LEVERAGING LARGE LANGUAGE MODELS FOR MULTIPLE CHOICE QUESTION ANSWERING

Anonymous authors
Paper under double-blind review

ABSTRACT

While large language models (LLMs) like GPT-3 have achieved impressive results on multiple choice question answering (MCQA) tasks in the zero, one, and few-shot settings, they generally lag behind the MCQA state of the art (SOTA). MCQA tasks have traditionally been presented to LLMs like cloze tasks. An LLM is conditioned on a question (without the associated answer options) and its chosen option is the one assigned the highest probability after normalization (for length, etc.). A more natural prompting approach is to present the question and answer options to the LLM jointly and have it output the symbol (e.g., "A") associated with its chosen answer option. This approach allows the model to explicitly compare answer options, reduces computational costs, and mitigates the effects of tokenization scheme and answer option representations on answer selection. For the natural approach to be effective the LLM it is used with must be able to associate answer options with the symbols that represent them. The LLM needs what we term multiple choice symbol binding (MCSB) ability. This ability varies greatly by model. We show that a model with high MCSB ability performs much better with the natural approach than with the traditional approach across 20 diverse datasets and largely closes the gap with the SOTA, suggesting that the MCQA ability of LLMs has been previously underestimated.

1 INTRODUCTION

Current state of the art (SOTA) methods on many multiple choice question answering (MCQA) tasks involve specialized models, prompt engineering, and individualized tuning in general. What if one model could do just as well as each of these models does individually?

This is part of a general vision for so-called foundation models (Bommasani et al., 2021). Foundation models include large pre-trained language models (LLMs) that have derived enough broad knowledge (spanning, for example, linguistic, factual, and commonsense (Liu et al., 2019; Amrami & Goldberg, 2018; Petroni et al., 2020; Bosselut et al.; Bouraoui et al.; Zuo et al., 2018; Bhagavatula et al., 2019)) to transfer from a simple language modelling objective to a huge array of natural language tasks.

Interestingly, while LLMs have achieved SOTA results on many tasks, they generally fall short on MCQA. Why is this the case, given their general language modelling prowess as suggested by the low cross-entropy loss they attain with all their parameters, data, and compute (Kaplan et al., 2020; Henighan et al., 2020; Hernandez et al., 2021)? Should they not excel, or at least be highly competitive?

In this paper, we argue that they fall short because dominant methods used with them conflate *probabilities of sentences* with *probabilities of correct answers*. We hypothesize that there are fundamental problems with the near-universal approach to MCQA for LLMs, which we refer to as "cloze prompting" (CP). Specifically, these problems include 1) the conflation of the grammaticality, commonality, and "naturalness" of a text and its likelihood qua question-answer, 2) the computational expense of scoring multiple candidate answers, 3) the fact that the LLM cannot explicitly reason about and compare different candidate answers, and 4) finicky normalization due to tokenization schemes. The centerpiece of our paper is an extensive investigation of an alternative: we explain how these problems might be solved by what we call *multiple choice prompting* (MCP). In MCP,

the language model receives both the question and also a list of candidate answers as on a multiple choice test, with each answer associated with (or "bound" to) a symbol such as "A", "B", "C", etc. We explain how this approach might be why MCP outperforms CP in Section 3.

More importantly, though, we demonstrate that when we prompt LLMs with MCP instead of CP, performance often dramatically improves – approaching or even surpassing SOTA performance. On a varied group of 20 datasets, we show that MCP outperforms CP on all but 4 of the datasets, with a mean gap of 9.7% on all tasks and a max gap of 44%. MCP surpasses old SOTA scores on 9 of 20 datasets (by as much as 15% on a single task), and averaged across all datasets, MCP scores fall 0.6% shy of SOTA.

This implies that the de facto method for prompting LLMs has led them to be considerably underestimated for MCQA, and that there exists a better general way to prompt a single LLM that scores within a percent of accuracy of all other previous SOTA scores, on average. For the 20 different datasets we consider, SOTA accuracy required 14 customized models and approaches – nearly three individualized setups for every four datasets. We argue that the fact that MCP is comparable to or surpasses SOTA, with no task-specific tuning, is evidence for the efficiency, generality, and overall promise of foundation models in MCQA.

Our primary contribution is three-fold: 1) We present an argument for multiple-choice prompting over cloze prompting and formally define multiple choice symbol binding (MCSB), a required ability for an LLM to benefit from MCP; 2) We show that not all LLMs are equally skilled in this regard; and 3) Across 20 diverse datasets, we show that the models most capable of MCSB can individually approach or beat SOTA on most of the considered tasks when prompted with multiple choice prompting instead of the near-universal approach of cloze prompting.

Code is available online.[1]

2 Related Work

Transformers (Vaswani et al., 2017) have revolutionized the field of NLP by allowing models to effectively absorb much larger datasets via massive scaling in parameter count and compute; these three factors are proportional to lower loss in models (Kaplan et al., 2020; Henighan et al., 2020; Hernandez et al., 2021). Parameter counts have quickly grown from 1.5B in 2018 (Radford et al., 2018) to 540B in 2022 (Chowdhery et al., 2022), and in general, larger models are tested on a more extensive suite of tasks to demonstrate their capacity for transfer. This invariably includes multiple choice question answering tasks, and nearly every LLM we know of uses cloze prompting for these tasks (Brown et al., 2020; Du et al., 2022; Smith et al., 2022; Chowdhery et al., 2022; Lieber et al., 2021).

It was, in part, these massive language models that prompted the coining of the phrase "foundation models" (Bommasani et al., 2021). This is a family of large models that are heavily trained on massive datasets in a self-supervised fashion. They derive general knowledge about a modality and can transfer with impressive sample efficiency to a great number of downstream tasks. A key part of the vision of these models is that they can be repurposed, avoiding the energy, storage, and human capital costs associated with ad hoc models. Our work supports this vision of LLMs as one such foundation model (if prompted appropriately), in contrast to previous views that they are unsuitable for large sets of MCQA tasks.

We are not the first to use the MCP approach. Rae et al. (2021) use this approach on just two out of their massive suite of tasks, namely TruthfulQA (Lin et al., 2021b) and RACE (Lai et al., 2017). Kadavath et al. (2022) use MCP while investigating to what extent language models are well-calibrated, and to what degree they "know what they know." In this work we explore what makes a model compatible with MCP and how prompting with MCP compares to prompting with CP in terms of LLM accuracy.

We are also not the first to notice that the method of prompting matters for LLM performance. A whole subfield of prompt engineering exists, with papers suggesting methods for ranking prompts based on mutual information (Sorensen et al., 2022), showing that LLMs suffer from majority label

[1] URL removed for anonymous submission, but code is attached

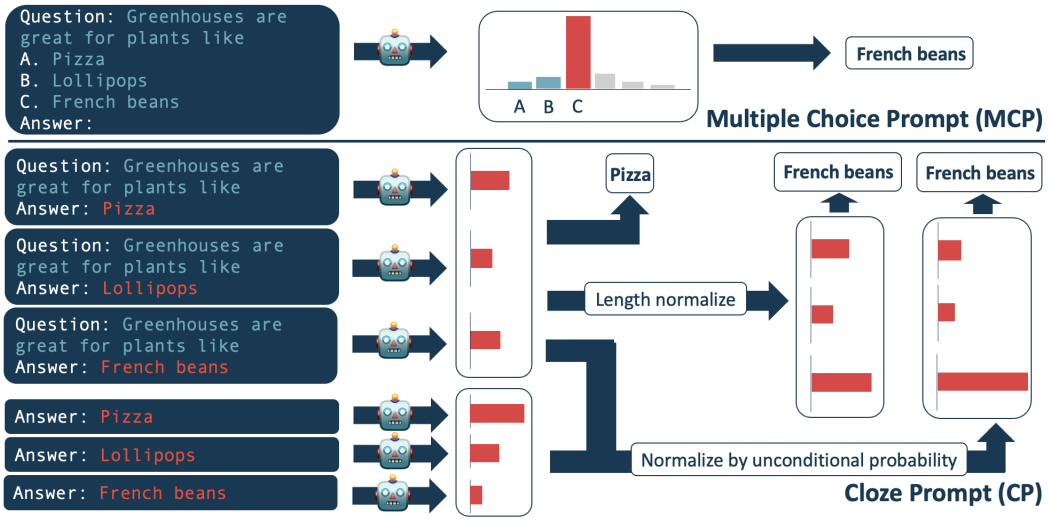

Figure 1: Visualization of the multiple choice prompt (MCP) and cloze prompt (CP) for an example question. Given the same raw data (question and candidate answers) the answer selected when using CP depends on choice of normalization strategy. Question taken from OpenBookQA (Mihaylov et al., 2018) with slight modification.

and recency biases (Zhao et al., 2021), and arguing that most few-shot learning is not, in fact, few-shot (Perez et al., 2021).

