"Let's Argue Both Sides": Argument Generation Can Force Models to Utilize Previously Inaccessible Reasoning Capabilities

Anonymous ACL submission

Abstract

Large Language Models (LLMs), despite achieving state-of-the-art results in a number of evaluation tasks, struggle to maintain their performance when logical reasoning is strictly required to correctly infer a prediction. In this work, we propose *Argument Generation* as a method of forcing models to utilize their reasoning capabilities when other approaches such as chain-of-thought reasoning prove insufficient. Our method involves the generation of arguments for each possible inference result, and asking the end model to rank the generated arguments. We show that *Argument Generation* can serve as an appropriate substitute for zero-shot prompting techniques without the requirement to add layers of complexity. Furthermore, we argue that knowledge-probing techniques such as chain-of-thought reasoning and *Argument Generation* are only useful when further reasoning is required to infer a prediction, making them auxiliary to more common zero-shot approaches. Finally, we demonstrate that our approach forces larger gains in smaller language models, showcasing a complex relationship between model size and prompting methods.

1 Introduction

Large Language Models, including state-of-the-art models such as Llama family of LLMs (Touvron et al., 2023), Mistral 7B (Jiang et al., 2023), and Phi-3 (Abdin et al., 2024) have shown to significantly outperform previous generation of models (Wang et al., 2023b) such as BERT (Devlin et al., 2019) in several mainly classification tasks (Chang et al., 2024). However, despite their seemingly human-like auto-regressive behavior, Large Language Models do not perform well when deep reasoning or analysis is required to effectively infer a prediction (Lee et al., 2023; Tao et al., 2023).

In order to bolster the reasoning capabilities of large language models, the research community has done extensive recent work in the form of chain-of-thought reasoning (Kojima et al., 2022; Wang et al., 2023a), Self-Reflection (Madaan et al., 2023), Multi-Agent Debate (Liang et al., 2023; Du et al., 2023), and Socratic prompting (Chang, 2023), demonstrating that prompting the model to generate the reasoning behind its answer, or generating a step-by-step guide to reach its response can help predict better results.

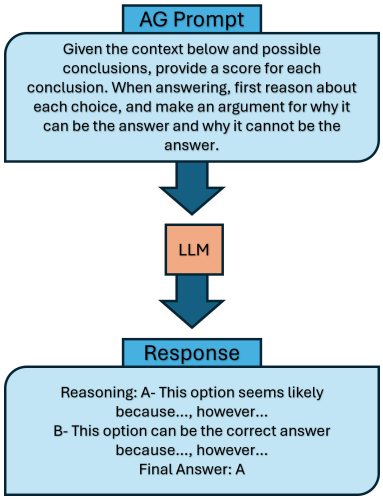

Figure 1: The general framework of Argument Generation Prompting

Taking inspiration from chain-of-thought reasoning, and motivated by the need to develop better prompt techniques with the goal of increasing model performance in reasoning tasks, we introduce *Argument Generation*, a single-pass prompting technique that aims to utilize the reasoning and argumentation capabilities of Large Language Models to generate better responses where deeper consideration of logic or reasoning is required to infer the correct result. *Argument Generation* involves a two-step process. We first prompt the model to generate possible reasoning for the truthfulness of each possible option, and then we ask the model to rank the generated arguments and map

its ranking to a final output in accordance with the task expectations.

We evaluate our method on a number of openly available state-of-the-art Large Language Models using five tasks of different natures. We find that *Argument Generation* at its weakest, does not perform significantly worse than chain-of-thought reasoning, and is able to significantly outperform both zero-shot reasoning and chain-of-thought reasoning when a deeper understanding of the task options is required. Furthermore, we note that in comparison to chain-of-thought reasoning, *Argument Generation* can be used as a stronger knowledge probing technique that is useful in instances where such probing is essential. However, our method does not necessarily increase the model performance for inputs that observe acceptable results under more common methods.

We make the following contributions: (1) We introduce *Argument Generation*, a novel prompting technique that aims to access the underlying reasoning capabilities of LLMs. (2) We show through a series of experiments that our method is able to effectively reason under conditions that fail chain-of-thought reasoning. (3) We show that our prompting method is more effective when used with smaller language models, eliciting further investigation into the relationship between prompting and model capabilities.

2 Background and Motivation

Argumentation is the cognitive capability of generating and evaluating "reasons" for deriving a conclusion (Mercier, 2016). It is a central aspect of human intelligence and is omnipresent in natural human communication. It extends the conception of reasoning in LLM-research (Yu et al., 2023a) by including the notion that conclusions drawn must be new. Indeed, it has been suggested that human reasoning evolved for the purposes of enabling humans to persuade each other (Mercier and Sperber, 2011) through arguments.

We hypothesize that many day-to-day arguments are evaluated by humans in an intuitive (fast, system 1) manner, without deep thought or "epistemic vigilance" (Sperber et al., 2010), unless they are from trusted sources and appear to contradict our own beliefs. Thus, because LLMs were pretrained with human communicative interactions, we hypothesize that LLMs are capable of fast argumentative thinking. By triggering argumentative thought, we hypothesize that LLMs can effectively generate reasons and assess conclusions, as well as improve core reasoning capabilities across a variety of domains, including commonsense, logical, and social.

