Why Can Large Language Models Generate Correct Chain-of-Thoughts?

Rasul Tutunov [*1] Antoine Grosnit [*1 2] Juliusz Ziomek [1] Jun Wang [1 3] Haitham Bou-Ammar [1 3]

Abstract

This paper delves into the capabilities of large language models (LLMs), specifically focusing on advancing the theoretical comprehension of chain-of-thought prompting. We investigate how LLMs can be effectively induced to generate a coherent chain of thoughts. To achieve this, we introduce a two-level hierarchical graphical model tailored for natural language generation. Within this framework, we establish a *compelling geometrical convergence rate* that gauges the likelihood of an LLM-generated chain of thoughts compared to those originating from the true language. Our findings provide a theoretical justification for the ability of LLMs to produce the correct sequence of thoughts (potentially) explaining performance gains in tasks demanding reasoning skills.

1. Introduction

Since their inception (OpenAI, 2023; Touvron et al., 2023; Jiang et al., 2023; Brown et al., 2020; Radford et al., 2019), large language models (LLMs) have revolutionised natural language processing (Zhang et al., 2023; Wang et al., 2023b; Ge et al., 2023; Lu et al., 2023a; Singhal et al., 2023; Tan et al., 2023; Guo et al., 2023a; Huang et al., 2023b; Bohnet et al., 2022; Robinson et al., 2022) and seen widespread applications in a variety of fields, including but not limited to robotics (Singh et al., 2023; Mai et al., 2023; Vemprala et al., 2023; Huang et al., 2023a; Shah et al., 2023; Liang et al., 2023; Ahn et al., 2022), medicine (Thirunavukarasu et al., 2023; Howard et al., 2023; Mbakwe et al., 2023; Gilbert et al., 2023; Nori et al., 2023; Liu et al., 2023b), search (Kamalloo et al., 2023; Vu et al., 2023), content creation (Orenstrakh et al., 2023; Gmeiner & Yildirim, 2023), code development (Liu et al., 2023a; Wang et al., 2022; Shen et al., 2023; Zheng et al., 2023; Christopoulou et al., 2022; Chang et al., 2023), customer support (Pesaru et al., 2023; Markapudi et al., 2021). LLMs have amazingly demonstrated many emergent capabilities, e.g., chain-of-thought prompting, instruction fine-tuning, and in-context learning, when model sizes and training data grew large (Jiang, 2023; Wei et al., 2022a). Even more interesting is that LLMs exhibit new capabilities unobserved in small models or data scales, in contrast to simple scaling laws that convey generalisation within data distributions while only being trained on next-token prediction, as noted in (Jiang, 2023; Lu et al., 2023b; Webb et al., 2023; Boiko et al., 2023; Noever & McKee, 2023; Teehan et al., 2022; Wei et al., 2022b).

Given those intriguing properties, machine learning researchers and practitioners have sought empirical and theoretical justifications that uncover the "mystery" behind those emergent behaviours. Empirically, works focused on understanding which training data properties may lead to emergent behaviour in LLMs and large-scale transformers. For instance, Chan et al. (2022) demonstrates that in-context learning arises from dynamic clusters of items rather than uniformly distributed ones. Moreover, the authors of (Min et al., 2022) show that contrary to common belief, LLMs do not demand ground truth demonstration data to attain satisfactory performance in classification tasks. They instead discover label space coverage, input text distributions and the format of the input sequence as the main driving factors of in-context learning. In another direction, the work in (Sanh et al., 2022) notices that one can induce zero-shot generalisation via explicit multitask learning across prompts, while the authors in (Razeghi et al., 2022) study the importance of term frequencies during pretraining.

Theoretical Attempts of In-Context Learning: Apart from the above empirical studies, many authors have also attempted to provide a cohesive theoretical justification for such emergent behaviour from LLMs (Xie et al., 2022; Wies et al., 2023; Hahn & Goyal, 2023; Jiang, 2023). In their seminal work, (Xie et al., 2022) formalised in-context learning as latent concept discovery. They showed that LLMs learn to perform implicit Bayesian marginalisation during pretraining and that those models infer shared concepts at test time despite distribution mismatches between

[*]Equal contribution [1]Huawei Noah's Ark [2]Technical University of Darmstadt [3]University College London. Correspondence to: Rasul Tutunov <rasul.tutunov@huawei.com>, Antoine Grosnit <antoine.grosnit2@huawei.com>.

Preprint.

prompts and pretraining data. While providing a valid theoretical justification, their framework analyses the setup when the pretraining data distribution consists of a mixture of hidden Markov models (HMMs) and when an infinite limit of in-context examples are accessible. The work in (Wies et al., 2023) addressed the second limitation (i.e., infinite in-context examples), deriving finite sample complexity results under a PAC framework. Extending the results in (Xie et al., 2022) to general data distributions, the work in (Jiang, 2023) explored sparsity properties present in joint distributions of languages and again demonstrated successful in-context learning when considering ambiguous and unambiguous latent language models.

Chain-of-Thoughts Prompting: While in-context learning is an important emergent property of LLMs, improving performance by carefully designed input prompts is another intriguing property that led to remarkable successes, especially in mathematics and general reasoning tasks. In those reasoning domains, standard in-context prompting in which we ask the LLM to generate the final answer immediately often fails to yield the correct solutions. To address such shortcomings, researchers proposed chain-of-thought (CoT) prompting that enabled LLMs to improve by triggering those models to output intermediate derivations (Wei et al., 2022b; Feng et al., 2023). CoT prompting is generally performed in one of two ways: *i)* by augmenting the input with specific phrases, e.g., "Let us think step by step" (Kojima et al., 2022), or *ii)* by providing a few-shot examples, see (Wei et al., 2022b). Albeit enabling impressive empirical successes, CoT prompting remains a mystery, demanding better in-depth theoretical justification. Only recently did the authors of (Feng et al., 2023) attempt to provide a (partial) validation to uncover why CoTs can succeed. They tackled this problem from an expressivity perspective, revealing two fundamental results for a class of arithmetic problems: *i)* if we constrain a transformer-based LLM to output an answer directly, then the model depth should grow super-polynomially compared to the input length to cover any problem, while *ii)* if we trigger the LLM to generate CoT, then a constant depth transformer is sufficient to solve any task from this class.

Contributions of This Study: The work above answers important (orthogonal) expressivity questions of auto-regressive transformer models but leaves unanswered *how LLMs can be triggered to produce CoTs from this auto-regressive process.* In answering this question, we revert our attention to Bayesian inference and study the properties of LLMs as marginal approximators of natural language. As mentioned earlier, we are not the first to consider such a setup for LLMs (Xie et al., 2022; Jiang, 2023). While similar in spirit to those works, directly applying the analysis of (Xie et al., 2022; Jiang, 2023) to justify CoTs is challenging and requires more realistic stochastic processes to account for the generation of a chain of thoughts. Consequently, we introduce a hierarchical graphical model with two-level latent variables denoting (unobserved) contexts - e.g., arithmetic operations or common-sense reasoning - and intentions. Importantly, we assume non-static latent intentions that evolve and generate the messages expressing the sequence of thoughts of the reasoning process for a true-to-practice setup. Notably, we assume that an intention $\boldsymbol{\theta}_i$ at step i is conditionally generated on *all previous messages and intentions and the context of the reasoning task* - a setup widely used in practice when producing CoTs from LLMs (Feng et al., 2023; Wei et al., 2022b). This two-level hierarchical model that allows for the conditional evolution of latent intentions requires us to introduce new ambiguity definitions when deriving valid upper bounds on quantities of interest. After presenting those, we further contribute by deriving a *geometric upper bound* between the likelihood of the sequence of thoughts generated by an LLM versus those that would have been generated from the actual language. Section 33 presents a formal exposure of our bounds. Now, we informally state our main result, providing intuition to the reader of what is to come:

$$\begin{aligned}|p_{\text{LLM}}(\texttt{CoT}|\texttt{Inp}, \texttt{CoT-Examples(N)}) \\ - q_{\texttt{True}}(\texttt{CoT}|\texttt{Inp}, \texttt{True-Context})| \leq \rho^N,\end{aligned}$$

with $\rho < 1$ is a function of the language ambiguities. In other words, we consider the difference between the likelihood $p_{\text{LLM}}(\texttt{CoT}|\texttt{Inp}, \texttt{CoT-Examples(N)})$ of generating $\texttt{CoT}$ without knowing the true context[1] but with a pre-prompt containing an input and a set of N chain-of-thought examples, and the likelihood $q_{\texttt{True}}(\texttt{CoT}|\texttt{Inp}, \texttt{True-Context})$ of generating this same $\texttt{CoT}$ from the *true* language conditioned on the same input and the *true-context*. Our results show that this difference is bounded by ρ^N, meaning that when triggered correctly, LLMs can generate the correct chain of thoughts that have been shown to improve performance on reasoning tasks.

2. Chain-of-Thoughts Formulation

2.1. Chains of Thoughts in Natural Language.

From a sensible natural language generation process viewpoint, a message is a sequence of tokens produced to convey information under a specific intention. Since the intention itself is not directly observable, it is a latent variable in the natural language generation process. Moreover, as problem complexity increases, it becomes more natural to

[1] Note that this is the standard setup under which LLMs operate to generate CoTs, i.e., from input and a set of CoT examples, the model tries to extrapolate a new sequence of thoughts that aid in solving the problem.

arrive to the solution step by step rather than only providing the answer in a single message. A natural language generative model should therefore account for the generation of step by step solutions also known as chains of thought. In a valid chain of thought prompt, each element of the chain is a message representing a thought, and the sequence of thoughts should be relevant and coherent to lead to the expected answer (Wang et al., 2023a). A message - and therefore its generative underlying intention - is relevant if it refers to some essential elements (precise names, figures, etc.) from the previous messages, and the sequence of thoughts is coherent if the i^{th} step can be logically derived from the initial question and the previous thoughts.

An example of CoT: To illustrate the notions of relevance and coherence introduced by Wang et al. (2023a) , we analyse the first example of CoT provided in (Wei et al., 2022b), which we rewrite below using one colour per intermediary reasoning step.

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can have 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5 + 6 = 11. The answer is 11.

As we show in Table 1, each message consists of a coherent and relevant step, which leads to the final answer. For instance, the last message of the reasoning, "5 + 6 = 11", proceeds from a coherent thought consisting in adding the number of balls Roger started with (in the message, the relevant number "5" is retrieved from the question) and the number of balls Roger acquired (retrieving the relevant number "6" from the previous reasoning step) in order to arrive to the total number of tennis balls Roger has now.

Probabilistic Graphical Model: Building from these observations and from previous natural language latent models (Jiang, 2023; Xie et al., 2022), we propose in Figure 1 a new hierarchical latent probabilistic graphical model for natural language generation that can account for the generation of the chains of thought we find in corpora. The process starts by sampling a context $\boldsymbol{c}$ from a prior distribution $q(\boldsymbol{c})$ over the finite space of contexts $\boldsymbol{C}$. The abstract notion of context corresponds to an unobserved variable which can induce a specific type of logic (e.g. generating code in some language, doing arithmetic operations, imitating commonsense reasoning, etc.[2]). The initial intention $\boldsymbol{\theta}_0$ is generated based on the context $\boldsymbol{c}$, and the first message $\boldsymbol{x}_0$, which can be the description of a problem to solve, is sampled from a conditional distribution $q(\boldsymbol{x}_0|\boldsymbol{\theta}_0)$ such that the message conveys the information encapsulated in the intention. At subsequent steps, the intention $\boldsymbol{\theta}_{i+1}$ is generated conditionally on all the previous messages $(\boldsymbol{x}_j)_{0\leq j\leq i}$ and intentions $(\boldsymbol{\theta}_j)_{0\leq j\leq i}$ to allow relevance, and on the overall context $\boldsymbol{c}$ responsible for the coherence of the chain. At the lowest level, each message $\boldsymbol{x}_i$ is sampled conditioned only on its underlying intention $\boldsymbol{\theta}_i$. Moreover, we assume the existence of a terminal intention $\boldsymbol{\theta}_{\text{END}}$ to which all sequences of messages converge, and such that $q(\text{"}\langle\text{END}\rangle\text{"}|\boldsymbol{\theta}_{\text{END}}) = 1$, where "⟨END⟩" is the stop token. Therefore, we have the following probabilistic process that can generate chains of messages of variable lengths:

$$\begin{aligned}&\boldsymbol{c} \sim q(\boldsymbol{c}),\ \boldsymbol{\theta}_0 \sim q(\cdot|\boldsymbol{c}),\ \boldsymbol{x}_0 \sim q(\cdot|\boldsymbol{\theta}_0)\\ &\boldsymbol{\theta}_i \sim q(\cdot|\boldsymbol{x}_{0:i-1}, \boldsymbol{\theta}_{0:i-1}, \boldsymbol{c}),\ \boldsymbol{x}_i \sim q(\cdot|\boldsymbol{\theta}_i) \quad \forall i \geq 1\\ &\text{until} \quad \boldsymbol{x}_i = \text{"}\langle\text{END}\rangle\text{"}\end{aligned}$$

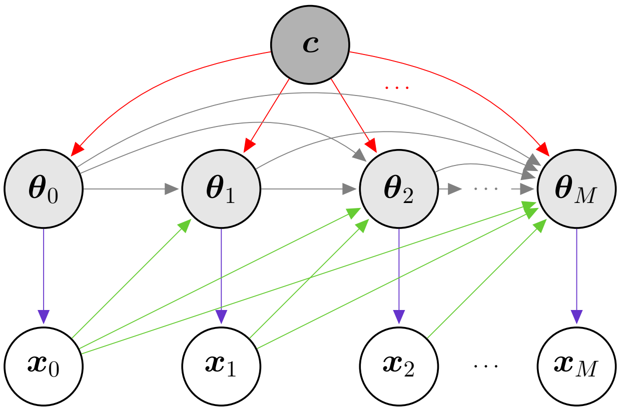

Figure 1. Probabilistic graphical model of natural language text generation that is compatible with the generation of chains of thoughts. $\boldsymbol{c}$ is a context, $(\boldsymbol{\theta}_i)_{0\leq i\leq M}$ is a sequence of intentions, and $(\boldsymbol{x}_i)_{0\leq i\leq M}$ is the sequence of messages corresponding to the formulated thoughts. The generation ends when the stop token is output x_M = "⟨END⟩".

We assume that the corpora on which LLMs are trained come from i.i.d samples following the generative process described in Figure 1. Notably, this does not imply that all the prompts from the pretraining datasets correspond to chains of thought since not all context variables in $\boldsymbol{C}$ induce coherent or relevant prompts. The context set should be broad enough to account for the diversity of the texts and speeches found in datasets.