Given that CP has, up to this point, been the de facto prompting method for LLMs on MCQA tasks, several works have endeavored to improve it with normalization schemes that outperform those used in Brown et al. (2020). These include Contextual Calibration (Zhao et al., 2021) and Domain Conditional PMI (Holtzman et al., 2021).

While in this work we consider methods for effectively leveraging LLMs for MCQA tasks, past work has successfully used many types of models for these tasks. A wide selection of these models are referenced in Table 2. Two notable examples of models that perform well across question answering datasets are UnifiedQA (Khashabi et al., 2020) and UNICORN Lourie et al. (2021).

3 Cloze vs. Multiple Choice Prompting

In this section, we specify what is meant by cloze prompting and multiple choice prompting, and enumerate the problems that are inherent to the former and ameliorated by the latter.

In cloze prompting, a question is passed to an LLM, the candidate answers are each independently scored by the model, and the answer option selected by the model is chosen to be the one assigned the highest probability. Recognizing that probabilities of answers might be skewed by especially common or uncommon tokens or sequences of varying length, Brown et al. (2020) made use of two different normalization procedures. In one, the sequence's probability is normalized for length by taking the nth root, or $P(x_1, x_2, ..., x_n) = \sqrt[n]{\prod_{i=1}^{n} P(x_i)}$. In the other, the answer's probability is normalized by the unconditional probability of the answer, or $\frac{P(\text{completion}|\text{context})}{P(\text{completion}|\text{answer_context})}$ where answer_context is the string "`Answer:  `". In the remainder of the the paper, we refer to length normalization as LN, unconditional normalization as UN, and neither as Raw. See Figure 1 for a visualization on how these strategies work.

In MCP, on the other hand, a question and its symbol-enumerated candidate answers are all passed to an LLM as a single prompt. The prompt is structured such that the LLM must only predict a single token (such as "A", "B", etc.). The answer choice associated with the token assigned the highest probability by the model is chosen to be the model's answer. The probabilities of these

symbols therefore serve as a proxy for each answer's probability. There are several problems with cloze prompting that do not apply to multiple choice prompting:

Conflation of likelihood as answer and likelihood as natural language One outstanding problem with CP is that the likelihood of the answer's text could be conflated with the likelihood of the text as an answer. For example, if asked which quote is from Shakespeare's *Macbeth*, the phrase `Give sorrow words; the grief that does not speak knits up o-er wrought heart and bids it break.` might be less common or grammatical than other sentences, which could artificially deflate its score. MCP does not face this problem because there is no grammar to alter the score.

Reliance on normalization procedures With CP, use of special normalization strategies are typically essential for achieving high performance. These often incur a computational cost or depend on choice of tokenization scheme. With MCP, there is no need for any normalization strategy.

No direct comparison between answers In CP, candidate answers are not compared to each other except implicitly through their final probabilistic scores. MCP gives LLMs the ability to explicitly compare and contrast different answer options. This also makes LLM+MCP more comparable to SOTA methods that typically have all answer options presented at once.

Expense Lastly, cloze prompting is computationally expensive. For a question with n possible answer choices CP requires n forward passes through the LLM for the raw or LN normalization strategies, and $2n$ forward passes for the UN strategy. MCP only requires a single pass (which is independently slightly cheaper than CP forward passes because the model only needs to generate a single output token).

4 THE CHALLENGE OF MULTIPLE CHOICE SYMBOL BINDING

When presenting a multiple choice question, the candidate answers must be enumerated in some order. Ideally, humans' answers to such questions are order-invariant. If an LLM exhibits the same characteristic, we say that it is capable of *multiple choice symbol binding* (MCSB). Interestingly, LLMs vary substantially in terms of this ability.

Consider the multiple choice prompt example from Figure 1. Given the answer order "Pizza", "Lollipop", "French beans" (as shown in the figure) GPT-3 (Davinci) Brown et al. (2020) assigns the highest probability to the token "A," which is associated with pizza. However, if we change the ordering to "French beans", "Lollipops", "Pizza", GPT-3 surprisingly still assigns the highest probability to "A," which is now associated with French beans. Simply changing the order of the candidate answers changes the model's answer.

How can we compare the relative symbol binding ability of two models? One way is to measure what we term *Proportion of Plurality Agreement* (PPA). Given a question with n answer options, there are $n!$ different ways these options can be associated with an ordered, fixed set of symbols. To measure PPA for a given model and question we present that question to the model with each different ordering and for each ordering record the answer assigned the highest probability by the model. PPA for that question is the proportion of orderings that chose the plurality answer among all orderings. For a dataset PPA is averaged over all questions.

Importantly, PPA measures order invariance irrespective of model ability to perform a task. If a model performs poorly on a task but answers consistently across possible orders of answer options it will still have a high PPA. For a dataset where each question has n answer choices, the baseline PPA is $1/n$.

Using PPA, we compare several popular LLMs' MCSB ability. The first is GPT-3 (Brown et al., 2020) (Davinci). We also consider Codex (Chen et al., 2021) (Davinci) and Instruct (Ouyang et al., 2022) (Curie and Davinci). These two models were trained, respectively, for code modelling and instruction following, both having been initialized from GPT-3 weights. We also evaluate the MCSB ability of GPT-2 Radford et al. (2019), CodeParrot (Tunstall et al., 2022) (GPT-2 with further training on code), and Jurassic-1 (Lieber et al., 2021) (Jumbo). The parameter count of these models

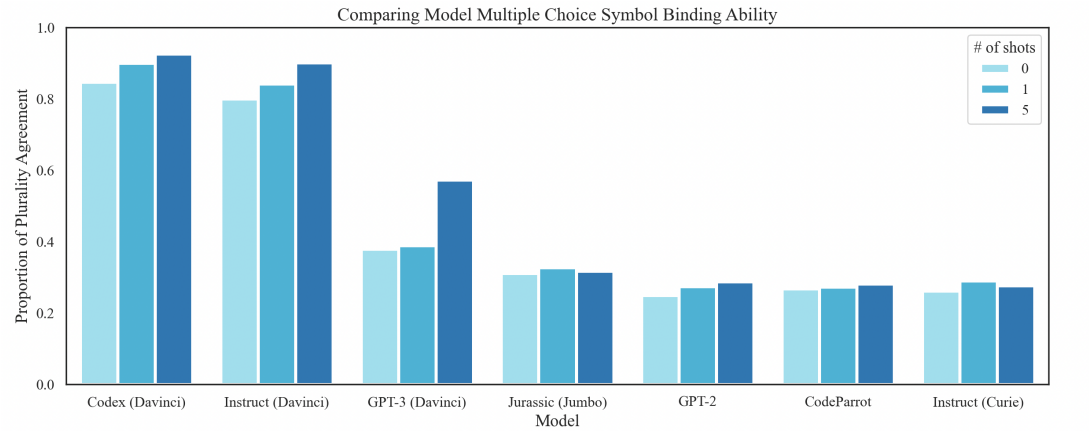

Figure 2: Comparison of the multiple choice symbol binding ability of different models as measured by PPA on a subset of OpenBookQA (Mihaylov et al., 2018).

spans from 1.5B (GPT-2) to 178B (Jurassic-1 Jumbo). We use API requests for GPT-3, Codex, Instruct, and Jurassic-1. For GPT-2 and CodeParrot we use the model checkpoints from Hugging Face Transformers (Wolf et al., 2020).