3 Related Work

General argumentation ability of LLMs have begun to be explored by researchers, with a focus on a number of computational argumentation subtasks such as argument mining, claim detection, evidence detection and type classification, argument generation, and summarization (Balikas, 2023; Chen et al., 2023; Holtermann et al., 2022; Ruiz-Dolz and Lawrence, 2023; Thorburn and Kruger; de Wynter and Yuan, 2023). Research suggests that LLMs "exhibit commendable performance" (Chen et al., 2023) in zero-shot and few-shot settings thereby supplying a foundation supporting our approach.

Delving deeper, we can explore two core aspects of argumentation. First, the ability to argue for/against all sides (thinking like a lawyer). Second, the ability to generate implicit assumptions (necessary or sufficient warrants) needed to support the argument.

Arguing all sides is related to "backward reasoning" suggested in (Yu et al., 2023a), where they discuss that it is "better to collect both supportive and opposing knowledge to compare the confidence of different conclusions for defeasible reasoning." Additionally (Wang et al., 2022) discuss the idea of allowing several different reasoning paths and choosing the "most consistent one". Another approach is contrastive chain-of-thought (Chia et al., 2023) where they consider both valid and invalid reasoning demonstrations alongside original prompt – a dual perspective approach. Additionally, work in multiagent debate, for example (Chia et al., 2023) uses a notion of a debate with multiple agents discussing and talking about the problem. However, none of these approaches attempt at *rationalizing* all sides of an argument. That is none of these offer up the best possible argument for/against each choice, and then evaluate the best argument (like for example, anticipatory reflection of plans in (Wang et al., 2024)).

Extracting implicit information relates to work in "knowledge-enhanced" (Qiao et al., 2023) strategies in which an implicit model generates knowledge and rationales. Also Yu et al. (2023a) dis-

cusses Leap-of-thought reasoning which uses implicit facts to answer questions. A related notion is that of decomposing implicit multi-hop questions down in connection with the general backward reasoning tactic of question-decomposition (see summary in (Yu et al., 2023a)). Work by (Sarathy et al., 2022) suggests extracting implicit assumptions from premise-conclusion pairs, however, that work does not explore how such endeavor influences an LLM's reasoning capability. Although there is a growing body of work in question decomposition, it is unclear to what extent they take implicit assumptions into account.

General LLM reasoning capabilities have been improving over the past several years with numerous datasets targeting different types of reasoning – logical, mathematical, commonsense, argumentation, and social reasoning (Qiao et al., 2023; Yu et al., 2023a; Huang and Chang, 2023; Yu et al., 2023b; Luo et al., 2023; Sahoo et al., 2024a). The methods have involved various techniques to evoke reasoning processes such as having the LLM explicate its chain of thought (Wei et al., 2022a), reflect on its own reasoning process (Wang and Zhao, 2023), decompose complex reasoning processes into simpler problems that can be solved more easily (Khot et al., 2023), explore many different reasoning paths and decide on one that wins a majority vote (Wang et al., 2022), and others. These various methods have shown improvements in various reasoning tasks, but none have shown cross-domain effectiveness. Moreover, their reasoning capabilities are limited when exposed to scenarios in which the model must resolve a disagreement (Lee et al., 2023), distinguish a correct phrase from an incorrect one (Riccardi and Desai, 2023), or assign a nondeterministic gender to a subject (Zakizadeh et al., 2023). Overall, Large Language Models have shown promising results in a variety of reasoning tasks while serious challenges and shortcomings still remain (Chang et al., 2024). What is missing is a cross-domain strategy to improve an LLM's zero-shot reasoning capability, which we believe to be enhanced by its latent capability for argumentative thinking.

4 Methodology

We now provide details regarding our approach, including the proposed zero-shot approach and the reasoning behind our choice of *Argument Generation* as a prompting technique.

Argument Generation involves two overall steps. Given an initial input x with possible answers $k_1, k_2, ...k_n$, we first prompt the model to generate arguments supporting and attacking each answer k_i, creating arguments $x'_1, x'_2, ...x'_n$ for each possible answer. We then ask the model to choose the answer with the strongest argument as the final output. More concretely, the Large Language Model is utilized as a proxy for an argument ranking function that chooses the most feasible options among arguments $x'_1, x'_2, ...x'_n$.

The rationale behind our approach is twofold. First, it has been shown that Large Language Models, when provided with a reasoning context towards the correct output, observe significantly improved performance (Wei et al., 2022b; Kojima et al., 2022), making the reasoning behind each choice an important contributor to model performance. Second, Large Language Models can act as effective rankers when provided with a list-wise input of possible options (Ma et al., 2023), indicating the feasibility of their possible utilization for the effective ranking of arguments. As a result, the proposed technique relies on the assumption that the correct answer k_i to the query x should logically have the strongest argument supporting it, forcing the ranker model to choose the argument that is directly mapped to the correct answer.

5 Evaluation

To empirically evaluate the effectiveness of our proposed method, we have tested the performance of *Argument Generation* in five datasets and across six models. For the remainder of this section, we focus on describing our evaluation setting.

5.1 Models

We test our approach using six models, including two families of models, and two independent, recently released LLMs. These include Llama 3 family of models[1] (8B and 70B), Gemma family of models (2B and 7B) (Mesnard et al., 2024), Phi-3 3.8B (Abdin et al., 2024), and Mistral 7B (Jiang et al., 2023).

5.2 Datasets

Our choice of datasets includes candidates from five different tasks, each representing a group of tasks that aim to quantify a specific aspect of a given model. We strive to cover tasks belonging to

[1] https://ai.meta.com/blog/meta-llama-3/

different domains, including question-answering, argumentation, reasoning, and bias evaluation.