2.2. Chain of Thoughts in LLMs

LLMs as marginal approximators State-of-the-art LLMs (Touvron et al., 2023; OpenAI, 2023) are large-scale autoregressive models based on transformer architectures (Vaswani et al., 2017) trained to perform text comple-

[2] These examples serve to give an idea of how to grasp this notion of context, but in reality, it is - as intentions - not observed and not as straightforward as what our examples show.

Table 1. Analysis of a chain of thoughts in terms of relevance and coherence.

Message	Relevance	Coherence
Roger started with 5 balls.	Refers to the “5 tennis balls” from the question.	Counts of the initial number of balls.
2 cans of 3 tennis balls each is 6 tennis balls.	Refers to ”2 cans” and “3 tennis balls” from the question.	Computes the number of balls bought on top of the ones counted before.
5 + 6 = 11	Refers to “5” and “6” from the two previous messages.	Uses previous steps counting independent sets of balls to get the total number of balls.

tion. Concretely, an autoregressive LLM learns a conditional distribution over tokens $p_{\text{LLM}}(\cdot|\texttt{prompt})$ that should give more weights on the tokens that are the most likely to follow the given $\texttt{prompt}$ according to the training data distribution. At inference time, given any task description prompt $\boldsymbol{x}$, the LLM can generate an answer by recursively predicting the sequence of next tokens from the learnt distribution p_{LLM} conditioned on the concatenation of $\boldsymbol{x}$ and of the tokens sampled so far. Interestingly, Jiang (2023) showed that as model and training scales increase, LLM pretraining allows to capture the marginal distribution of the natural language, implying that for all sequences of messages $\boldsymbol{X} = (\boldsymbol{x}_0, \ldots, \boldsymbol{x}_m)$ of at most T tokens, $p_{\text{LLM}}(\boldsymbol{X})$ can be a good approximation of $q_{\texttt{True}}(\boldsymbol{X})$. This result relies on the fact that large enough transformer-based models can approximate any continuous sequence-to-sequence functions with arbitrary precision (Yun et al., 2020), and in particular serve as approximators of $\boldsymbol{X} \mapsto q_{\texttt{True}}(\boldsymbol{X})$. Moreover, Jiang (2023) established that if $p_{\text{LLM},n}$ maximises the empirical log-likelihood function $\log(p_{\text{LLM},n}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n))$ with $(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ being n i.i.d. samples drawn from the natural language generative process, then the marginal distribution $p_{\text{LLM},n}(\boldsymbol{X})$ converges to $q_{\texttt{True}}(\boldsymbol{X})$ for all sequences $\boldsymbol{X}$ as n goes to infinity.

In light of this theoretical result and considering that p_{LLM} effectively matches the marginal distribution of $q_{\texttt{True}}$, we can analyse the empirical successes of the latest LLMs which benefited from a drastic increase in terms of model and training data sizes. Indeed, these enhancements allowed the emergence of capabilities unobserved with weaker models, such as the ability to trigger the generation of valid intermediate reasoning steps to solve tasks like arithmetic problems or symbolic reasoning (Wei et al., 2022b). This feature is highly desirable for at least two reasons. Firstly, it allows the LLM to provide the correct solution to complex problems for which direct generation of the answer would fail, as documented in many previous works (Wei et al., 2022b; Wang et al., 2023a; Yang et al., 2023; Kojima et al., 2022; Suzgun et al., 2023). Secondly, outputting a sequence of intermediary reasoning steps allows users to understand and validate the LLM’s final answer and distinguish valid solutions from lucky guesses (or potentially indistinguishable unlucky ones). This capacity can be elicited via proper conditioning of the LLM obtained with CoT prompting.

Inferring context from CoT prompting To solve a task using an LLM generally consists of formulating the task in natural language to form an initial prompt $\boldsymbol{x}_0$ and to output an answer by recursively sampling one token at a time from the learnt conditional distribution. Nevertheless, sampling the answer starting directly from $p_{\text{LLM}}(\cdot|\boldsymbol{x}_0)$ can lead to worse performance than starting from $\text{warp}(\boldsymbol{x}_0)$, where warp is some strategy enriching the initial task description. A growing line of work is precisely dedicated to the design of prompt crafting strategies (Fernando et al., 2023; Chen et al., 2023; Guo et al., 2023b; Chen et al., 2022; Cobbe et al., 2021; Lester et al., 2021), and endeavours to find warping functions warp that favour the sampling of the expected answer. For reasoning tasks, Wei et al. (2022b) showed that few-shot CoT prompting improves the success rate by favouring the generation of valid CoTs.

Formally, given a task $\boldsymbol{x}_0$ and its step-by-step solution $(\boldsymbol{x}_r)_{1\leq r\leq m}$ generated in natural language from the latent context $\boldsymbol{c}^*$, CoT prompting relies on the warping $\text{wrap}_{N-\texttt{shot}}$ to tighten the gap between $p_{\text{LLM}}((\boldsymbol{x}_r)_{1\leq r\leq m}|\,\text{wrap}_{N-\texttt{shot}}(\boldsymbol{x}_0))$ and $q_{\texttt{True}}((\boldsymbol{x}_r)_{1\leq r\leq m}|\boldsymbol{x}_0, \boldsymbol{c}^*)$. Concretely, the $\texttt{wrap}_{N-\texttt{shot}}$ strategy consists in adding to $\boldsymbol{x}_0$ a prompt made of N examplar problems with their respective step-by-step answers $\boldsymbol{Z}_k = (\boldsymbol{z}_{k,r})_{0\leq r\leq m_k}$ also coming from the same context $\boldsymbol{c}^*$, actually conditioning p_{LLM} on $\boldsymbol{x}_0, (\boldsymbol{Z}_k)_{1\leq k\leq N})$. Besides, 0-shot-CoT prompting does not use any related CoT examples in its warping, but it elicits CoT output by adding an instruction prompt $\langle\texttt{INST}\rangle$ to $\boldsymbol{x}_0$ (e.g. “Let’s think step by step” (Kojima et al., 2022), “Take a deep breath and work on this problem step-by-step” (Yang et al., 2023), etc.), which corresponds to having $\texttt{warp}_{0-\text{shot}}(\boldsymbol{x}_0) = (\boldsymbol{x}_0, \langle\texttt{INST}\rangle)$. Since zero-shot-CoT does not perform on par with few-shot-CoT prompting (Kojima et al., 2022), *we focus on the latter in* *this paper.*

Natural Language Ambiguity: In the next section, we will show that the success of CoT prompting comes from the capacity of the LLMs to infer the true reasoning context from the provided series of examples $(\boldsymbol{Z}_k)_{1\leq k\leq N}$. Nevertheless, the intrinsic ambiguity of natural language is a factor impeding the inference ability of LLMs since, by definition of ambiguity, it is not possible to fully identify the underlying intention or context of an ambiguous message or a sequence of messages. Formally, for a chain of messages $(\boldsymbol{x}_i)_{0\leq i\leq m}$ generated from our graphical model with latent context $\boldsymbol{c}^*$ and intentions $(\boldsymbol{\theta}_i^*)_{0\leq i\leq m}$, we express the ambiguity of the chain $\epsilon((\boldsymbol{x}_i)_{0\leq i\leq m})$ as the complement of the likelihood of the context $\boldsymbol{c}^*$ and intentions $(\boldsymbol{\theta}_i^*)_{0\leq i\leq m}$ conditioned on $(\boldsymbol{x}_i)_{0\leq i\leq m}$, i.e. we have

$$q_{\texttt{True}}(\boldsymbol{c}^*, (\boldsymbol{\theta}_i^*)_{0\leq i\leq m}|(\boldsymbol{x}_i)_{0\leq i\leq m}) = 1 - \epsilon((\boldsymbol{x}_i)_{0\leq i\leq m}).$$

This formulation extends the definition of ambiguity that Jiang (2023) introduced to study in-context learning [3]. For the in-context learning setup it is sufficient to define the ambiguity of isolated prompts with respect to their unique underlying intentions $\epsilon(\boldsymbol{\theta}_{\boldsymbol{x}}^*|\boldsymbol{x})$ as the examples provided in in-context learning to the prompt are independent messages generated from the same intention. On the other hand, in our setup the LLM is conditioned on sequences of prompts generated from a coherent chain of intentions guided by a hidden context, which requires to take into account the ambiguity of an entire sequence of messages with respect to all their underlying generative variables. Nevertheless, since ambiguity goes against the core functional purpose of language (conveying information), it is reasonable to assume as in (Jiang, 2023) that language evolved to decrease the overall level of ambiguity.