We evaluate on the OpenBookQA dataset (Mihaylov et al., 2018), a multiple choice QA dataset composed of 500 science questions. We randomly sample 100 instances from the dataset to reduce computational cost. Each question has four answer options, meaning that the PPA random baseline is 25%. We choose to use OpenBookQA because of its relatively small size and because it is a traditional multiple choice dataset, with questions modeled after human multiple choice exams.

The results of our evaluation can be found in Figure 2. The most immediately interesting result is that Codex (Davinci) and Instruct (Davinci) significantly outperform the other models in terms of PPA, showing remarkable invariance to answer ordering. GPT-3 seems to perform about half as well, interestingly outperforming the larger Jurassic-1 (Jumbo) model. GPT-2, CodeParrot, and Instruct (Curie) all have PPA scores close to the 25% baseline.

Another apparent trend is that providing exemplars to a model increases PPA consistently across models. This is especially notable for GPT-3. While we do not explore what about these models causes high MCSB ability, it appears that model size could be an important factor because Instruct Davinci has a much higher PPA than Instruct Curie. This hypothesis is in line with the work of Wei et al. (2022).

It also seems like further training (on source code or with reinforcement learning based on human preferences) is essential (Codex and Instruct both outperform GPT-3 substantially even though they are basically fine-tuned versions of GPT-3). Intuitively, it makes sense that training on source code increases symbol binding ability because performing well on a next token prediction task for code requires grasping symbol binding as it is used in e.g., dictionaries and list indexing. Training on source code alone does not seem sufficient in and of itself, though, because CodeParrot didn't achieve notably better PPA than did GPT-2.

Given the strong multiple choice symbol binding ability of Codex and Instruct, a natural next question is whether using these models with MCPs results in higher accuracy. We address this question in the next section.

5 Experimental Setup

We evaluate the performance of a model with strong MCSB ability and multiple choice prompts across a set of 20 diverse datasets. In this section we discuss how that model was chosen (Section 5.1), and which datasets we use for evaluation (Section 5.2). We also address the possibility of dataset leakage into model training data (Section 5.3) and explain our prompt engineering choices (Section 5.4).

Dataset	GPT-3				Instruct				Codex			
	Raw	LN	UN	MCP	Raw	LN	UN	MCP	Raw	LN	UN	MCP
OpenBookQA	35.0	46.8	**57.4**	41.4	41.8	49.6	58.4	**77.4**	43.0	51.4	65.6	**83.0**
StoryCloze	75.2	**76.4**	75.6	70.8	78.0	78.8	82.4	**97.6**	80.8	83.6	84.0	**97.4**
RACE-m	55.6	**57.2**	56.6	50.2	63.2	64.8	66.8	**89.6**	63.4	67.0	63.8	**89.2**

Table 1: Comparing the performance of large language models across prompting strategies. For cloze prompts there are three normalization strategies. Raw is the strategy of selecting an answer choice based on raw probabilities. LN and UN are selecting an answer choice after normalizing probabilities based on length or unconditional completion probability respectively (see Brown et al. (2020)). MCP is multiple choice prompt. The best accuracy for each model and dataset is bolded.

5.1 MODELS

In Section 4 we observed that models like Codex (Davinci) (Chen et al., 2021) and Instruct (Davinci) (Ouyang et al., 2022) demonstrated much stronger MCSB ability than did GPT-3. In this section we explore whether higher MCSB ability leads to higher multiple choice task accuracy. We evaluate model performance on a commonsense reasoning dataset (OpenBookQA (Mihaylov et al., 2018)), a cloze/completion dataset (StoryCloze (Mostafazadeh et al., 2016)), and a reading comprehension dataset (RACE-m Lai et al. (2017)). See Section 5.2 for more information on these datasets. We randomly sample 500 instances from StoryCloze and RACE-m each to reduce computational costs.

The results of our comparison can be found in Table 1. We note that whereas choosing answer options based on cloze prompts (with max raw probability (Raw), max raw probability after length normalization (LN), or max raw probability after unconditional normalization (UN)) performs best for GPT-3, MCP always performs best for Instruct and Codex, and it does so by a sizeable margin. Because these models have high MCSB ability they are able to effectively leverage MCP prompts to achieve higher accuracy. Instruct and Codex both outperform GPT-3 by large margins across all tasks.

For all of our further experiments we choose to use Codex (Davinci) because its training setup seems to explicitly promote MCSB ability, it had the highest PPA value of all models we tested, and it was the most computationally cheap (since it is currently in free beta). Additionally, Codex should not add any dataset leakage issues that were not already present in GPT-3 since it was exclusively fine-tuned on Python files.

5.2 DATASETS

We compare multiple choice prompts and cloze prompts across a diverse array of datasets. Examples of questions from each of these datasets can be found in Appendix A.

Common sense reasoning ARC (Clark et al., 2018) consists of grade-school science questions. Its challenge set is filtered down to questions not correctly answered by simple retrieval and co-occurence solvers. CODAH (Chen et al., 2019) questions were adversarially generated to fool BERT-base Devlin et al. (2019) fine-tuned on SWAG (Zellers et al., 2018). CommonsenseQA (Talmor et al., 2019) questions are based on ConceptNet (Speer et al., 2017) concepts, allowing for diversity and incorporation of world knowledge into questions. COPA (Roemmele et al., 2011) is a causal reasoning dataset that tasks models with determining the cause or effect of a premise. Composed of questions about metaphors, Fig-QA (Liu et al., 2022a) is designed to test nonliteral reasoning abilities. MedMCQA (Pal et al., 2022) has questions from many areas of the medical domain drawn from medical exams (see Appendix F for a full list of subjects covered). The Massive Multitask Language Understanding (MMLU) benchmark (Hendrycks et al., 2021) consists of diverse tasks from STEM, the humanities, and the social sciences (a full list of tasks can be found in Appendix G). OpenBookQA (Mihaylov et al., 2018) endeavors to test world knowledge and reasoning using science questions. PIQA (Bisk et al., 2020) questions focus on the area of commonsense reasoning related to understanding of physical interactions with the world. RiddleSense (Lin et al., 2021a) and Social IQa (Sap et al., 2019) are designed to test for model ability to reason about riddles and social situations respectively.

Natural language inference ANLI (Nie et al., 2020) is a dataset of adversarially generated NLI questions. These questions were generated by annotators in three rounds, with annotators seeking to generate questions challenging for increasingly capable models trained for NLI.

Cloze and completion tasks HellaSwag (Zellers et al., 2019) tasks models with predicting the best continuation for a video caption or WikiHow article. StoryCloze (Mostafazadeh et al., 2016) is also a continuation prediction task, but with short, four-sentence stories.

Text classification AG News (Zhang et al., 2015) is a news classification task.

Winograd-style tasks Winogrande (Sakaguchi et al., 2021) is a set of winograd schema questions that was carefully crowdsourced and then filtered with a bias mitigation algorithm. Because we are doing evaluation without fine-tuning in this work we evaluate in the smallest training data setting (XS), but also include results in the largest setting (XL) to facilitate comparison with prior work.

Reading comprehension Cosmos QA (Huang et al., 2019) is a commonsense reasoning based reading comprehension task designed to test model ability to "read between the lines." The questions in the DREAM (Sun et al., 2019) dataset are based on dialogues. LogiQA (Liu et al., 2020) tasks models with answering questions from a human exam that tests critical thinking. RACE (Lai et al., 2017) is a widely used reading comprehension dataset with questions taken from middle and high school English exams for Chinese students.

5.3 Addressing Dataset Leakage

A concern in all work with LLMs is that an LLM being used for evaluation may have been exposed to the contents of an evaluation dataset while it was being pre-trained. We only evaluate on Codex, which was initialized from GPT-3's weights and fine-tuned exclusively on Python files. This means that the risks of dataset leakage we face are effectively the same of those faced by the authors of the GPT-3 paper (Brown et al., 2020).

The dataset leakage risk of the 9 of our datasets GPT-3 was tested on in the GPT-3 paper was evaluated extensively by the authors of that paper. The risk of leakage remains for the other 11 datasets. Although we weren't able to acquire the GPT-3 training data to perform a manual collision check, there are several reasons why we suspect dataset leakage is not meaningfully impacting our results.