CommonSenseQA (Talmor et al., 2019) includes multiple choice questions that quantify the common sense reasoning of the LLM that is being tested. The goal of these questions is to evaluate the capabilities of a model in connecting a question with complex semantics to a likely answer. We report the simple accuracy score for CommonSenseQA.

DiFair (Zakizadeh et al., 2023) is designed as a gender bias quantification dataset and tasks the model with the determination of gender for a given noun/pronoun in a setting where the presence of sufficient information for gender assignment is not guaranteed. We report the Gender Invariance score as defined by the paper.

IBM-30K (Gretz et al., 2020) contains 30,000 topic-argument pairs with the goal of determining if the argument is sufficiently valid in a given topic. They utilize the weighted average of the annotator score as the gold label and report the ranking correlation against the gold label. In place of the original metric, we report $1 - MAE$ as the main score to be consistent with other metrics and to showcase the model response quality per individual instance.

TruthfulQA (Lin et al., 2022) measures the model truthfulness by quantifying the model likelihood of mimicking human falsehoods as seen in text. We utilize the MC1 task definition of TruthfulQA which involves 880 multiple-choice questions designed to surface the model misinformation. We additionally generate 60 questions by randomly sampling $\%15$ of the original dataset and replacing the correct option with 'None of the Answers are Correct'. This is done in order to further evaluate model performance when no clear answer exists. We report the accuracy score of the generated dataset.

StereoSet (Nadeem et al., 2021) strives to measure the stereotypical bias inherent in language models by requiring the model to select among stereotypical, anti-stereotypical, and unrelated options given a context sentence. They argue that an ideal model should never choose an unrelated option as it directly contradicts the language modeling capabilities of the model. In addition, an ideally fair model should rank a stereotypical answer over an anti-stereotypical one only 50% of the time, demonstrating the fact that it does not prefer a commonly associated stereotypical or anti-stereotypical view of the context. We run our tests on the subset of the dataset that targets racial and gender bias and report the Idealized Cat (icat) score of this subset.

5.3 Argument Generation

Algorithm 1 Argument Generation

Require: Input x, List of Possible Answers K
Ensure: Final Response k_i

procedure GENERATION(x, K)
  function IMPLICITASSUMPTION(x, K)
    Let $A := \emptyset$
    for all $k_i \in K$ **do**
      $A := A \cup$ ASSUMPTION(x, k_i)
    Let $Ranking :=$ RANKING(A)
    return $Ranking[0]$ ▷ Return the Top Ranking Answer
  function ARGUMENTGENERATION(x, K)
    Let $A := \emptyset$
    for all $k_i \in K$ **do**
      $A := A \cup \{$ARGUMENT(x, k_i), ARGUMENT(x, $\neg k_i$)$\}$
    Let $Ranking :=$ LWR(A)
    return $Ranking[0]$ ▷ Return the Top Ranking Answer

We perform our evaluations using two different *Argument Generation* settings. In the first approach, given an input x and a possible answer k, we explicitly ask the model to generate an **implicit assumption** under which k is a valid response to x. An implicit assumption is a set of logical propositions P such that every proposition in P must hold in order for the answer to follow logically from x. We then ask the model to rank these implicit assumptions by the feasibility of all $p_i \in P$ to hold simultaneously. We finally take the implicit assumption with the highest feasibility ranking as the final answer to the input.

In the second approach, given an input x and a possible answer k, we ask the model to both generate an argument for accepting k as a correct answer to x and generate an argument for rejecting k as a correct answer to x. We then apply this process to all candidate answers k_1 through k_n such that n tuples of arguments are generated by the model. We finally prompt the model to rank these tuples and generate the final answer to input

Model	Prompt	CommonenseQA	DiFair	IBM-30K	TruthfulQA	StereoSet
Gemma 7B	Zero-Shot	**43.24**%	0.0%	59.46%	20.63%	**63.70**%
	Chain of Thought	41.85%	12.65%	49.98%	18.61%	36.17%
	Argument Generation w/ Implicit Assumptions	37.18%	34.54%	62.63%	**47.97**%	44.97%
	Argument Generation	39.80%	**55.39**%	**80.93**%	31.27%	34.3%
Gemma 7B	Zero-Shot	**69.28**%	0.0%	70.85%	28.93%	**88.87**%
	Chain of Thought	69.12%	32.52%	63.14%	**41.48**%	66.98%
	Argument Generation w/ Implicit Assumptions	66.33%	47.51%	69.31%	33.05%	64.05%
	Argument Generation	66.66%	**55.84**%	**72.94**%	25.21%	73.88%
Llama3 8B	Zero-Shot	71.33%	22.19%	60.51%	47.97%	42.47%
	Chain of Thought	**71.41**%	10.80%	66.03%	44.57%	54.36%
	Argument Generation w/ Implicit Assumptions	63.22%	55.88%	71.22%	**51.70**%	**55.73**%
	Argument Generation	64.12%	**58.57**%	**73.50**%	33.93%	45.90%
Llama3 70B	Zero-Shot	79.85%	78.08%	76.04%	69.04%	41.91%
	Chain of Thought	**80.26**%	**82.79**%	64.46%	**70.53%**	39.04%
	Argument Generation w/ Implicit Assumptions	74.44%	72.45%	**76.98**%	56.91%	**73.44**%
	Argument Generation	75.34%	79.16%	76.13%	68.93%	52.05%
Phi3 3.8B	Zero-Shot	**67.97**%	6%	63.04%	47.55%	56.0%
	Chain of Thought	66.66%	**71.59**%	62.57%	51.48%	61.52%
	Argument Generation w/ Implicit Assumptions	66.91%	57.24%	**69.50**%	51.70%	61.15%
	Argument Generation	**67.97**%	52.39%	69.17%	**52.12**%	**61.67**%
Mistral 7B	Zero-Shot	67.81%	45.66%	64.83%	8%	46.61%
	Chain of Thought	**67.89**%	62.19%	59.82%	**55.95**%	41.10%
	Argument Generation w/ Implicit Assumptions	64.29%	63.44%	66.58%	50.63%	46.28%
	Argument Generation	64.70%	**66.51**%	**66.85**%	51.27%	**49.24**%

Table 1: Prompting results using Argument Generation, Chain of Thought Reasoning, and Zero-Shot Prompting in five different datasets.

x.

