.45%, while remains flat with unknown-shuffle at 40.80%. For the smallest LLaMA-7B, its performances stay flat across

---

[9]In Wei et al. (2023), experiments using in-context exemplars with flipped labels show that `text-davinci-002` experiences the largest drop on binary classification tasks. We further include this model in this set of experiments.
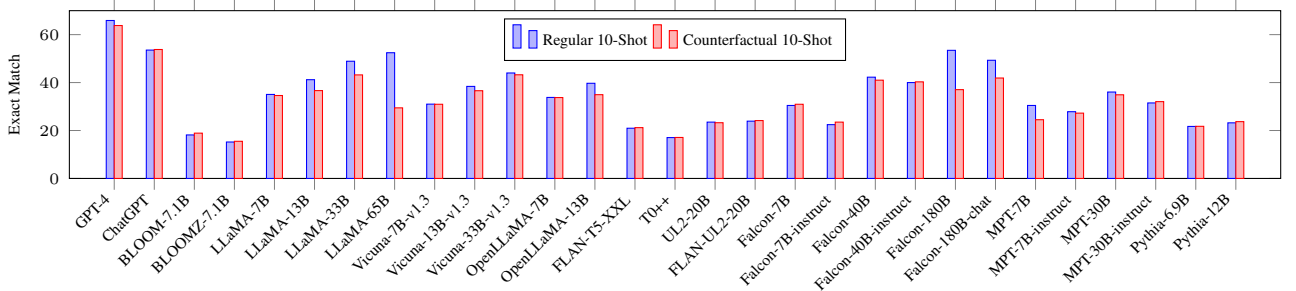
Figure 6: Comparison of regular 10-shot and counterfactual 10-shot by Exact Match. LLaMA-65B experiences a major drop with counterfactual exemplars, followed by Falcon-180B and LLaMA-33B, while the performance of smaller models remains flat and unaffected.
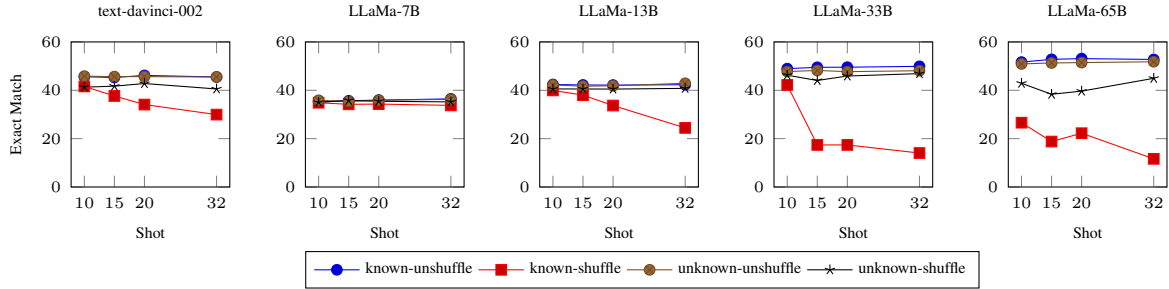

Figure 7: Counterfactual few-shot with known and unknown knowledge, evaluated by Exact Match. Result shows that the degradation in factual knowledge recall is primarily due to exemplars that contradict models' *known* knowledge, and the number of such exemplars.

different settings.

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 and even smaller models (LLaMA-13B in our experiments) can suffer from significant performance drops. In practical applications, it is therefore important to 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-7B to conduct experiments.

5.1 Regular fine-tuning

Experimental setup. We fine-tune LLaMA-7B 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 *contains* metric instead of EM since the predictions of pretraining-only LLaMA are verbose. Table 2 shows that our fine-tuned LLaMA underperforms Vicuna, and both models underperform the pretraining-only LLaMA. Results of this experiment further verify the hypothesis in Zhou et al. (2023a) that a model's knowledge is mostly learned from pretraining, and instruction-tuning only helps align the answer format.

Models	zero-shot		
	EM	F1	Contains
LLaMA-7B	14.65	27.66	35.20
Vicuna-7B	24.65	33.33	33.25
LLaMA-7B (fine-tuned)	28.75	35.22	29.85

Table 2: Comparison of LLaMA, Vicuna and our fine-tuned LLaMA.

5.2 Counterfactual fine-tuning

Experimental setup. In the counterfactual ICL experiments (Section 4), our experiment results indicate that LLaMA-7B is mostly unaffected 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 (11.25% EM vs 29.1%).

Setup for fine-tuning	zero-shot		
	EM	F1	Contains
Regular fine-tuning	28.75	35.22	29.85
Counterfactual fine-tuning	10.75	15.61	12.45

Table 3: Fine-tuning LLaMA-7B with counterfactual knowledge.

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

Experimental setup. We fine-tune LLaMA-7B 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 consistently outperforms training with mixed knowledge, and training with unknown knowledge leads to the worst performance. The results verify the claim in Schulman (2023) that fine-tuning on knowledge unknown to the model teaches the model to hallucinate.