First, we note (anectdotally) that Codex+MCP errors in evaluation occur disproportionately on questions that are ill-defined (not solvable by humans, having multiple answers, etc.) (see, for example, non-cherry-picked examples in Appendix D). If there were training set leakage, this set would be more independent of being well-formed than it is; that is, failures by the model seem to be due more to the deck being stacked against good reasoning than to the model having failed to memorize the data. Second, shuffling the order of the symbol-enumerated candidate answers doesn't systematically harm performance (see Appendix C). Third, CP+MCP performance on publicly available test sets and private test sets is generally comparable.

The strongest argument that MCP's performance is not determined by training set leakage is CP's meaningfully inferior performance for the same model; if the model had memorized the training data, it would have ascribed more probability to the correct answer regardless of the prompting method.

5.4 Prompt Engineering

In an effort to keep the comparison of cloze prompts and multiple choice prompts as simple as possible we systematically avoid prompt engineering. Unlike Brown et al. (2020) we do not tune K, the number of few-shot exemplars, based on a development set, nor do we develop highly task specific prompt phrasings. K is always chosen to be as high as possible while respecting Codex's 4,000 token context limit.

Our prompt phrasing is always consistent: We prefix the raw question with `Question:`, list answer options with associated letters (like `A. Lollipop`), and finish prompts with `Answer:`. We mea-

Dataset	N	K	Zero-Shot		One-Shot		Few-Shot		Server	SOTA
			CP	MCP	CP	MCP	CP	MCP		
AG News	4	38	68.2	**83.5**	77.6	**87.1**	**90.1**	89.4		95.6[a]
ANLI R1	3	27	**45.3**	33.2	35.6	**61.7**	58.4	**64.2**		75.5[b]
ANLI R2	3	26	**39.2**	33.6	35.7	**53.0**	51.8	**55.2**		58.6[c]
ANLI R3	3	26	**37.8**	34.3	35.5	**47.8**	54.2	**54.5**		53.4[c]
ARC (Challenge)	4	50	58.9	**81.7**	64.1	**82.8**	66.6	**86.1**		86.5[d]
ARC (Easy)	4	57	84.2	**93.1**	85.9	**93.5**	87.8	**94.7**		94.8[d]
CODAH	4	63	56.8	**76.0**	65.4	**87.8**	73.6	**91.9**		84.3[e]
CommonsenseQA	5	79	68.5	**72.0**	73.1	**78.9**	78.6	**83.2**	76.6	79.1[f]
COPA	2	113	**92.0**	89.0	95.0	**99.0**	96.0	**100.0**	—	99.2[d]
Cosmos QA	4	24	43.0	**75.5**	44.0	**81.8**	38.1	**82.4**	83.5	91.8[g]
DREAM	3	7	72.7	**91.3**	82.5	**93.3**	84.3	**94.1**		92.6[h]
Fig-QA	2	99	79.6	**84.7**	82.4	**86.7**	82.5	**94.0**	93.1	90.3[i]
HellaSwag	4	16	—	71.0	—	75.1	—	73.6	—	93.9[g]
LogiQA	4	16	36.6	**44.5**	37.5	**45.3**	37.8	**47.3**		42.5[j]
MedMCQA	4	58	37.8	**52.1**	42.1	**53.9**	41.2	**54.4**	58.0	41.0[k]
MMLU	4	5	49.5	**62.1**	—	68.2	—	69.5		67.5[l]
OpenBookQA	4	83	63.2	**72.0**	64.0	**81.6**	71.2	**87.0**		87.2[f]
PIQA	2	35	**83.7**	73.7	**84.1**	81.8	**86.1**	84.5	—	90.1[g]
RACE-h	4	4	52.3	**82.1**	53.2	**85.1**	55.2	**86.2**		89.8[m]
RACE-m	4	8	67.5	**85.4**	70.5	**89.3**	71.7	**90.3**		92.8[m]
RiddleSense	5	59	**79.8**	67.6	**89.1**	77.1	**91.3**	83.9	80.0	68.8[f]
Social IQa	3	72	52.1	**64.4**	58.1	**72.2**	62.4	**74.9**	76.0	83.2[g]
StoryCloze	2	44	80.3	**97.5**	83.4	**98.3**	88.2	**98.5**		89.0[n]
Winogrande (XL)	2	102	62.5	**64.5**	**71.6**	**71.6**	**75.5**	72.1	72.3	91.3[g]
Winogrande (XS)	2	102	63.0	**64.8**	71.0	**71.3**	**76.2**	73.6	73.8	79.2[g]

Table 2: Comparison of multiple choice prompt (MCP) and cloze prompt (CP) with Codex. N is the number of answer options for each question. K exemplars are provided in the few-shot setting. SOTA values come from [a]Yang et al. (2019), [b]Wang et al. (2021), [c]Lan et al. (2019), [d]Zoph et al. (2022), [e]Yang et al. (2020), [f]Khashabi et al. (2020), [g]Lourie et al. (2021), [h]Zhang & Yamana (2022), [i]Liu et al. (2022b), [j]Jiao et al. (2022), [k]Gu et al. (2020), [l]Hoffmann et al. (2022), [m]Jiang et al. (2020), and [n]Chowdhery et al. (2022). These values are computed using a private test set when it exists (for rows with a Server value) or on a public test set otherwise. The best prompt method for each dataset and exemplar count is bolded. The SOTA including our experimental results is underlined. Values marked with — could not be computed - mostly due to computational restraints (see Appendix B).

sure model probability for an answer via the probability of the symbol associated with it. When a passage is included as part of a question (as in reading comprehension) we insert the passage before the question and prefix it with `Passage:` (`Story:` for StoryCloze (Mostafazadeh et al., 2016) and `Dialogue:` for DREAM (Sun et al., 2019)). See Appendix A for examples. Our prompts are simple and modelled after those in Brown et al. (2020).

Our goal in this work is to provide a fair comparison between cloze prompts and multiple choice prompts, and not to maximize accuracy by careful prompt engineering. However, we can say anecdotally that multiple choice prompts seem very robust to different wordings and symbol choices. Prompt engineering for MPCs could be interesting future work.

6 RESULTS

The results of our experiments can be found in Table 2. In the table the CP columns represent cloze prompting with the best possible strategy on the test set. That is, for each dataset and cloze prompts we calculated test accuracy based on raw accuracy, raw accuracy with length normalization, and raw accuracy with unconditional normalization. The CP column contains *the highest accuracy of*

any strategy. This is an unrealistically strong baseline, since no single normalization scheme is universally optimal. The results of each individual scheme on all datasets can be found in Table 4.

The results in Table 2 validate our hypothesis that for a model with high MCSB ability (Codex) using MCP outperforms CP. This holds consistently across datasets and number of exemplars. MCP increases accuracy over CP by 8.3, 12.2, and 9.7 percentage points on average in the zero, one, and few-shot settings, respectively.

The dataset with the largest gain between the cloze prompts and multiple choice prompts is Cosmos QA. For this dataset using MCP instead of CP increases accuracy by 32.5, 37.8, and 44.3 pecentage points in the zero, one, and few-shot settings, respectively. This substantial improvement on task performance is likely due to the fact that Cosmos QA questions (including their answer options) have somewhat irregular spacing[2]. This poses no issue for MCPs, but is a serious issue for CPs that rely on the linguistic representation of answer options.

There are four datasets where CP outperforms MCP - AG News, PIQA, RiddleSense, and Winogrande. One thing in common between AG News, Winogrande, and RiddleSense is they tend to have short, often one word answers. For these datasets CP is acting more like MCP because answer option length and wording have less impact. Some PIQA questions have answer options much longer than normal ones, perhaps making MCSB more challenging. Additionally, PIQA questions sometimes appear very "cloze-friendly" (see example prompt in Appendix A).

In addition to consistently outperforming Codex+CP, Codex+MCP sets a new state of the art for 9 datasets. For MedMCQA Codex+MCP has an accuracy 13.4% above the old SOTA model, PubmedBERT (Gu et al., 2020). This seems to suggest that, in contrast to prior work (Moradi et al. (2021)), large language models may have high potential in the biomedical domain. The key is prompting them in a way that effectively aligns them to the task.