Next, we provide an upper bound on the difference between the likelihood of the reasoning steps under the true language $q_{\texttt{True}}(\cdot|\boldsymbol{x}_0, \boldsymbol{c}^*)$ and under the LLM conditioned via k-shot-CoT, and we express this bound as a product of ambiguity terms.

3. LLMs can Produce Correct CoTs

This section presents our main results that shed light on how LLMs can effectively produce the correct sequence of thoughts, which, as mentioned earlier, has significantly enhanced their performance in various reasoning tasks. Before detailing our proof, we introduce and discuss an assumption on the prior distribution of the context $q(\boldsymbol{c})$.

Assumption 3.1. The prior distribution $q(\boldsymbol{c})$ associated with contexts $\boldsymbol{c} \in \boldsymbol{C}$ is uniform.

The above assumption implies that natural language does not discriminate in some contexts over others. While this is valid when using large enough and well-balanced datasets, it can be violated in general, especially if data is imbalanced or skewed. In those cases, either more data can be collected, or self-supervised data augmentation can be adopted for skew correction (Yao et al., 2021; Reed et al., 2021; Liu et al., 2021). With that being said, in Section 4, we offer an alternative route that relaxes this assumption and can (potentially) lead us to a new bound for CoT generation with LLMs.

Main Theorem: Equipped with the above assumption, we now present our main result, which considers N-shot-CoT. In formalising our theorem, we follow the setup described in the previous section that we briefly summarise here for ease of exposure:

Problem Setup

- The LLM is provided with N *varying length* chain of thought examples $\boldsymbol{Z}_k = (\boldsymbol{z}_{k,r})_{0\leq r\leq m_k}$ with m_k denoting the length of the chain $\boldsymbol{Z}_k$ and each $\boldsymbol{z}_{k,r}$ being a sequence of tokens representing a thought or one reasoning step;
- Those examples are designed to aid the LLM in producing correct answers via CoT generation and thus, for all k, $\boldsymbol{Z}_k$ are generated with true intentions $(\boldsymbol{\theta}_{k,r}^\star)_{0\leq r\leq m_k}$ and context $\boldsymbol{c}^\star$ in mind;
- Given $\boldsymbol{Z}_k$ and a task $\boldsymbol{x}_0$, the LLM then generates $(\boldsymbol{x}_r)_{1\leq r\leq m}$ messages.

At this stage, we can define the following: *i)* the likelihood of the generated thoughts with our model, i.e. $p_{\texttt{LLM}}((\boldsymbol{x}_r)_{1\leq r\leq m}|\boldsymbol{x}_0, \boldsymbol{Z}_k)$, and *ii)* the likelihood of the generated thoughts if we had known $\boldsymbol{c}^\star$, i.e. $q_{\texttt{True}}((\boldsymbol{x}_r)_{1\leq r\leq m}|\boldsymbol{x}_0, \boldsymbol{c}^\star)$. We then bound their differences in terms of the language's ambiguity to arrive at the following result:

Theorem 3.2. *Consider a collection of N varying length chain-of-thought examples $\boldsymbol{Z}_k = (\boldsymbol{z}_{k,r})_{0\leq r\leq m_k}$ generated from $(\boldsymbol{\theta}_{k,r}^*)_{0\leq r\leq m_k}$ with a context $\boldsymbol{c}^* \sim q(\boldsymbol{c})$ that satisfies Assumption 3.1. Furthermore, let $\boldsymbol{x}_0 \sim q(\cdot|\boldsymbol{\theta}_0^\star)$ be the input message or task generated from $\boldsymbol{\theta}_0^\star \sim q(\cdot|\boldsymbol{c}^\star)$. Then, for any sequence of messages $(\boldsymbol{x}_r)_{1\leq r\leq m}$ we have:*

$$\begin{aligned} &|p_{LLM}((\boldsymbol{x}_r)_{1\leq r\leq m}|\boldsymbol{x}_0, (\boldsymbol{Z}_k)_{1\leq k\leq N}) \\ &\quad - q_{True}((\boldsymbol{x}_r)_{1\leq r\leq m}|\boldsymbol{x}_0, \boldsymbol{c}^*)| \leq \eta \prod_{k=1}^{N} \frac{\boldsymbol{\epsilon}(\boldsymbol{Z}_k)}{1-\boldsymbol{\epsilon}(\boldsymbol{Z}_k)}, \end{aligned} \tag{1}$$

with $\eta = 2\left({}^{\epsilon(\boldsymbol{x}_0)}/_{1\,-\,\epsilon(\boldsymbol{x}_0)}\right)$ depending on the ambiguity of the input task.

[3] In (Jiang, 2023), the Section 6 discussing CoT only considers the case of unambiguous languages.

Informally, this result implies that a large language model prompted with chain-of-thought examples $(\boldsymbol{Z}_k)_{1\leq k\leq N}$ is capable of approximating the true natural language distribution equipped with a true context knowledge. The accuracy of this approximation is guided by ambiguity properties of provided examples $(\boldsymbol{\epsilon}(\boldsymbol{Z}_k))_{1\leq k\leq N}$ and input task $\epsilon(\boldsymbol{x}_0)$. Notably, our result in Theorem 3.2 holds for any sequence of messages $(\boldsymbol{x}_r)_{1\leq r\leq m}$. Consequently, this bound will also be valid for those messages generated by the LLM when prompted with $\boldsymbol{Z}_k$ and $\boldsymbol{x}_0$ in accordance with our "Problem Setup".

Proof. Appendix A details the proof of Theorem 3.2. Here, we provide a *proof sketch* that presents the main steps needed to achieve our bound. Starting from $p_{\texttt{LLM}}((\boldsymbol{x}_r)_{1\leq r\leq m}|\boldsymbol{x}_0, \boldsymbol{Z}_{k=1}^N)$ and noticing that the LLMs' marginal distribution matches the true marginal, we can write:

$$p_{\texttt{LLM}}((\boldsymbol{x}_r)_{1\leq r\leq m}|\boldsymbol{x}_0, \boldsymbol{Z}_{k=1}^N) = \frac{q_{\texttt{True}}((\boldsymbol{x}_r)_{1\leq r\leq m}|\boldsymbol{x}_0, \boldsymbol{c}^\star) + \mathcal{A}}{1+\mathcal{B}},$$

where $\mathcal{A}$ and $\mathcal{B}$ are defined as:

$$\mathcal{A} = \frac{\sum\limits_{\boldsymbol{c}\neq\boldsymbol{c}^*} q_{\texttt{True}}((\boldsymbol{x}_r)_{0\leq r\leq m}, \boldsymbol{Z}_{k=1}^N, \boldsymbol{c})}{q_{\texttt{True}}(\boldsymbol{Z}_{k=1}^N, \boldsymbol{x}_0, \boldsymbol{c}^*)}$$

$$\mathcal{B} = \sum_{\boldsymbol{c}\neq\boldsymbol{c}^*} \frac{q_{\texttt{True}}(\boldsymbol{x}_0, \boldsymbol{Z}_{k=1}^N, \boldsymbol{c})}{q_{\texttt{True}}(\boldsymbol{Z}_{k=1}^N, \boldsymbol{x}_0, \boldsymbol{c}^*)}.$$