Algorithm 1 showcases both of the aforementioned techniques, where ASSUMPTION(x, K) refers to the generation of implicit assumptions for each candidate answer, and ranking them via a list-wise ranking technique, and ARGUMENT(x, K) refers to the generation of tuples of arguments for each candidate answer that both support and attack the corresponding candidate answer, and then ranking them via a list-wise ranking approach.

We acknowledge that it is possible to extend our approach to a multi-agent setting, where the argument generation is done by an external model that is separate from the ranking model. However, we focus on single-pass prompting for the purpose of this study to (i) provide a single-pass, easy-to-implement approach that is comparable to zero-shot chain-of-thought reasoning both in performance, and running time, and (ii) refrain from unnecessarily increasing the computational requirement of the approach, as seen in other multi-agent techniques. However, we hypothesize that generalizing our algorithm to utilize multiple agents is both simple and observes an increase in performance.

6 Evaluation Results

We now showcase our results as tested against the datasets mentioned in section 5. We additionally show that *Argument Generation*, when outperforming zero-shot chain-of-thought reasoning, demonstrates significantly higher performance gain, and suffers smaller losses in cases where it does not result in increased performance. We finally provide a model size analysis to better understand the relationship between prompting methods and the number of parameters present in a given Large Language Model.

6.1 Performance Analysis

Table 1 showcases the evaluation results when using *Argument Generation* against zero-shot chain-of-thought prompting (Kojima et al., 2022) and common zero-shot prompting (Radford et al., 2019).

We observe that our method is able to outperform both zero-shot prompting and chain-of-thought reasoning in 18 of the 30 test settings, amounting to a win rate of 60%. Additionally, our approach outperforms chain-of-thought reasoning in 20 of the 30 settings, showcasing that *Argument Generation* yields better results in 66.66% of the test cases. Among the 20 cases where our proposed method performs better, there are 15 cases (75%) in which both proposed approaches outperform chain-of-thought reasoning, while *Argument Generation* with implicit assumptions is able to yield better re-

sults in 18 cases (90%), and *Argument Generation* without implicit assumptions has a better performance in 17 cases (85%), demonstrating that both methods have better overall results in comparison to zero-shot chain-of-thought reasoning.

With respect to individual datasets, we find that our method enjoys a significant performance boost when tested against instances of IBM-30K (Gretz et al., 2020), with both methods showing improved results over the two other baselines in all models. This behavior is expected as IBM-30K measures a model's capability to correctly discern a valid argument from an invalid one, and our approach operates via generating arguments that both support and attack the given input, meaning that invalid arguments will have weaker support, allowing the model to correctly rank the inputs based on their strength.

Additionally, we observe that *Argument Generation* is able to increase model performance for 8 out of 12 instances (66.66%) against all methods, and for 10 out of 12 instances (83.33%) against chain-of-thought reasoning in DiFair (Zakizadeh et al., 2023) and StereoSet (Nadeem et al., 2021) datasets, showcasing that argumentation might serve as a reliable debiasing method for Large Language Models. Interestingly, the correlation between our approach's improving effects and a given model's general capability is not strictly positive in this case, meaning that it is possible for larger models to observe lower, or no gains when prompted with *Argument Generation*. We attribute this observation to the possibility of more capable models deceiving themselves via supporting an incorrect candidate when the initial knowledge is sufficient to make a prediction, meaning that *Argument Generation* might force an artificial and unwanted decrease in model confidence. We provide further details and analysis in section 6.3.

6.2 Performance Difference Analysis

In order to observe the expected performance metric difference, we define Δ_{min} as the mean difference between chain-of-thought reasoning and the worst-performing *Argument Generation* method when chain-of-thought reasoning is performing better than our approach, and Δ_{max} as the mean difference between chain-of-thought reasoning and the best-performing *Argument Generation* method when chain-of-thought reasoning is performing better than our approach. Conversely, we define Γ_{min} and Γ_{max} similarly for cases in which *Argument Generation* is performing better than chain-of-thought reasoning. More concretely, Δ values show the performance decrease of *Argument Generation* with respect to chain-of-thought reasoning when the second approach is able to outperform our method, while Γ values demonstrate the performance increase when *Argument Generation* produces better results in comparison to chain-of-thought reasoning.

Model Name	Δ_{min}	Δ_{max}	Γ_{min}	Γ_{max}
Gemma 2B	4.67	2.05	**15.73**	**34.35**
Gemma 7B	9.53	5.44	**10.57**	**16.56**
Llama3 8B	8.18	7.28	**25.13**	**27.61**
Llama3 70B	9.92	3.38	**12.34**	**23.46**
Phi3 3.8B	**19.2**	**14.35**	2.35	2.96
Mistral 7B	4.45	3.93	**4.39**	**6.49**
Overall	8.98	5.26	**10.90**	**17.77**

Table 2: Observed results of Δ_{min}, Δ_{max}, Γ_{min}, and Γ_{max} for all tested models. We find that in cases where our method performs better, it generally holds that it has a larger performance gain in comparison to the cases where Chain-of-Thought reasoning outperforms our approach.