Setup for fine-tuning	zero-shot		
	EM	F1	Contains
Known knowledge	33.00	39.54	33.85
Unknown knowledge	27.55	34.10	28.75
Mixed knowledge	29.30	36.36	30.25

Table 4: Fine-tuning LLaMA-7B 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 LLMs on knowledge at different popularity levels. 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 31 models from 10 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.

References

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

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.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A suite for analyzing large language models across training and scaling, 2023.

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.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing GPT-4 with 90%* chatgpt quality, March 2023.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models, 2022.

Xinyang Geng and Hao Liu. OpenLLaMA: An open reproduction of LLaMA, May 2023.

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.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023.

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.

MosaicML NLP Team. Introducing MPT-7B: A new standard for open-source, commercially usable llms, 2023. Accessed: 2023-05-05.

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.

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.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Stella Biderman, Leo Gao, Tali Bers, Thomas Wolf, and Alexander M. Rush. Multitask prompted training enables zero-shot task generalization, 2021.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, and Colin Raffel. BLOOM: A 176b-parameter open-access multilingual language model, 2023.

John Schulman. Reinforcement learning from human feedback: Progress and challenges. The Berkeley AI lectures, April 2023. URL https://news.berkeley.edu/2023/04/24/berkeley-talks-chatgpt-developer-john-schulman.

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.

Yi Tay, Mostafa Dehghani, Vinh Q. Tran, Xavier Garcia, Jason Wei, Xuezhi Wang, Hyung Won Chung, Siamak Shakeri, Dara Bahri, Tal Schuster, Huaixiu Steven Zheng, Denny Zhou, Neil Houlsby, and Donald Metzler. UL2: Unifying language learning paradigms, 2023.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and efficient foundation language models, 2023.

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.

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.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. LIMA: Less is more for alignment, 2023a.

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*, 2023b.

A   Task Instructions

Table 5, 6, 7, and 8 show the prompts for question generation, zero-shot, 10-shot, and counterfactual 10-shot, respectively.

B   Experimental Settings

B.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.

B.2   Fine-tuning Experiments

For fine-tuning LLaMA-7B, 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.

B.3   Full Benchmarking Results on PREMIUM2K

Table 9 shows zero-shot, 1-shot, 6-shot, and 10-shot results on PREMIUM2K for all 31 models we consider.

B.4   Full Benchmarking Results on 15k Evaluation Set

Table 10 shows zero-shot and 10-shot results on the full 15k evaluation set for all 31 models we consider.

C   List of Properties

Table 11, 12, and 13 show the 134 property types by answer type (date, number, and entity).

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 5: Prompt for question generation.

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

Table 6: 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 7: Prompt for few-shot (10-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: 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 8: Prompt for counterfactual few-Shot (10-shot).