We notice that although the input message $\boldsymbol{x}_0$ and the CoT examples $\boldsymbol{Z}_{k=1}^N$ share the same context $\boldsymbol{c}^*$, they are independent when conditioned on $\boldsymbol{c}^\star$. This observation allows us to establish the following bounds on $\mathcal{A}$ and $\mathcal{B}$:

$$\mathcal{A} \text{ and } \mathcal{B} \leq \frac{\boldsymbol{\epsilon}(\boldsymbol{x}_0)}{1-\boldsymbol{\epsilon}(\boldsymbol{x_0})} \prod_{k=1}^N \frac{\boldsymbol{\epsilon}(\boldsymbol{Z}_k)}{1-\boldsymbol{\epsilon}(\boldsymbol{Z}_k)}.$$

We arrive at the theorem's statement upon substituting those results in the absolute likelihood difference expression. □

4. What CoT Examples Will Work?

This section investigates sufficient requirements for CoT examples to guarantee vanishing bound in Equation 1. We start with the case when the CoT examples $\boldsymbol{Z}_k$ exhibit low ambiguity:

Condition 4.1. Imagine that the chain-of-thought examples $\boldsymbol{Z}_k = (\boldsymbol{z}_{k,r})_{0\leq r\leq m_k}$ generated from $(\boldsymbol{\theta}^*_{k,r})_{0\leq r\leq m_k}$ with a context $\boldsymbol{c}^* \sim q(\boldsymbol{c})$ have a bounded ambiguity measure such that:

$$\boldsymbol{\epsilon}(\boldsymbol{Z}_k) = q_{\texttt{True}}(\boldsymbol{c}^*, (\boldsymbol{\theta}^*_{k,r})_{0\leq r\leq m_k}|(\boldsymbol{z}_{k,r})_{0\leq r\leq m_k}) \leq \delta, \tag{2}$$

where $\delta \in [0, \frac{1}{2})$.

This condition implies that the LLM model is prompted by carefully chosen CoT examples $\boldsymbol{Z}_k$ such that the true context $\boldsymbol{c}^*$ can be recovered from $\boldsymbol{Z}_k$ with reasonably high certainty (the probability the $\boldsymbol{c}^*$ is behind $\boldsymbol{Z}_k$ is strictly greater than one half). Provided with such CoT examples, one can establish a geometrical convergence rate for the bound in Equation 1 as the number of examples grows large:

$$\begin{aligned}|p_{\texttt{LLM}}((\boldsymbol{x}_r)_{1\leq r\leq m}|\boldsymbol{x}_0, (\boldsymbol{Z}_k)_{1\leq k\leq N}) \\ - q_{\texttt{True}}((\boldsymbol{x}_r)_{1\leq r\leq m}|\boldsymbol{x}_0, \boldsymbol{c}^*)| \leq \eta\rho^{\mathrm{N}},\end{aligned} \tag{3}$$

where $\rho = \frac{\delta}{1-\delta} \in [0, 1)$.

Those examples described in Condition 4.1 should be carefully selected to guarantee low ambiguity requirements. In practice, however, it could be challenging to collect such chain-of-thought examples, as there is no rigorous procedure allowing us to measure ambiguity for a given sequence of thoughts. To remedy these strict requirements, we consider the following relaxed sufficient conditions:

Condition 4.2. For the chain-of-thought examples $\boldsymbol{Z}_k = (\boldsymbol{z}_{k,r})_{0\leq r\leq m_k}$ generated from $(\boldsymbol{\theta}^*_{k,r})_{0\leq r\leq m_k}$ with a context $\boldsymbol{c}^* \sim q(\boldsymbol{c})$, the associated ambiguity measure $\epsilon((\boldsymbol{z_{k,r}})_{0\leq r\leq m_k})$ vanishes as length of sequence grows large:

$$\lim_{\ell\to\infty} \epsilon((\boldsymbol{z_{k,r}})_{0\leq r\leq \ell}) = 0. \tag{4}$$

Intuitively, the above condition implies that uncertainty over true context $\boldsymbol{c}^*$ and associated intentions $(\boldsymbol{\theta}^*_{k,r})_{0\leq r\leq m_k}$ for a sequence of thoughts is diminishing when more of these thoughts are collected.

The next lemma demonstrates that this asymptotic requirement is sufficient to guarantee low ambiguity measure for long enough CoT examples:

Lemma 4.3. *Let us consider CoT examples* $\{\boldsymbol{Z}_k = (\boldsymbol{z}_{k,r})_{0\leq r\leq m_k}$ *satisfying condition 4.2. Then, for any fixed* $\delta \in [0, \frac{1}{2})$ *there is a length threshold* $m^*_{k,\delta} \in \mathbb{N}$, *such that for any* $m_k \geq m^*_{k,\delta}$*:*

$$\epsilon(\boldsymbol{Z}_k) \leq \delta$$

Appendix B details the proof of Lemma 4.3. This result implies in particular, that a geometrical convergence rate 3 can also be established when the LLM model is prompted with CoT examples $\boldsymbol{Z}_{k=1}^N$ of sufficient length ($m_k \geq m^*_{k,\delta}$) satisfying Condition 4.2. In contrast to the low ambiguity requirement, CoT examples with low asymptotic ambiguity can be more attainable. Indeed, being provided with high ambiguity CoT example $\boldsymbol{Z}_k$ satisfying Condition 4, one could break it to a longer sequence of thoughts $\boldsymbol{Z}^{'}_k$ consisting of more refined reasoning steps such that ambiguity of $\boldsymbol{Z}^{'}_k$ satisfies a sufficient threshold $\delta \in [0, \frac{1}{2})$.

Non-Uniform Context Priors: We stated our main result under the assumption that the prior distribution over the set of contexts is uniform. Relaxing this assumption, we provide a similar bound on the discrepancy between the true and the predicted likelihood of a CoT. To account for the potentially non-uniform distribution of contexts in training datasets, we introduce a skewness parameter $\gamma(\boldsymbol{c}^*)$, which we define as:

$$\gamma(\boldsymbol{c}^*) = \sup_{\boldsymbol{c}\in\boldsymbol{C}} \frac{q_{\texttt{True}}(\boldsymbol{c}^*)}{q_{\texttt{True}}(\boldsymbol{c})}$$

Giving this measure of skewness, we establish a modified bound on the parameters $\mathcal{A}$ and $\mathcal{B}$:

$$\mathcal{A} \text{ and } \mathcal{B} \leq \frac{\gamma^N(\boldsymbol{c}^*)\boldsymbol{\epsilon}(\boldsymbol{x}_0)}{1-\boldsymbol{\epsilon}(\boldsymbol{x_0})} \prod_{k=1}^{N} \frac{\boldsymbol{\epsilon}(\boldsymbol{Z}_k)}{1-\boldsymbol{\epsilon}(\boldsymbol{Z}_k)}.$$

This, in turn, leading us to a modified bound of the following form:

$$|p_{\texttt{LLM}}((\boldsymbol{x}_r)_{1\leq r\leq m}|\boldsymbol{x}_0, (\boldsymbol{Z}_k)_{1\leq k\leq N}) \\ - q_{\texttt{True}}((\boldsymbol{x}_r)_{1\leq r\leq m}|\boldsymbol{x}_0, \boldsymbol{c}^*)| \leq \hat{\eta} \prod_{k=1}^{N} \frac{\boldsymbol{\epsilon}(\boldsymbol{Z}_k)}{1-\boldsymbol{\epsilon}(\boldsymbol{Z}_k)}, \tag{5}$$

with $\hat{\eta} = 2\gamma^N(\boldsymbol{c}^*)\frac{\boldsymbol{\epsilon}(\boldsymbol{x}_0)}{1-\boldsymbol{\epsilon}(\boldsymbol{x_0})}$ being a constant that depends on the ambiguity of the input task and skewness parameter $\gamma(\boldsymbol{c}^*)$.