Table 2 showcases our empirical results. We find that except for the Phi3 3.8B model, all LLMs demonstrate significantly higher performance in instances where our method outperforms zero-shot chain-of-thought reasoning. Most significantly, Llama3 8B has a mean performance difference of 25.13% between the worst-performing *Argument Generation* approach and zero-shot chain-of-thought reasoning (Γ_{min}) in tasks that our method performs better. Looking at Γ_{max}, the best-performing proposed method is able to boost Gemma 2B model performance by 34.35%, and Llama3 8B performance by 27.61%, showcasing that overall when such an increase in model performance is observed, the increase is significant. Conversely, Phi3 3.8B, when prompted using our method, only has an increased output value of 2.96% at best, while performing 19.2% better than the worst-performing *Argument Generation* approach, and 14.35% better than the best-performing approach in instances that chain-of-thought reasoning yields better results. We attribute this behavior to the model's lower argument ranking capabilities, meaning that Phi3 cannot effectively rank the arguments based on their validity. This notion is further bolstered by the model's relatively low per-

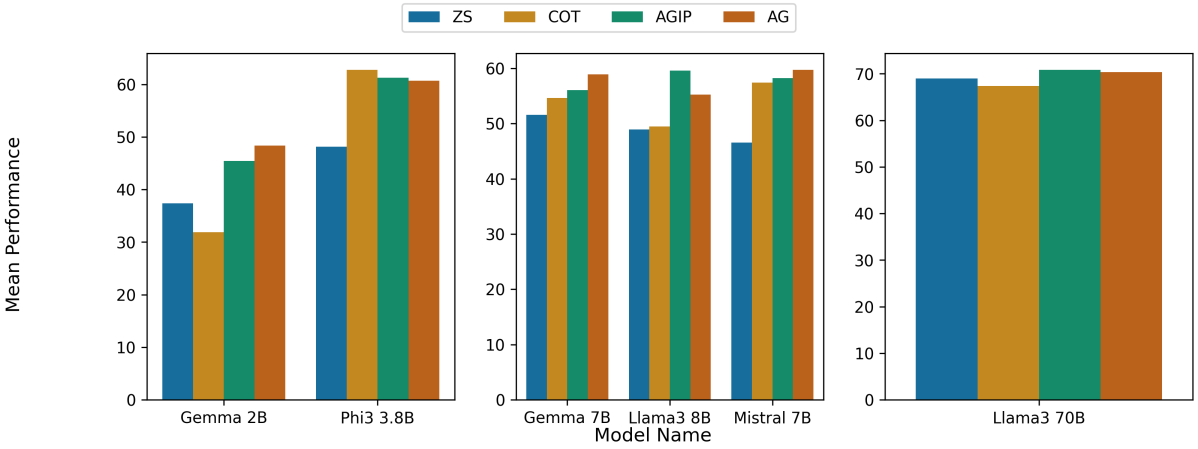

Figure 2: Mean Performance in models of different size

formance in the IBM-30K dataset as seen in Table 1. Additionally, Phi family of models enjoy a significant performance boost when paired with chain-of-thought reasoning [2], which we believe contributes to the observation that our approach does not significantly increase the model performance in this instance relative to other models.

6.3 Model Size Analysis

We now provide our results on the effects of prompting on models of different sizes. In order to conduct our evaluation, we divide the models under test into three subcategories. The first category constitutes Gemma 2B and Phi3 3.8B and is demonstrative of small language models (below 7 billion parameters). The second category contains Gemma 7B, Llama3 8B, and Mistral 7B, and showcases language models of medium size. Finally, Llama 3 70B is the sole member of the third category and acts as a sample member for the largest of language models by parameter count.

Figure 2 demonstrates the mean performance of the four prompting methods across different sizes, grouped by the aforementioned categorization where ZS, COT, AGIP, and AG stand for zero-shot, chain-of-thought, Argument Generation with Implicit Assumptions, and Argument Generation, respectively. Our findings show that generally, all models experience a performance increase when prompted either with chain-of-thought reasoning, or *Argument Generation* with the exception of Gemma 2B and Llama3 70B for COT. We observe that models of smaller sizes (medium and small) experience a significant performance boost when prompted via *Argument Generation* (for 100% of the models) and chain-of-thought reasoning (for 80% of the models).

Furthermore, smaller models show a higher performance gain when compared to the largest Llama 3 instance. More specifically, the mean performance gain when utilizing *Argument Generation* compared to zero-shot prompting is 12.06% for small models, and 10.36% for medium models, while the performance gain for the 70B model is 1.86%. We hypothesize that the reason behind the lower performance gain in larger models is due to their already impressive capability to infer the correct results without the requirement to introduce further information probing techniques such as chain-of-thought reasoning and *Argument Generation*. More concretely, forcing the model to perform self-reasoning or rank the validity of arguments and responses does not expose the model to previously hidden information, and does not necessarily increase the performance when additional information is not strictly required to respond to the input. This phenomenon is especially observable in CommonSenseQA and TruthfulQA as seen in table 1, where the introduction of prompting does not improve the model performance in all instances. These observations are in line with those reported by Kojima et al. (2022) and lead us to believe that knowledge probing prompting methods are only useful in cases where this additional information is required to make strong predictions.