7 Conclusion

In this work we have argued for multiple choice prompting over the universally-practiced cloze prompting, formalized the idea of multiple choice symbol binding (MCSB), showed that large language models vary greatly in their MCSB ability, and demonstrated that for a model with high MCSB ability like OpenAI Codex (Chen et al., 2021) multiple choice prompts generally elicit much more accurate responses than do cloze prompts. This approach led to a new state-of-the-art on 9 popular datasets, and on average scored within a percentage point of previous SOTA, all using a single model and single prompting approach. This demonstrates the power of LLMs as foundation models and is a strong case study in how these models might be used broadly in the future.

The future for symbol-binding is exciting, with potential to teach LLMs new concepts and symbols representing them and have language models fold these new frameworks and ideas into their already-rich world models.

There is also a concrete lesson to be drawn from seeing such drastic improvement in LLM performance with such a simple change to prompting: an intuitive, sensible change to the way we train and use our models can quickly unlock their potential in ways that we have previously been blind to as a field. This should fill us with motivation to seek out such improvements.

Promising directions for future work include prompt engineering for multiple choice prompts; evaluating prompts on more datasets, tasks, and models; and assessing what factors are responsible for high MCSB ability. Our results suggest that the performance of large language models on multiple choice question answering tasks has been previously underestimated, and we hope that increased adoption of multiple choice prompts will lead to more effective probing and utilization of large language models.

Reproducibility Statement Source code for replicating all experiment results can be found at [URL removed for anonymous submission, but code is attached]. Names of all model checkpoints and API endpoints used can be found in `constants.py`. The only randomness in our experiments comes from the selection of few-shot exemplars, strong shuffling, and downsampling datasets. To ensure reproducibility, all of these use a single random seed that can be found in `constants.py`.

[2] `https://huggingface.co/datasets/cosmos_qa`

REFERENCES

Asaf Amrami and Yoav Goldberg. Word Sense Induction with Neural biLM and Symmetric Patterns. pp. 4860–4867, 2018.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Scott Wen-tau Yih, and Yejin Choi. Abductive commonsense reasoning. *arXiv preprint arXiv:1908.05739*, 2019.

Yonatan Bisk, Rowan Zellers, Ronan Le bras, Jianfeng Gao, and Yejin Choi. Piqa: Reasoning about physical commonsense in natural language. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7432–7439, Apr. 2020. doi: 10.1609/aaai.v34i05.6239. URL https://ojs.aaai.org/index.php/AAAI/article/view/6239.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. COMET : Commonsense Transformers for Automatic Knowledge Graph Construction.

Zied Bouraoui, Jose Camacho-collados, and Steven Schockaert. Inducing Relational Knowledge from BERT.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *CoRR*, abs/2005.14165, 2020. URL https://arxiv.org/abs/2005.14165.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harrison Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Joshua Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code. *CoRR*, abs/2107.03374, 2021. URL https://arxiv.org/abs/2107.03374.

Michael Chen, Mike D'Arcy, Alisa Liu, Jared Fernandez, and Doug Downey. CODAH: An adversarially-authored question answering dataset for common sense. In *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, pp. 63–69, Minneapolis, USA, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-2008. URL https://aclanthology.org/W19-2008.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas

Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways, 2022. URL https://arxiv.org/abs/2204.02311.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge, 2018. URL https://arxiv.org/abs/1803.05457.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Nan Du, Yanping Huang, Andrew M Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, et al. Glam: Efficient scaling of language models with mixture-of-experts. In *International Conference on Machine Learning*, pp. 5547–5569. PMLR, 2022.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *CoRR*, abs/2007.15779, 2020. URL https://arxiv.org/abs/2007.15779.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 107–112, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2017. URL https://aclanthology.org/N18-2017.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B Brown, Prafulla Dhariwal, Scott Gray, et al. Scaling laws for autoregressive generative modeling. *arXiv preprint arXiv:2010.14701*, 2020.

Danny Hernandez, Jared Kaplan, Tom Henighan, and Sam McCandlish. Scaling laws for transfer. *arXiv preprint arXiv:2102.01293*, 2021.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models, 2022. URL https://arxiv.org/abs/2203.15556.

Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. Surface form competition: Why the highest probability answer isn't always right. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7038–7051, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.564. URL https://aclanthology.org/2021.emnlp-main.564.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2391–2401, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1243. URL https://aclanthology.org/D19-1243.

Yufan Jiang, Shuangzhi Wu, Jing Gong, Yahui Cheng, Peng Meng, Weiliang Lin, Zhibo Chen, and Mu Li. Improving machine reading comprehension with single-choice decision and transfer learning. *CoRR*, abs/2011.03292, 2020. URL https://arxiv.org/abs/2011.03292.

Fangkai Jiao, Yangyang Guo, Xuemeng Song, and Liqiang Nie. MERIt: Meta-Path Guided Contrastive Learning for Logical Reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 3496–3509, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.276. URL https://aclanthology.org/2022.findings-acl.276.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. UNIFIEDQA: Crossing format boundaries with a single QA system. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1896–1907, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.171. URL https://aclanthology.org/2020.findings-emnlp.171.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 785–794, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1082. URL https://aclanthology.org/D17-1082.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A lite BERT for self-supervised learning of language representations. *CoRR*, abs/1909.11942, 2019. URL http://arxiv.org/abs/1909.11942.

Opher Lieber, Or Sharir, Barak Lenz, and Yoav Shoham. Jurassic-1: Technical details and evaluation. *White Paper. AI21 Labs*, 2021.

Bill Yuchen Lin, Ziyi Wu, Yichi Yang, Dong-Ho Lee, and Xiang Ren. Riddlesense: Reasoning about riddle questions featuring linguistic creativity and commonsense knowledge. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL-IJCNLP 2021): Findings*, 2021a. to appear.

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

Emmy Liu, Chenxuan Cui, Kenneth Zheng, and Graham Neubig. Testing the ability of language models to interpret figurative language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4437–4452, Seattle, United States, July 2022a. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.330. URL https://aclanthology.org/2022.naacl-main.330.

Emmy Liu, Chenxuan Cui, Kenneth Zheng, and Graham Neubig. Testing the ability of language models to interpret figurative language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4437–4452, Seattle, United States, July 2022b. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.330. URL https://aclanthology.org/2022.naacl-main.330.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning. *CoRR*, abs/2007.08124, 2020. URL https://arxiv.org/abs/2007.08124.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. Linguistic knowledge and transferability of contextual representations. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 1:1073–1094, 2019. doi: 10.18653/v1/n19-1112.

Nicholas Lourie, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Unicorn on rainbow: A universal commonsense reasoning model on a new multitask benchmark. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13480–13488, May 2021. doi: 10.1609/aaai.v35i15.17590. URL https://ojs.aaai.org/index.php/AAAI/article/view/17590.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2381–2391, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1260. URL https://aclanthology.org/D18-1260.

Milad Moradi, Kathrin Blagec, Florian Haberl, and Matthias Samwald. GPT-3 models are poor few-shot learners in the biomedical domain. *CoRR*, abs/2109.02555, 2021. URL https://arxiv.org/abs/2109.02555.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 839–849, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1098. URL https://aclanthology.org/N16-1098.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4885–4901, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.441. URL https://aclanthology.org/2020.acl-main.441.

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, 2022. URL https://arxiv.org/abs/2203.02155.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In Gerardo Flores, George H Chen, Tom Pollard, Joyce C Ho, and Tristan Naumann (eds.), *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pp. 248–260. PMLR, 07–08 Apr 2022. URL https://proceedings.mlr.press/v174/pal22a.html.

Ethan Perez, Douwe Kiela, and Kyunghyun Cho. True few-shot learning with language models. *Advances in Neural Information Processing Systems*, 34:11054–11070, 2021.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. Language models as knowledge bases? *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pp. 2463–2473, 2020. doi: 10.18653/v1/d19-1250.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*, 2021.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S. Gordon. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*, 2011.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. *Commun. ACM*, 64(9):99–106, aug 2021. ISSN 0001-0782. doi: 10.1145/3474381. URL https://doi.org/10.1145/3474381.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. Social IQa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4463–4473, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1454. URL https://aclanthology.org/D19-1454.

Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, et al. Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model. *arXiv preprint arXiv:2201.11990*, 2022.

Taylor Sorensen, Joshua Robinson, Christopher Michael Rytting, Alexander Glenn Shaw, Kyle Jeffrey Rogers, Alexia Pauline Delorey, Mahmoud Khalil, Nancy Fulda, and David Wingate. An information-theoretic approach to prompt engineering without ground truth labels. *arXiv preprint arXiv:2203.11364*, 2022.

Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1), Feb. 2017. doi: 10.1609/aaai.v31i1.11164. URL https://ojs.aaai.org/index.php/AAAI/article/view/11164.

Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. DREAM: A challenge data set and models for dialogue-based reading comprehension. *Transactions of the Association for Computational Linguistics*, 7:217–231, 2019. doi: 10.1162/tacl_a_00264. URL https://aclanthology.org/Q19-1014.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4149–4158, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1421. URL https://aclanthology.org/N19-1421.

Lewis Tunstall, Leandro von Werra, and Thomas Wolf. *Natural Language Processing with Transformers: Building Language Applications with Hugging Face*. O'Reilly Media, Incorporated, 2022. ISBN 1098103246. URL https://books.google.ch/books?id=7hhyzgEACAAJ.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Boxin Wang, Shuohang Wang, Yu Cheng, Zhe Gan, Ruoxi Jia, Bo Li, and Jingjing Liu. Infobert: Improving robustness of language models from an information theoretic perspective. In *International Conference on Learning Representations*, 2021.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models, 2022. URL https://arxiv.org/abs/2206.07682.

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. URL https://aclanthology.org/2020.emnlp-demos.6.

Yiben Yang, Chaitanya Malaviya, Jared Fernandez, Swabha Swayamdipta, Ronan Le Bras, Ji-Ping Wang, Chandra Bhagavatula, Yejin Choi, and Doug Downey. Generative data augmentation for commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1008–1025, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.90. URL https://aclanthology.org/2020.findings-emnlp.90.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/dc6a7e655d7e5840e66733e9ee67cc69-Paper.pdf.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 93–104, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1009. URL https://aclanthology.org/D18-1009.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4791–4800, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1472. URL https://aclanthology.org/P19-1472.

Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *NeurIPS*, 2015.

Yuxiang Zhang and Hayato Yamana. HRCA+: Advanced multiple-choice machine reading comprehension method. In *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, pp. 6059–6068, Marseille, 2022. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2022/pdf/2022.lrec-1.651.pdf.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, pp. 12697–12706. PMLR, 2021.

Barret Zoph, Irwan Bello, Sameer Kumar, Nan Du, Yanping Huang, Jeff Dean, Noam Shazeer, and William Fedus. St-moe: Designing stable and transferable sparse expert models, 2022. URL https://arxiv.org/abs/2202.08906.

Yukun Zuo, Quan Fang, Shengsheng Qian, Xiaorui Zhang, and Changsheng Xu. Representation Learning of Knowledge Graphs with Entity Attributes and Multimedia Descriptions. *2018 IEEE 4th International Conference on Multimedia Big Data, BigMM 2018*, pp. 2659–2665, 2018. doi: 10.1109/BigMM.2018.8499179.

A Prompts Used For Each Dataset

In this section we provide examples of the prompts used for each dataset. Any typos in questions are present in the datasets they are drawn from. The examples are multiple choice prompts (MCPs).

```
Article: At Disney, Mending Fences or Moving On? Without Michael D.
Eisner at the helm of the Walt Disney Company, will Harvey Weinstein and
Steven P. Jobs stay as partners?
Question: What is the best classification for this article?
A. World
B. Sports
C. Business
D. Sci/Tech
Answer:
```

Figure 3: Prompt example for the AG News dataset.

```
Premise: press release: Did you know that Marquette University owns the
original manuscripts for J. R. R. Tolkien's The Hobbit and The Lord of
the Rings? William Fliss, Archivist in Marquette's Department of Special
Collections and University Archives, will share the remarkable tale of
how these literary treasures came to Wisconsin, and he will explain what
these manuscripts can tell us about one of the most iconic authors of the
twentieth century. Cost: Suggested
donation of $3/person
Hypothesis: Attendees will pay $3.
A. Hypothesis is definitely true given premise
B. Hypothesis might be true given premise
C. Hypothesis is definitely not true given premise
Answer:
```

Figure 4: Prompt example for the ANLI dataset. Wording was taken from Gururangan et al. (2018).

Removing the answer options from an example question yields the cloze prompt (CP) we used for that question.

```
Question: What adaptation is necessary in intertidal ecosystems but not
in reef ecosystems?
A. the ability to live in salt water
B. the ability to use oxygen in respiration
C. the ability to cope with daily dry periods
D. the ability to blend into the surroundings
Answer:
```

Figure 5: Prompt example for the ARC dataset.

```
Question: Two friends are looking at cakes in a bakery. They
A. start throwing water at each other.
B. pay the bartender for the cake and leave the pub.
C. run a marathon.
D. order a cheesecake.
Answer:
```

Figure 6: Prompt example for the CODAH dataset.

```
Question: How does a bishop move from one place to another?
A. chess game
B. church
C. in a car
D. queen
E. cathedral
Answer:
```

Figure 7: Prompt example for the CommonsenseQA dataset.

```
Question: The pond froze over for the winter so
A. People skated on the pond.
B. People brought boats to the pond.
Answer:
```

Figure 8: Prompt example for the COPA dataset.

```
Passage: Today I went to the new Trader Joe 's on Court Street . It is so
pretty . It 's inside what appears to be an old bank . It was spacious
and there were no NYU students wearing velour sweatpants .
Question: What was the narrator very impressed with ?
A. None of the above choices .
B. The grocery store .
C. The NYU campus .
D. The bank workers .
Answer:
```

Figure 9: Prompt example for the Cosmos QA dataset.

```
Dialogue: M: I want to send this package by first-class mail. W: Do you
want it insured? M: Yes, for 50 dollars, please. I'd also like some
stamps--a book of 22 and three airmail. W: You'll have to get those at
the stamp window over there, next to general delivery. M: Can I get money
orders there, too? W: No, that's to the left, three windows down the hall.
Question: Where can the man get money orders?
A. At the stamp window.
B. Next to general delivery.
C. Three windows down the hall.
Answer:
```

Figure 10: Prompt example for the DREAM dataset.

```
Question: The chihuahua believes it is a wolf, meaning
A. The small dog thinks it is undefeatable
B. The small dog always stays on your lap
Answer:
```

Figure 11: Prompt example for the Fig-QA dataset.

```
Passage: [header] How to get around london easily [title] Know how you're
going to travel. [step] The easiest method of travel in london is the
tube. For this, it is easiest to buy what is called an' oyster card' or a
get a travelcard for all zones from one of the automated machines in a
tube station.
Question: Which choice best continues the passage?
A. People take an oyster card (this is a permanent, digital card) for
optimal services and there are a number of reputable card companies that
buy oyster cards. [title] Firstly, when considering destination, are you
travelling with a package? [step] Do you want to surprise your friends
and family at london.
B. These cover buses, tubes, trams and overground trains throughout
the city. This is usually the best option, especially for tourists,
as you can travel as much as you'd like in one day with one flat fare.
C. [title] Know the locations of the railway stations you are going
to. [step] Look for normal bus lines around london.
D. The card lets you ride on the tube without the added cost of any
rail, bus, or train stops. You can also travel by car (train makes
easier to return for rides in london if you're travelling as
non-railway cars), train from the station, or post office.
Answer:

Passage: (Kayaking) Man is kayaking in a calm river. Man is standing in
te seasore talking to the camera and showing the kayak.
Question: Which choice best continues the passage?
A. man is getting in the sea and sits in a kayak.
B. man is kayaking in rafts and going through mountains.
C. man is kayaking on a snowy river.
D. man is returning in a river with a land trail and a shop.
Answer:
```