CoT examples $(\boldsymbol{Z}_k)_{1\leq k\leq N}$ and such context $\boldsymbol{c}$. Thus, to avoid such a situation, either the dataset of samples from $q_{\texttt{True}}(\cdot)$ should be diverse enough in terms of covering context set $\mathcal{C}$, and hence, guaranteeing small values for $\gamma(\boldsymbol{c}^*)$, or provided CoT examples should have small enough ambiguity, so the model with high certainty could guess the true context $\boldsymbol{c}^*$ from them.

5. Conclusion

While the success of CoT prompting is well supported empirically, the emergent ability of LLMs to solve intricate tasks through intermediary steps is still not fully understood. We claim that our paper sheds light on one important aspect of the CoT mystery. While Feng et al. (2023) studied the task-solving coverage of fixed-depth transformer-based LLMs, showing that the autoregressive generation of several messages permits significant expressivity gains, our paper gives insights into the role of the few-shot CoT prompting technique itself in the exploitation of this expressivity gain. To do so, we have introduced a hierarchial latent language model that accounts for the generation of realistic chains of thought, i.e. of coherent and relevant series of messages. As emergent behaviours, such as LLM step-by-step reasoning ability, appear when the scale of architectures and training datasets increases, we followed the justification of (Jiang, 2023) to consider that LLM predictive distribution $p_{\texttt{LLM}}$ trained on next-token completion is a perfect approximator of the marginal distribution of the language model $q_{\texttt{True}}$. By framing the CoT-prompting as a wrapper of the task input that appends exemplary chains of thought, we showed that the LLM can leverage the examples to infer the underlying reasoning context. Therefore, conditioned on the task prompt and these examples, we derived an upper bound between the likelihood of the expected step-by-step answers under the, expressed as the product of ambiguity factors. Finally, we thoroughly discussed the practical understanding of our theoretical bound, considering the introduction of different conditions on CoT ambiguity and providing a more general bound for non-uniform context priors. We hope that from these discussions, the success of CoT prompting appears less mysterious.

Future Work: We plan to consider many interesting future research avenues. First, we wish to perform a rigorous experimental study to empirically validate our results in Section 3. Second, we will investigate a more complete analysis than this paper by taking into account the separators in the actual prompts as part of our graphical model. Finally, we plan to generalise our analysis to cover recent discoveries in smart prompting, including but not limited to graph and tree of thoughts (Besta et al., 2023; Yao et al., 2023).

6. Acknowledgements

This work is supported by the CSTT on Generalisable Robot Learning via Machine Learning Models 2100332-GB and by the 2030 ”New Generation of AI” - Major Project of China under grant No. 2022ZD0116408.

References

Ahn, M., Brohan, A., Brown, N., Chebotar, Y., Cortes, O., David, B., Finn, C., Fu, C., Gopalakrishnan, K., Hausman, K., et al. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022.

Besta, M., Blach, N., Kubicek, A., Gerstenberger, R., Gianinazzi, L., Gajda, J., Lehmann, T., Podstawski, M., Niewiadomski, H., Nyczyk, P., et al. Graph of thoughts: Solving elaborate problems with large language models. *arXiv preprint arXiv:2308.09687*, 2023.

Bohnet, B., Tran, V. Q., Verga, P., Aharoni, R., Andor, D., Soares, L. B., Eisenstein, J., Ganchev, K., Herzig, J., Hui, K., et al. Attributed question answering: Evaluation and modeling for attributed large language models. *arXiv preprint arXiv:2212.08037*, 2022.

Boiko, D. A., MacKnight, R., and Gomes, G. Emergent autonomous scientific research capabilities of large language models. *arXiv preprint arXiv:2304.05332*, 2023.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

Chan, S. C., Santoro, A., Lampinen, A. K., Wang, J. X., Singh, A. K., Richemond, P. H., McClelland, J., and Hill, F. Data distributional properties drive emergent in-context learning in transformers. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022.

Chang, J. D., Brantley, K., Ramamurthy, R., Misra, D., and Sun, W. Learning to generate better than your llm. *arXiv preprint arXiv:2306.11816*, 2023.

Chen, A., Dohan, D. M., and So, D. R. Evoprompting: Language models for code-level neural architecture search. *arXiv preprint arXiv:2302.14838*, 2023.

Chen, W., Ma, X., Wang, X., and Cohen, W. W. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *arXiv preprint arXiv:2211.12588*, 2022.

Christopoulou, F., Lampouras, G., Gritta, M., Zhang, G., Guo, Y., Li, Z., Zhang, Q., Xiao, M., Shen, B., Li, L., Yu, H., Yan, L., Zhou, P., Wang, X., Ma, Y., Iacobacci, I., Wang, Y., Liang, G., Wei, J., Jiang, X., Wang, Q., and Liu, Q. Pangu-coder: Program synthesis with function-level language modeling, 2022.

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Feng, G., Zhang, B., Gu, Y., Ye, H., He, D., and Wang, L. Towards revealing the mystery behind chain of thought: A theoretical perspective, 2023.

Fernando, C., Banarse, D., Michalewski, H., Osindero, S., and Rocktäschel, T. Promptbreeder: Self-referential self-improvement via prompt evolution. *arXiv preprint arXiv:2309.16797*, 2023.

Ge, Y., Hua, W., Ji, J., Tan, J., Xu, S., and Zhang, Y. Openagi: When llm meets domain experts. *arXiv preprint arXiv:2304.04370*, 2023.

Gilbert, S., Harvey, H., Melvin, T., Vollebregt, E., and Wicks, P. Large language model ai chatbots require approval as medical devices. *Nature Medicine*, pp. 1–3, 2023.

Gmeiner, F. and Yildirim, N. Dimensions for designing llm-based writing support. In *In2Writing Workshop at CHI*, 2023.

Guo, J., Li, J., Li, D., Tiong, A. M. H., Li, B., Tao, D., and Hoi, S. From images to textual prompts: Zero-shot visual question answering with frozen large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10867–10877, 2023a.

Guo, Q., Wang, R., Guo, J., Li, B., Song, K., Tan, X., Liu, G., Bian, J., and Yang, Y. Connecting large language models with evolutionary algorithms yields powerful prompt optimizers. *arXiv preprint arXiv:2309.08532*, 2023b.

Hahn, M. and Goyal, N. A theory of emergent in-context learning as implicit structure induction. *CoRR*, abs/2303.07971, 2023. doi: 10.48550/arXiv.2303.07971.

Howard, A., Hope, W., and Gerada, A. Chatgpt and antimicrobial advice: the end of the consulting infection doctor? *The Lancet Infectious Diseases*, 23(4):405–406, 2023.