To further investigate the effects of prompting on model performance, and its relationship with the number of model parameters, we report the mean performance across the number of parameters in figure 3. We find that although both our proposed method and chain-of-thought reasoning

[2] Open COT Leaderboard

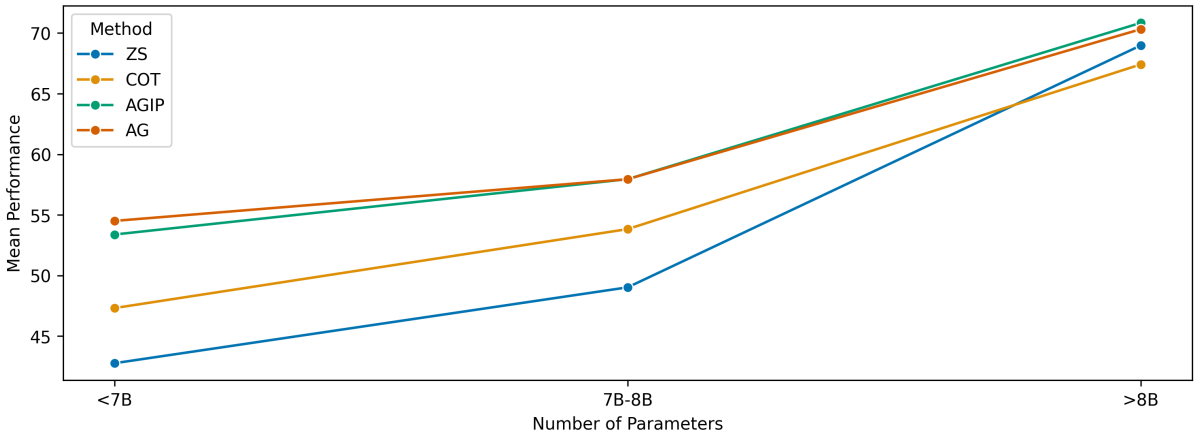

Figure 3: Mean Performance trend across model parameters

provide improved performance in models of larger size, their impact diminishes as the models grow larger. More specifically, we find that the mean difference between zero-shot prompting and *Argument Generation* methods is 11.81% for models with less than 7 billion parameters, 8.92% for models of 7 billion to 8 billion parameters, and 1.59% for the largest model. Further investigation is required to fully confirm our observations, however, this finding bolsters the previous hypothesis that *Argument Generation* as a prompting technique, is more effective in increasing the performance of smaller models. This behavior may stem from the fact that large models are able to generate convincing arguments for incorrect options, making the task of discerning an invalid argument from a valid one difficult. Conversely, smaller models are not able to generate arguments of high quality for incorrect candidates, thus goading the model to rank the valid argument over the incorrect one. Similarly, the observed mean differences between *Argument Generation* and chain-of-thought reasoning are 6.63%, 4.12%, and 3.16% respectively for models of small (<7B), medium (7B and 8B), and large (>8B) sizes.

Based on the above observation, a multi-agent technique to increase performance might be to generate arguments using a less capable model, while utilizing a more performant model to rank the arguments. We delegate further analysis to future work.

7 Discussion and Future Work

Prompting has been proposed as a method of improving model performance in either task-specific settings or broader, task-agnostic environments (Sahoo et al., 2024b). Despite the visible gains of employing prompting to yield better model results, the literature showcasing how, and when prompting works is limited (Petrov et al., 2024). We observe that the proposed method is able to significantly boost the model performance in smaller models while gaining marginal improvements as the model size increases, which is contrary to the previous work showing that larger models have higher gains through prompting (Wei et al., 2022b). This leads us to believe that the relationship between prompting and the nature of the model is complex, and might be affected both by the model size, and its relative task-specific knowledge and capabilities. Further work is required to demonstrate the effects of prompting when models hold knowledge of varying degrees with respect to a task description. Investigation of the learning resources used in model training can provide invaluable insight into the relationship between prompting and model reasoning.

8 Conclusion

In this work, we have proposed *Argument Generation* as a novel, zero-shot prompting technique. Through empirical evaluation using a number of datasets, we observe that our method is able to outperform both zero-shot prompting and zero-shot chain-of-thought reasoning in the majority of the conducted tests, making it a likely candidate when improving the model performance in a zero-shot setting is required. Furthermore, we show that our approach yields larger gains in smaller models, both offering an effective method that can be used in small models and providing a possible future direction to better understand the relationship between model capabilities and prompting.

9 Limitations

Despite the observation that *Argument Generation* is able to generally outperform other common zero-shot prompting methods, its reliance on the existence of a predefined number of options from which the model can arguments is an inherent limitation of our work. While it is true that all questions can be modified to behave as either a multi-choice question or a yes-no question, this conversion relies on the background knowledge of the user that is interacting with the model, meaning that in cases where the user has no information regarding the possible answer for an open question, the correct formulation of the input to fit our criteria can prove difficult. We plan to address this limitation in future work via the possible introduction of automatic option generation.

In addition, while we have made the best effort to cover datasets pertaining to different tasks that evaluate various model capabilities, it is possible that other task-agnostic prompting methods outperform our approach in a number of yet untested metrics. Further investigation is required to fully confirm the effects of our approach on different models and tasks.