Figure 12: Prompt examples for the HellaSwag dataset. We include a WikiHow example (top) and an ActivityNet example (bottom) because they are formatted slightly differently.

```
Question: Statistics show that train accidents in a country mostly occur
in the southern region, so it is safer to travel by train in the northern
region. Which of the following can best refute the above argument?
A. Slower train speeds in the north of the country
B. There are many more train lines in the south of the country than in
the north
C. Many lines in the south of the country already use EMUs
D. Most of the northern part of the country is mountainous and is
more suitable for car driving
Answer:
```

Figure 13: Prompt example for the LogiQA dataset.

```
Question: Keratin in skin is softer than keratin in nail because keratin
in skin has -
A. Less number of disulphide bonds
B. Less number of salt bridges
C. High sodium content
D. Different affinity for water
Answer:
```

Figure 14: Prompt example for the MedMCQA dataset.

```
Question: A state has recently enacted a statute prohibiting the disposal
of any nuclear wastes within the state. This law does not contravene or
conflict with any federal statutes. A man operates a company in the state
that is engaged in the disposal of nuclear wastes. Subsequent to the
passage of the state statute, the man, not yet aware of the new law,
entered into contracts with many out-of-state firms to dispose of their
nuclear wastes in the state. On account of this new law, however, the
man will be unable to perform these contracts. Assume that the man has
standing to challenge this state law. Which of the following presents his
strongest constitutional grounds to challenge the state law prohibitin
the disposal of nuclear wastes within the state?
A. The commerce clause.
B. The equal protection clause of the Fourteenth Amendment.
C. The privileges and immunities clause of Article IV, Section 2.
D. The contract clause.
Answer:
```

Figure 15: Prompt example for the MMLU dataset.

```
Question: Greenhouses are great for plants like
A. Pizza
B. Lollipops
C. Candles
D. French beans
Answer:
```

Figure 16: Prompt example for the OpenBookQA dataset.

```
Question: To clear snot out of your nose,
A. place a tissue over your nose and blow the snot out.
B. place a tissue over your nose and suck the snot in.
Answer:
```

Figure 17: Prompt example for the PIQA dataset.

```
Passage: Food is very important.Everyone needs to eat well if he wants to
have a strong body.Our minds also need a kind of food.This kind of
food is knowledge .
When we are very young,we start getting knowledge.Kids like watching and
listening.Color pictures especially interest them.When kids are older,
they enjoy reading.When something interests them,they love to ask
questions.
Our minds,like our bodies,always need the best food.Studying on our own
brings the most knowledge.
If someone is always telling us answers,we never learn well.When we study
correctly and get knowledge on our own,we learn more and understand
better.
Question: We start getting knowledge   _  .
A. when we are old
B. when we are very young
C. when we are pupils
D. when we are parents
Answer:
```

Figure 18: Prompt example for the RACE dataset.

```
Question: This is an ancient suit that is not worn with a tie
A. shirt
B. armor
C. helmet
D. shirt and trousers
E. hair
Answer:
```

Figure 19: Prompt example for the RiddleSense dataset.

```
Question: Cameron returned home with a bag of candy to eat all night
long. What will Others want to do next?
A. great
B. buy the candy to eat
C. bored
Answer:
```

Figure 20: Prompt example for the Social IQa dataset.

```
Story: Jon loved the night sky. He would spend many of his nights looking
at the stars. His mom saw that he loved the night sky. His mom bought him
a telescope.
Question: Which sentence best completes the story?
A. Jon then watched germs with his microscope.
B. Jon used his telescope often.
Answer:
```

Figure 21: Prompt example for the StoryCloze dataset.

```
Question: So _ plays video games because Leslie has a lot of free time
while Nelson has to work all the time.
A. Leslie
B. Nelson
Answer:
```

Figure 22: Prompt example for the Winogrande dataset.

B COMPUTATIONAL CONSTRAINTS

While the OpenAI Codex Beta [3] being free enabled the high volume of experiments we performed, we were limited by its maximum 20 API requests per minute limit (which we werent't able to hit in practice). Computing just the zero-shot CP value for MMLU (Hendrycks et al., 2021) in Table 2 took over a week.

C RESULTS UNDER STRONG SHUFFLE OF ANSWER OPTIONS

[3] https://openai.com/blog/openai-codex/

Dataset	N	K	Zero-Shot		One-Shot		Few-Shot	
			No	Shuffle	No	Shuffle	No	Shuffle
AG News	4	38	83.5	—	87.1	—	89.4	—
ANLI R1	3	27	33.2	**38.2**	**61.7**	54.0	64.2	**66.2**
ANLI R2	3	26	33.6	**34.8**	**53.0**	48.1	55.2	**57.1**
ANLI R3	3	26	34.3	**35.4**	47.8	**50.1**	54.5	**56.3**
ARC (Challenge)	4	50	81.7	**82.2**	82.8	**83.0**	**86.1**	85.6
ARC (Easy)	4	57	**93.1**	92.8	**93.5**	93.4	**94.7**	94.4
CODAH	4	63	**76.0**	75.9	**87.8**	87.5	91.9	**92.5**
CommonsenseQA	5	79	**72.0**	71.2	**78.9**	78.0	**83.2**	82.9
COPA	2	113	**89.0**	86.0	**99.0**	96.0	**100.0**	99.0
Cosmos QA	4	24	75.5	**76.3**	81.8	**82.8**	82.4	**83.7**
DREAM	3	7	91.3	**91.8**	93.3	**94.1**	**94.1**	94.0
Fig-QA	2	99	**84.7**	81.8	**86.7**	86.0	**94.0**	92.5
HellaSwag	4	16	**71.0**	70.6	75.1	—	73.6	—
LogiQA	4	16	**44.5**	39.3	**45.3**	42.9	**47.3**	41.6
MedMCQA	4	58	**52.1**	49.4	**53.9**	52.5	**54.4**	51.9
MMLU	4	5	62.1	—	68.2	—	69.5	—
OpenBookQA	4	83	72.0	**74.8**	81.6	**82.0**	**87.0**	**87.0**
PIQA	2	35	73.7	**73.9**	**81.8**	80.5	84.5	**86.2**
RACE-h	4	4	**82.1**	81.5	85.1	**85.3**	86.2	**86.5**
RACE-m	4	8	85.4	**86.3**	**89.3**	89.2	**90.3**	90.1
RiddleSense	5	59	**67.6**	67.3	**77.1**	74.6	**83.9**	82.6
Social IQa	3	72	**64.4**	64.3	**72.2**	71.3	74.9	**75.0**
StoryCloze	2	44	97.5	**98.1**	**98.3**	98.2	98.5	**98.7**
Winogrande (XL)	2	102	**64.5**	59.5	**71.6**	67.8	**72.1**	66.7
Winogrande (XS)	2	102	**64.8**	59.4	**71.3**	66.3	**73.6**	66.1

Table 3: Effect of shuffling answer options on the performance of Codex with MCP. Strong shuffle ensures the index associated with the correct answer choice changes. N is the number of answer options for each dataset. K exemplars are provided in the few-shot setting. Values marked with — could not be computed due to computational restraints (see Appendix B). Note that a slight drop in accuracy when shuffling is to be expected for many of the datasets where answer options refer to ordering (e.g., when an answer option is "both B and D are correct.")

D NON-CHERRY-PICKED MISSED QUESTIONS FROM COMMONSENSEQA

```
Question: James was looking for a good place to buy farmland.  Where
might he look?
A. midwest
B. countryside
C. estate
D. farming areas
E. illinois
Answer:

Question: What would vinyl be an odd thing to replace?
A. pants
B. record albums
C. record store
D. cheese
E. wallpaper
Answer:

Question: Aside from water and nourishment what does your dog need?
A. bone
B. charm
C. petted
D. lots of attention
E. walked
Answer:

Question: Though the thin film seemed fragile, for it's intended purpose
it was actually nearly what?
A. indestructible
B. durable
C. undestroyable
D. indestructible
E. unbreakable
Answer:

Question: What is someone who isn't clever, bright, or competent called?
A. clumsy
B. ineffectual
C. dull
D. clumsy
E. stupid
Answer:

Question: Blue read material outside of his comfort zone because he
wanted to gain what?
A. new perspective
B. entertained
C. understanding
D. hunger
E. tired eyes
Answer:

Question: What must someone do before they shop?
A. get money
B. have money
C. bring cash
D. go to market
E. bring cash
Answer:
```

Figure 23: Examples of CommonsenseQA questions missed by Codex with MCP (not cherry-picked). Answers are A, E, D, D, E, A, and A. Model selections were D, B, E, E, C, C, and B.