Huang, C., Mees, O., Zeng, A., and Burgard, W. Visual language maps for robot navigation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 10608–10615. IEEE, 2023a.

Huang, Z., Zhou, J., Xiao, G., and Cheng, G. Enhancing in-context learning with answer feedback for multi-span question answering. *arXiv preprint arXiv:2306.04508*, 2023b.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. Mistral 7b, 2023.

Jiang, H. A Latent Space Theory for Emergent Abilities in Large Language Models. *arXiv e-prints*, art. arXiv:2304.09960, April 2023. doi: 10.48550/arXiv.2304.09960v3.

Kamalloo, E., Jafari, A., Zhang, X., Thakur, N., and Lin, J. Hagrid: A human-llm collaborative dataset for generative information-seeking with attribution. *arXiv preprint arXiv:2307.16883*, 2023.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., and Iwasawa, Y. Large language models are zero-shot reasoners. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 22199–22213. Curran Associates, Inc., 2022.

Lester, B., Al-Rfou, R., and Constant, N. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021.

Liang, J., Huang, W., Xia, F., Xu, P., Hausman, K., Ichter, B., Florence, P., and Zeng, A. Code as policies: Language model programs for embodied control. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 9493–9500. IEEE, 2023.

Liu, H., HaoChen, J. Z., Gaidon, A., and Ma, T. Self-supervised learning is more robust to dataset imbalance. *arXiv preprint arXiv:2110.05025*, 2021.

Liu, J., Xia, C. S., Wang, Y., and Zhang, L. Is your code generated by chatgpt really correct? rigorous evaluation of large language models for code generation. *arXiv preprint arXiv:2305.01210*, 2023a.

Liu, Z., Yu, X., Zhang, L., Wu, Z., Cao, C., Dai, H., Zhao, L., Liu, W., Shen, D., Li, Q., et al. Deid-gpt: Zero-shot medical text de-identification by gpt-4. *arXiv preprint arXiv:2303.11032*, 2023b.

Lu, Q., Sun, X., Long, Y., Gao, Z., Feng, J., and Sun, T. Sentiment analysis: Comprehensive reviews, recent advances, and open challenges. *IEEE Transactions on Neural Networks and Learning Systems*, 2023a.

Lu, S., Bigoulaeva, I., Sachdeva, R., Madabushi, H. T., and Gurevych, I. Are emergent abilities in large language models just in-context learning? *arXiv preprint arXiv:2309.01809*, 2023b.

Mai, J., Chen, J., Li, B., Qian, G., Elhoseiny, M., and Ghanem, B. Llm as a robotic brain: Unifying egocentric memory and control. *arXiv preprint arXiv:2304.09349*, 2023.

Markapudi, B., Latha, K. J., and Chaduvula, K. A new hybrid classification algorithm for predicting customer churn. In *2021 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES)*, pp. 1–4. IEEE, 2021.

Mbakwe, A. B., Lourentzou, I., Celi, L. A., Mechanic, O. J., and Dagan, A. Chatgpt passing usmle shines a spotlight on the flaws of medical education, 2023.

Min, S., Lyu, X., Holtzman, A., Artetxe, M., Lewis, M., Hajishirzi, H., and Zettlemoyer, L. Rethinking the role of demonstrations: What makes in-context learning work? In *EMNLP*, 2022.

Noever, D. and McKee, F. Numeracy from literacy: Data science as an emergent skill from large language models. *arXiv preprint arXiv:2301.13382*, 2023.

Nori, H., King, N., McKinney, S. M., Carignan, D., and Horvitz, E. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*, 2023.

OpenAI. Gpt-4 technical report. *ArXiv*, abs/2303.08774, 2023.

Orenstrakh, M. S., Karnalim, O., Suarez, C. A., and Liut, M. Detecting llm-generated text in computing education: A comparative study for chatgpt cases. *arXiv preprint arXiv:2307.07411*, 2023.

Pesaru, A., Gill, T. S., and Tangella, A. R. Ai assistant for document management using lang chain and pinecone. *International Research Journal of Modernization in Engineering Technology and Science*, 2023.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners, 2019.

Razeghi, Y., IV, R. L. L., Gardner, M., and Singh, S. Impact of pretraining term frequencies on few-shot numerical reasoning. In Goldberg, Y., Kozareva, Z., and Zhang, Y. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pp. 840–854. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.findings-emnlp.59.

Reed, C. J., Metzger, S., Srinivas, A., Darrell, T., and Keutzer, K. Selfaugment: Automatic augmentation policies for self-supervised learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2674–2683, 2021.

Robinson, J., Rytting, C. M., and Wingate, D. Leveraging large language models for multiple choice question answering. *arXiv preprint arXiv:2210.12353*, 2022.

Sanh, V., Webson, A., Raffel, C., Bach, S., Sutawika, L., Alyafeai, Z., Chaffin, A., Stiegler, A., Raja, A., Dey, M., Bari, M. S., Xu, C., Thakker, U., Sharma, S. S., Szczechla, E., Kim, T., Chhablani, G., Nayak, N., Datta, D., Chang, J., Jiang, M. T.-J., Wang, H., Manica, M., Shen, S., Yong, Z. X., Pandey, H., Bawden, R., Wang, T., Neeraj, T., Rozen, J., Sharma, A., Santilli, A., Fevry, T., Fries, J. A., Teehan, R., Scao, T. L., Biderman, S., Gao, L., Wolf, T., and Rush, A. M. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*, 2022.

Shah, D., Osiński, B., Levine, S., et al. Lm-nav: Robotic navigation with large pre-trained models of language, vision, and action. In *Conference on Robot Learning*, pp. 492–504. PMLR, 2023.

Shen, B., Zhang, J., Chen, T., Zan, D., Geng, B., Fu, A., Zeng, M., Yu, A., Ji, J., Zhao, J., et al. Pangu-coder2: Boosting large language models for code with ranking feedback. *arXiv preprint arXiv:2307.14936*, 2023.

Singh, I., Blukis, V., Mousavian, A., Goyal, A., Xu, D., Tremblay, J., Fox, D., Thomason, J., and Garg, A. Progprompt: Generating situated robot task plans using large language models. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 11523–11530. IEEE, 2023.

Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Hou, L., Clark, K., Pfohl, S., Cole-Lewis, H., Neal, D., et al. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*, 2023.

Suzgun, M., Scales, N., Schärli, N., Gehrmann, S., Tay, Y., Chung, H. W., Chowdhery, A., Le, Q. V., Chi, E., Zhou, D., and Wei, J. Challenging big-bench tasks and whether chain-of-thought can solve them. In Rogers, A., Boyd-Graber, J. L., and Okazaki, N. (eds.), *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pp. 13003–13051. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-acl.824.

Tan, Y., Min, D., Li, Y., Li, W., Hu, N., Chen, Y., and Qi, G. Evaluation of chatgpt as a question answering system for answering complex questions. *arXiv preprint arXiv:2303.07992*, 2023.

Teehan, R., Clinciu, M., Serikov, O., Szczechla, E., Seelam, N., Mirkin, S., and Gokaslan, A. Emergent structures and training dynamics in large language models. In *Proceedings of BigScience Episode# 5–Workshop on Challenges & Perspectives in Creating Large Language Models*, pp. 146–159, 2022.

Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., and Ting, D. S. W. Large language models in medicine. *Nature medicine*, 29(8): 1930–1940, 2023.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. Llama: Open and efficient foundation language models, 2023. cite arxiv:2302.13971.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Vemprala, S., Bonatti, R., Bucker, A., and Kapoor, A. Chatgpt for robotics: Design principles and model abilities. *Microsoft Auton. Syst. Robot. Res*, 2:20, 2023.