10 Ethical Considerations

Previous work has shown that Large Language Models are limited in their capability to understand their own lack of knowledge (Yin et al., 2023). As such, it is possible to generate prompts that exacerbate model hallucinations, and even force models to generate misinformation. The proposed method can especially be prone to attacks of a similar kind as a malicious agent can force the model to showcase generally unwanted behavior by providing the model with incorrect, and even dangerous options. As such, we encourage the research community to continue the work in hallucination reduction and use all prompting methods both responsibly and skeptically.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Qin Cai, Martin Cai, Caio César Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Yen-Chun Chen, Yi-Ling Chen, Parul Chopra, Xiyang Dai, Allie Del Giorno, Gustavo de Rosa, Matthew Dixon, Ronen Eldan, Victor Fragoso, Dan Iter, Mei Gao, Min Gao, Jianfeng Gao, Amit Garg, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Jamie Huynh, Mojan Javaheripi, Xin Jin, Piero Kauffmann, Nikos Karampatziakis, Dongwoo Kim, Mahoud Khademi, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Ce Liu, Mengchen Liu, Weishung Liu, Eric Lin, Zeqi Lin, Chong Luo, Piyush Madan, Matt Mazzola, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Xin Wang, Lijuan Wang, Chunyu Wang, Yu Wang, Rachel Ward, Guanhua Wang, Philipp Witte, Haiping Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Sonali Yadav, Fan Yang, Jianwei Yang, Ziyi Yang, Yifan Yang, Donghan Yu, Lu Yuan, Chengruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

Georgios Balikas. 2023. John-arthur at semeval-2023 task 4: Fine-tuning large language models for arguments classification. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, page 1428–1432, Toronto, Canada. Association for Computational Linguistics.

Edward Y. Chang. 2023. Prompting large language models with the socratic method. In *2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC)*, pages 0351–0360.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.*, 15(3).

Guizhen Chen, Liying Cheng, Luu Anh Tuan, and Lidong Bing. 2023. Exploring the potential of large language models in computational argumentation. (arXiv:2311.09022). ArXiv:2311.09022 [cs].

Yew Ken Chia, Guizhen Chen, Luu Anh Tuan, Soujanya Poria, and Lidong Bing. 2023. Contrastive chain-of-thought prompting. (arXiv:2311.09277). ArXiv:2311.09277 [cs].

Adrian de Wynter and Tommy Yuan. 2023. I wish to have an argument: Argumentative reasoning in large language models. (arXiv:2309.16938). ArXiv:2309.16938 [cs].

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*.

Shai Gretz, Roni Friedman, Edo Cohen-Karlik, Assaf Toledo, Dan Lahav, Ranit Aharonov, and Noam Slonim. 2020. A large-scale dataset for argument quality ranking: Construction and analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7805–7813.

Carolin Holtermann, Anne Lauscher, and Simone Paolo Ponzetto. 2022. Fair and argumentative language modeling for computational argumentation. (arXiv:2204.04026). ArXiv:2204.04026 [cs].

Jie Huang and Kevin Chen-Chuan Chang. 2023. Towards reasoning in large language models: A survey. (arXiv:2212.10403). ArXiv:2212.10403 [cs].

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. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. 2023. Decomposed prompting: A modular approach for solving complex tasks. (arXiv:2210.02406). ArXiv:2210.02406 [cs].

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Noah Lee, Na Min An, and James Thorne. 2023. Can large language models capture dissenting human voices? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4569–4585, Singapore. Association for Computational Linguistics.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. 2023. Encouraging divergent thinking in large language models through multi-agent debate. *Preprint*, arXiv:2305.19118.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Man Luo, Shrinidhi Kumbhar, Ming shen, Mihir Parmar, Neeraj Varshney, Pratyay Banerjee, Somak Aditya, and Chitta Baral. 2023. Towards logiglue: A brief survey and a benchmark for analyzing logical reasoning capabilities of language models. (arXiv:2310.00836). ArXiv:2310.00836 [cs].

Xueguang Ma, Xinyu Crystina Zhang, Ronak Pradeep, and Jimmy J. Lin. 2023. Zero-shot listwise document reranking with a large language model. *ArXiv*, abs/2305.02156.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, pages 46534–46594. Curran Associates, Inc.

Hugo Mercier. 2016. The argumentative theory: Predictions and empirical evidence. *Trends in Cognitive Sciences*, 20(9):689–700.

Hugo Mercier and Dan Sperber. 2011. Why do humans reason? arguments for an argumentative theory. *Behavioral and Brain Sciences*, 34(2):57–74.

Gemma Team Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, L. Sifre, Morgane Riviere, Mihir Kale, J Christopher Love, Pouya Dehghani Tafti, L'eonard Hussenot, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Am'elie H'eliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Cl'ement Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikula, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Pier Giuseppe Sessa, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vladimir Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Brian Warkentin, Ludovic Peran,

Minh Giang, Cl'ement Farabet, Oriol Vinyals, Jeffrey Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024. Gemma: Open models based on gemini research and technology. *ArXiv*, abs/2403.08295.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online. Association for Computational Linguistics.

Aleksandar Petrov, Philip H. S. Torr, and Adel Bibi. 2024. When do prompting and prefix-tuning work? a theory of capabilities and limitations. *Preprint*, arXiv:2310.19698.

Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, and Huajun Chen. 2023. Reasoning with language model prompting: A survey. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5368–5393, Toronto, Canada. Association for Computational Linguistics.

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

Nicholas Riccardi and Rutvik H. Desai. 2023. The two word test: A semantic benchmark for large language models. *Preprint*, arXiv:2306.04610.

Babak Rokh, Ali Azarpeyvand, and Alireza Khanteymoori. 2023. A comprehensive survey on model quantization for deep neural networks in image classification. *ACM Trans. Intell. Syst. Technol.*, 14(6).