Models	zero-shot			1-shot			6-shot			10-shot		
	EM	F1	Contains	EM	F1	Contains	EM	F1	Contains	EM	F1	Contains
GPT-4	58.60	65.99	64.65	59.85	65.7	63.2	63.35	69.17	66.45	65.90	72.01	69.15
GPT-3.5-turbo	49.75	55.77	52.60	51.25	57.41	53.70	52.65	59.36	55.80	53.55	60.19	56.40
BLOOM-7.1B	3.20	14.31	20.30	18.95	24.90	19.95	17.85	24.98	19.90	18.15	24.76	19.75
BLOOMZ-7.1B	18.00	24.08	19.45	14.05	18.02	15.35	14.40	19.65	17.05	15.20	20.49	17.70
LLaMA-7B	14.65	27.66	35.20	33.25	39.73	34.15	35.55	42.88	37.15	35.05	41.96	36.75
LLaMA-13B	21.35	33.92	39.95	36.45	42.81	37.30	41.15	47.98	42.75	41.20	48.48	42.95
LLaMA-33B	27.25	40.48	46.55	45.25	51.70	46.70	48.30	55.03	50.30	48.90	55.65	51.10
LLaMA-65B	35.25	46.67	49.20	47.15	53.33	48.45	52.15	58.44	53.80	52.45	58.98	54.10
Vicuna-7B-v1.3	24.65	33.33	33.25	31.15	38.30	33.80	30.10	39.17	35.05	31.00	39.23	34.65
Vicuna-13B-v1.3	32.95	39.74	35.15	36.45	42.63	37.60	38.00	46.14	41.20	38.40	45.79	41.15
Vicuna-33B-v1.3	34.30	43.94	44.15	41.39	49.38	44.75	44.10	52.36	48.10	44.00	52.44	48.05
OpenLLaMA-7B	14.05	26.62	32.30	31.75	38.03	32.80	32.55	40.19	34.70	33.80	41.23	35.95
OpenLLaMA-13B	25.70	35.98	37.35	37.05	43.21	38.40	38.75	45.98	40.70	39.70	47.07	41.55
FLAN-T5-XXL (11B)	20.60	27.63	21.60	20.45	27.31	21.45	21.05	27.93	22.15	20.95	27.87	22.00
T0++ (11B)	16.05	24.46	21.25	16.75	23.40	19.95	16.80	24.43	20.00	17.05	24.64	19.85
UL2 (20B)	3.40	13.83	23.55	23.50	28.91	24.40	24.15	30.92	25.75	23.50	30.41	25.00
FLAN-UL2 (20B)	24.05	30.58	25.20	24.10	30.64	25.25	24.10	30.47	25.30	23.90	30.26	24.95
Falcon-7B	23.60	31.03	30.05	30.25	36.35	31.90	30.70	37.19	32.60	30.45	37.19	32.25
Falcon-7B-instruct	10.85	20.59	25.10	21.75	29.25	24.60	22.45	30.51	25.45	22.45	30.10	25.20
Falcon-40B	26.55	31.81	30.90	39.10	45.68	40.50	42.05	48.70	43.60	42.25	48.95	43.80
Falcon-40B-instruct	21.95	32.98	40.25	38.85	45.42	40.75	40.40	47.04	42.20	40.00	47.14	41.85
Falcon-180B	44.90	50.34	47.45	49.25	55.80	50.60	53.55	59.89	55.05	53.45	59.94	55.00
Falcon-180B-chat	39.95	48.14	47.10	47.00	54.14	49.30	49.05	56.08	51.50	49.30	56.17	51.60
MPT-7B	3.45	14.89	30.35	28.85	34.72	29.85	29.75	36.00	31.15	30.45	36.51	31.55
MPT-7B-instruct	3.55	11.34	30.40	21.55	27.84	29.25	26.35	32.95	29.30	27.85	34.31	29.60
MPT-30B	25.30	34.18	35.00	34.35	40.73	35.55	35.80	42.98	37.55	36.05	43.03	37.75
MPT-30B-instruct	19.05	28.41	33.50	28.80	35.70	31.20	31.00	38.06	33.50	31.50	38.30	33.85
Pythia-6.9B	11.00	14.11	13.15	21.20	27.06	22.45	21.85	28.41	23.05	21.70	28.32	23.25
Pythia-12B	15.25	22.98	22.00	22.75	28.71	23.70	22.95	29.44	24.35	23.20	29.73	24.65
Mistral-7B	28.45	31.34	29.25	38.90	45.38	39.8	40.45	47.43	41.85	40.75	47.96	42.60
Mistral-7B-instruct	26.00	32.97	29.30	26.80	34.68	30.05	26.80	35.06	30.35	27.20	35.43	30.75

Table 9: Full benchmarking results on PREMIUM2K.

Models	Zero-shot			10-shot		
	EM	F1	Contains	EM	F1	Contains
GPT-4	52.31	59.83	58.47	58.30	64.72	61.39
GPT-3.5-turbo	44.36	50.21	47.19	48.60	55.21	51.55
BLOOM-7.1B	3.37	14.09	18.87	17.05	23.58	18.57
BLOOMZ-7.1B	16.54	22.78	17.96	14.57	20.10	17.01
LLaMA-7B	12.53	25.16	32.69	32.24	39.16	33.99
LLaMA-13B	18.53	31.25	36.79	37.33	44.30	39.16
LLaMA-33B	24.94	37.51	43.29	44.31	51.05	46.21
LLaMA-65B	31.19	42.55	45.62	47.39	53.67	49.02
Vicuna-7B-v1.3	22.46	31.07	31.09	28.14	36.72	32.18
Vicuna-13B-v1.3	29.81	36.17	32.42	34.21	42.09	37.45
Vicuna-33B-v1.3	30.23	40.07	40.69	38.69	47.06	42.97
OpenLLaMA-7B	12.06	24.47	30.79	31.11	38.28	33.12
OpenLLaMA-13B	23.06	33.31	34.45	35.19	42.21	37.03
FLAN-T5-XXL	19.07	25.88	20.42	19.32	26.28	20.73
T0++	14.59	22.97	19.45	15.39	23.54	18.47
UL2	3.37	13.48	22.05	21.85	28.72	23.29
FLAN-UL2	21.16	27.97	22.62	21.03	27.86	22.53
Falcon-7B	21.29	28.84	27.50	28.00	34.95	29.75
Falcon-7B-instruct	9.50	19.14	23.15	19.62	27.34	22.81
Falcon-40B	23.41	28.56	28.57	37.82	44.64	39.37
Falcon-40B-instruct	18.79	29.79	36.87	36.53	43.14	38.32
MPT-7B	2.95	14.23	28.04	27.85	34.39	29.13
MPT-7B-instruct	3.72	11.31	28.39	25.89	32.55	27.61
MPT-30B	22.69	31.58	31.87	33.59	40.21	35.15
MPT-30B-instruct	16.51	26.01	30.29	29.37	36.35	31.53
Pythia-6.9B	9.26	12.55	11.93	19.83	26.39	21.43
Pythia-12B	13.81	21.48	20.19	21.40	28.17	22.88

Table 10: Full benchmarking results on 15K evaluation set.

- 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 11: 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 12: 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 13: List of 90 entity properties.