E TRADITIONAL PROMPT PERFORMANCE BY NORMALIZATION METHOD

Dataset	N	K	Zero-Shot			One-Shot			Few-Shot		
			Raw	LN	UN	Raw	LN	UN	Raw	LN	UN
AG News	4	38	**68.2**	66.4	44.8	73.0	**77.6**	38.9	**90.1**	89.4	26.4
ANLI R1	3	27	41.2	**45.3**	41.8	**35.6**	35.5	34.5	58.0	**58.4**	34.7
ANLI R2	3	26	36.3	37.9	**39.2**	**35.7**	35.5	34.1	51.4	**51.8**	34.3
ANLI R3	3	26	34.5	**37.8**	37.0	35.2	**35.5**	34.5	**54.2**	54.1	30.4
ARC (Challenge)	4	50	55.7	58.6	**58.9**	61.4	63.7	**64.1**	63.1	**66.6**	64.9
ARC (Easy)	4	57	**84.2**	82.9	78.0	85.8	**85.9**	81.2	87.1	**87.8**	82.2
CODAH	4	63	55.7	**56.8**	55.7	61.1	**65.4**	62.0	69.0	**73.6**	69.9
CommonsenseQA	5	79	65.4	57.9	**68.5**	70.2	70.9	**73.1**	77.1	**78.6**	77.6
COPA	2	113	**92.0**	86.0	90.0	**95.0**	92.0	92.0	**96.0**	**96.0**	**96.0**
Cosmos QA	4	24	32.6	**43.0**	34.6	35.9	**44.0**	36.4	30.1	31.3	**38.1**
DREAM	3	7	**72.7**	71.2	71.5	81.5	**82.5**	80.0	84.2	**84.3**	82.0
Fig-QA	2	99	73.2	74.0	**79.6**	76.8	79.5	**82.4**	79.0	81.8	**82.5**
HellaSwag	4	16	—	—	—	—	—	—	—	—	—
LogiQA	4	16	25.5	30.0	**36.6**	26.1	29.5	**37.5**	25.7	30.9	**37.8**
MedMCQA	4	58	34.1	37.6	**37.8**	37.8	41.1	**42.1**	38.0	**41.2**	**41.2**
MMLU	4	5	46.1	48.9	**49.5**	—	—	—	—	—	—
OpenBookQA	4	83	36.0	47.0	**63.2**	41.8	51.0	**64.0**	46.4	57.0	**71.2**
PIQA	2	35	82.7	**83.7**	68.6	83.2	**84.1**	67.4	84.5	**86.1**	70.7
RACE-h	4	4	48.5	**52.3**	49.3	49.5	**53.2**	52.3	51.3	**55.2**	53.0
RACE-m	4	8	64.2	**67.5**	63.3	67.3	**70.5**	66.0	68.9	**71.7**	67.5
RiddleSense	5	59	**79.8**	68.7	77.4	**89.1**	84.4	86.3	**91.3**	86.4	88.2
Social IQa	3	72	51.1	50.7	**52.1**	53.7	**58.1**	55.5	59.1	**62.4**	58.2
StoryCloze	2	44	76.3	**80.3**	79.5	80.1	**83.4**	82.4	84.7	**88.2**	86.9
Winogrande (XL)	2	102	**62.5**	62.4	60.1	**71.6**	70.8	66.3	**75.5**	**75.5**	68.7
Winogrande (XS)	2	102	**63.0**	62.7	59.8	**71.0**	69.9	64.6	**76.2**	75.8	69.4

Table 4: Effect of normalization strategy on the performance of Codex with cloze prompts. Raw is the strategy of selecting an answer choice based on raw probabilities. LN and UN are selecting an answer choice after normalizing probabilities based on length or unconditional completion probability respectively (see Brown et al. (2020)). N is the number of answer options for each dataset. K exemplars are provided in the few-shot setting. The best strategy for each dataset and exemplar count is bolded. Values marked with — could not be computed due to computational restraints (see Appendix B).

F MEDMCQA TEST PERFORMANCE BY SUBJECT

Subject	Accuracy
Anaesthesia	52.5
Anatomy	47.5
Biochemistry	64.8
Dental	55.9
ENT	57.0
FM	56.8
Medicine	64.5
Microbiology	54.5
O&G	55.5
Ophthalmology	57.6
PSM	60.1
Pathology	65.2
Pediatrics	53.7
Pharmacology	64.4
Physiology	60.6
Psychiatry	33.3
Radiology	63.0
Skin	61.7
Surgery	52.9
Unknown	58.5

Table 5: Test accuracy of Codex with MCP on the MedMCQA test set by subject.

G MMLU PERFORMANCE BY TASK

Subject	Zero-Shot	One-Shot	Five-Shot
abstract_algebra	29.0	29.0	**31.0**
anatomy	62.2	59.3	**65.9**
astronomy	73.0	79.6	**81.6**
business_ethics	69.0	**72.0**	71.0
clinical_knowledge	70.2	70.9	**71.7**
college_biology	75.0	78.5	**81.2**
college_chemistry	**50.0**	45.0	42.0
college_computer_science	53.0	54.0	**57.0**
college_mathematics	35.0	**39.0**	37.0
college_medicine	65.3	71.7	**72.3**
college_physics	42.2	**46.1**	45.1
computer_security	75.0	75.0	**79.0**
conceptual_physics	63.8	65.1	**66.0**
econometrics	44.7	50.0	**50.9**
electrical_engineering	55.9	64.8	**69.7**
elementary_mathematics	48.9	53.4	**54.8**
formal_logic	27.8	49.2	**54.0**
global_facts	34.0	**48.0**	45.0
high_school_biology	80.0	81.3	**85.2**
high_school_chemistry	53.2	**56.2**	55.2
high_school_computer_science	79.0	77.0	**80.0**
high_school_european_history	80.0	85.5	**86.7**
high_school_geography	79.8	**84.8**	84.3
high_school_government_and_politics	89.1	92.2	**93.8**
high_school_macroeconomics	66.7	**73.1**	71.5
high_school_mathematics	37.4	**40.7**	40.4
high_school_microeconomics	69.3	70.6	**74.4**
high_school_physics	39.1	**44.4**	41.7
high_school_psychology	85.9	87.3	**89.9**
high_school_statistics	54.2	59.7	**61.6**
high_school_us_history	83.8	83.3	**87.7**
high_school_world_history	81.0	84.4	**86.1**
human_aging	**74.0**	77.1	**74.0**
human_sexuality	80.9	**81.7**	80.9
international_law	76.9	82.6	**85.1**
jurisprudence	80.6	81.5	**85.2**
logical_fallacies	**81.0**	74.2	79.1
machine_learning	45.5	50.0	**54.5**
management	74.8	83.5	**86.4**
marketing	61.5	87.2	**89.7**
medical_genetics	72.0	69.0	**74.0**
miscellaneous	85.8	85.4	**87.7**
moral_disputes	13.6	75.1	**80.1**
moral_scenarios	24.7	**47.5**	46.3
nutrition	72.2	72.2	**76.5**
philosophy	74.3	74.6	**75.9**
prehistory	74.1	78.7	**81.2**
professional_accounting	**51.8**	50.7	49.3
professional_law	52.2	53.9	**54.8**
professional_medicine	73.5	**73.9**	72.1
professional_psychology	69.9	74.2	**74.5**
public_relations	66.4	70.0	**73.6**
security_studies	67.3	73.5	**75.5**
sociology	80.6	86.6	**87.6**
us_foreign_policy	**87.0**	86.0	**87.0**
virology	50.0	51.8	**54.8**
world_religions	52.0	84.2	**85.4**

Table 6: Test accuracy of Codex with MCP on the MMLU test set by task.