Ramon Ruiz-Dolz and John Lawrence. 2023. Detecting argumentative fallacies in the wild: Problems and limitations of large language models. In *Proceedings of the 10th Workshop on Argument Mining*, page 1–10, Singapore. Association for Computational Linguistics.

Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. 2024a. A systematic survey of prompt engineering in large language models: Techniques and applications. (arXiv:2402.07927). ArXiv:2402.07927 [cs].

Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. 2024b. A systematic survey of prompt engineering in large language models: Techniques and applications. *Preprint*, arXiv:2402.07927.

Vasanth Sarathy, Mark Burstein, Scott Friedman, Robert Bobrow, and Ugur Kuter. 2022. A neuro-symbolic cognitive system for intuitive argumentation. In *Advances in Cognitive Systems (ACS)*.

Dan Sperber, Fabrice Clement, Christophe Heintz, Olivier Mascaro, Hugo Mercier, Gloria Origgi, and Deirdre Wilson. 2010. Epistemic vigilance. *Mind & Language*, 25(4):359–393.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. 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)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Zhengwei Tao, Zhi Jin, Xiaoying Bai, Haiyan Zhao, Yanlin Feng, Jia Li, and Wenpeng Hu. 2023. Eveval: A comprehensive evaluation of event semantics for large language models. *Preprint*, arXiv:2305.15268.

Luke Thorburn and Ariel Kruger. Optimizing language models for argumentative reasoning.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Haoyu Wang, Tao Li, Zhiwei Deng, Dan Roth, and Yang Li. 2024. Devil's advocate: Anticipatory reflection for llm agents. (arXiv:2405.16334). ArXiv:2405.16334 [cs].

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

Xuezhi Wang, Jason Wei, Dale Schuurmans, and Quoc V Le. 2023a. H. chi, sharan narang, aakanksha chowdhery, and denny zhou. self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, volume 1.

Yuqing Wang and Yun Zhao. 2023. Metacognitive prompting improves understanding in large language models. (arXiv:2308.05342). ArXiv:2308.05342 [cs].

Zengzhi Wang, Qiming Xie, Zixiang Ding, Yi Feng, and Rui Xia. 2023b. Is chatgpt a good sentiment analyzer? a preliminary study. *arXiv preprint*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022a. Chain of thought prompting elicits reasoning in large language models. (arXiv:2201.11903). ArXiv:2201.11903 [cs].

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022b. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. 2023. Do large language models know what they don't know? In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8653–8665, Toronto, Canada. Association for Computational Linguistics.

Fei Yu, Hongbo Zhang, Prayag Tiwari, and Benyou Wang. 2023a. Natural language reasoning, a survey. (arXiv:2303.14725). ArXiv:2303.14725 [cs].

Zihan Yu, Liang He, Zhen Wu, Xinyu Dai, and Jiajun Chen. 2023b. Towards better chain-of-thought prompting strategies: A survey. (arXiv:2310.04959). ArXiv:2310.04959 [cs].

Mahdi Zakizadeh, Kaveh Miandoab, and Mohammad Pilehvar. 2023. DiFair: A benchmark for disentangled assessment of gender knowledge and bias. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1897–1914, Singapore. Association for Computational Linguistics.

A Model Details

We utilize the Ollama framework [3] to conduct all evaluations described in the paper. Generally, we make use of the 4-bit quantized (Rokh et al., 2023) versions of the tested models to maintain consistency, and due to hardware limitations. Table 3 demonstrates all the tested models, their Ollama hub links, as well as their quantization methods.

Model Name	Hub Link	Quantization Method
Gemma 2B	Link	Q4
Gemma 7B	Link	Q4
Llama3 7B	Link	Q4
Llama3 80B	Link	Q4
Phi3 3.8B	Link	Q5
Mistral 7B	Link	Q4

Table 3: All model sources as well as their quantization method.

Additionally, in order to minimize output variance and generate reproducible evaluations, all tests were performed with a model temperature of 0 and a random seed of 42. Furthermore, our test setting involved a workstation containing an Nvidia A6000, and an Nvidia RTX 4090, with 128 GB of available RAM. All testing code will be made publicly available upon the publication of the work.

B Evaluation Method and Prompt Strings

Table 4 lists the tested prompting methods as well as the special instruction used for each prompt. A special instruction is a text string that is appended to the end of the input question and aims to guide the model behavior while responding to that specific input.

For the case of zero-shot prompting, we simply ask the model to only respond with the correct answer without providing any instructions to reason about the input. Chain-of-thought reasoning is additionally employed via the guidelines provided by Kojima et al. (2022). Finally, we showcase the special instructions for the proposed method, both containing the implicit assumption generation, and common argument generation.

[3] https://github.com/ollama/ollama-python

Prompting Method	Special Instruction
Zero-Shot	Only respond with the correct answer
Chain-of-Thought	Let's think about each option step by step
Argument Generation w/ Implicit Assumptions	When answering, first reason about each choice, and make an argument for why it can be the answer and why it cannot be the answer. Then identify, for each choice, what implicit assumptions you might be making for each of your arguments. By implicit assumption, we mean those propositions that are necessary so that the choice logically follows the question. Then select one of the choices based on the strongest argument
Argument Generation	When answering, first reason about each choice, and make an argument for why it can be the answer and why it cannot be the answer. Then select one of the choices based on the strongest argument.

Table 4: Special model instructions corresponding to each prompting method.