Vu, T., Iyyer, M., Wang, X., Constant, N., Wei, J., Wei, J., Tar, C., Sung, Y.-H., Zhou, D., Le, Q., et al. Freshllms: Refreshing large language models with search engine augmentation. *arXiv preprint arXiv:2310.03214*, 2023.

Wang, B., Min, S., Deng, X., Shen, J., Wu, Y., Zettlemoyer, L., and Sun, H. Towards understanding chain-of-thought prompting: An empirical study of what matters. In *The 61st Annual Meeting of the Association for Computational Linguistics*, 2023a.

Wang, X., Li, S., and Ji, H. Code4struct: Code generation for few-shot structured prediction from natural language. *arXiv preprint arXiv:2210.12810*, 2022.

Wang, Z., Xie, Q., Ding, Z., Feng, Y., and Xia, R. Is chatgpt a good sentiment analyzer? a preliminary study. *arXiv preprint arXiv:2304.04339*, 2023b.

Webb, T., Holyoak, K. J., and Lu, H. Emergent analogical reasoning in large language models. *Nature Human Behaviour*, pp. 1–16, 2023.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., and Fedus, W. Emergent abilities of large language models, 2022a.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., ichter, b., Xia, F., Chi, E., Le, Q. V., and Zhou, D. Chain-of-thought prompting elicits reasoning in large language models. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837. Curran Associates, Inc., 2022b.

Wies, N., Levine, Y., and Shashua, A. The learnability of in-context learning. *arXiv preprint arXiv:2303.07895*, 2023.

Xie, S. M., Raghunathan, A., Liang, P., and Ma, T. An explanation of in-context learning as implicit bayesian inference. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.

Yang, C., Wang, X., Lu, Y., Liu, H., Le, Q. V., Zhou, D., and Chen, X. Large language models as optimizers. *CoRR*, abs/2309.03409, 2023. doi: 10.48550/arXiv.2309.03409.

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., and Narasimhan, K. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*, 2023.

Yao, T., Yi, X., Cheng, D. Z., Yu, F., Chen, T., Menon, A., Hong, L., Chi, E. H., Tjoa, S., Kang, J., et al. Self-supervised learning for large-scale item recommendations. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pp. 4321–4330, 2021.

Yun, C., Bhojanapalli, S., Rawat, A. S., Reddi, S., and Kumar, S. Are transformers universal approximators of sequence-to-sequence functions? In *International Conference on Learning Representations*, 2020.

Zhang, W., Deng, Y., Liu, B., Pan, S. J., and Bing, L. Sentiment analysis in the era of large language models: A reality check. *arXiv preprint arXiv:2305.15005*, 2023.

Zheng, Q., Xia, X., Zou, X., Dong, Y., Wang, S., Xue, Y., Wang, Z., Shen, L., Wang, A., Li, Y., et al. Codegeex: A pre-trained model for code generation with multilingual evaluations on humaneval-x. *arXiv preprint arXiv:2303.17568*, 2023.

A. Proof of Theorem 3.2

Proof. We will give the proof for the general case of the prior distribution over contexts, and then provide the effect of Assumption 3.1 on the obtained bound. Let us fix any sequence of messages $(\boldsymbol{x}_r)_{1\leq r\leq m}$ and consider the probability of a large language model to output $(\boldsymbol{x}_r)_{1\leq r\leq m}$ while being prompted with input message $\boldsymbol{x}_0$ and chain-of-thoughts examples $\boldsymbol{Z}_{k=1}^N = (\boldsymbol{Z}_k)_{1\leq k\leq N}$:

$$p_{\text{LLM}}((\boldsymbol{x}_r)_{1\leq r\leq m}|\boldsymbol{x}_0, \boldsymbol{Z}_{k=1}^N) = \frac{q_{\texttt{True}}((\boldsymbol{x}_r)_{0\leq r\leq m}, \boldsymbol{Z}_{k=1}^N)}{q_{\texttt{True}}(\boldsymbol{x}_0, \boldsymbol{Z}_{k=1}^N)}$$
$$= \frac{q_{\texttt{True}}((\boldsymbol{x}_r)_{0\leq r\leq m}, \boldsymbol{Z}_{k=1}^N, \boldsymbol{c}^*) + \sum_{\boldsymbol{c}\neq\boldsymbol{c}^*} q_{\texttt{True}}((\boldsymbol{x}_r)_{0\leq r\leq m}, \boldsymbol{Z}_{k=1}^N, \boldsymbol{c})}{q_{\texttt{True}}(\boldsymbol{x}_0, \boldsymbol{Z}_{k=1}^N, \boldsymbol{c}^*) + \sum_{\boldsymbol{c}\neq\boldsymbol{c}^*} q_{\texttt{True}}(\boldsymbol{x}_0, \boldsymbol{Z}_{k=1}^N, \boldsymbol{c})}$$

Notice,

$$q_{\texttt{True}}((\boldsymbol{x}_r)_{0\leq r\leq m}, \boldsymbol{Z}_{k=1}^N, \boldsymbol{c}^*) = q_{\texttt{True}}((\boldsymbol{x}_r)_{1\leq r\leq m}|\boldsymbol{x}_0, \boldsymbol{c}^*) q_{\texttt{True}}(\boldsymbol{Z}_{k=1}^N, \boldsymbol{x}_0, \boldsymbol{c}^*)$$

Hence, after diving both numerator and denominator on $q_{\texttt{True}}(\boldsymbol{Z}_{k=1}^N, \boldsymbol{x}_0, \boldsymbol{c}^*)$ gives:

$$p_{\text{LLM}}((\boldsymbol{x}_r)_{1\leq r\leq m}|\boldsymbol{x}_0, \boldsymbol{Z}_{k=1}^N) = \frac{q_{\texttt{True}}((\boldsymbol{x}_r)_{1\leq r\leq m}|\boldsymbol{x}_0, \boldsymbol{c}^*) + \frac{\sum_{\boldsymbol{c}\neq\boldsymbol{c}^*} q_{\texttt{True}}((\boldsymbol{x}_r)_{0\leq r\leq m}, \boldsymbol{Z}_{k=1}^N, \boldsymbol{c})}{q_{\texttt{True}}(\boldsymbol{Z}_{k=1}^N, \boldsymbol{x}_0, \boldsymbol{c}^*)}}{1 + \sum_{\boldsymbol{c}\neq\boldsymbol{c}^*} \frac{q_{\texttt{True}}(\boldsymbol{x}_0, \boldsymbol{Z}_{k=1}^N, \boldsymbol{c})}{q_{\texttt{True}}(\boldsymbol{Z}_{k=1}^N, \boldsymbol{x}_0, \boldsymbol{c}^*)}} \tag{6}$$
$$= \frac{q_{\texttt{True}}((\boldsymbol{x}_r)_{1\leq r\leq m}|\boldsymbol{x}_0, \boldsymbol{c}^*) + \mathcal{A}}{1 + \mathcal{B}}$$

where in the last step we use the notation:

$$\mathcal{A} = \frac{\sum_{\boldsymbol{c}\neq\boldsymbol{c}^*} q_{\texttt{True}}((\boldsymbol{x}_r)_{0\leq r\leq m}, \boldsymbol{Z}_{k=1}^N, \boldsymbol{c})}{q_{\texttt{True}}(\boldsymbol{Z}_{k=1}^N, \boldsymbol{x}_0, \boldsymbol{c}^*)} \quad \text{and} \quad \mathcal{B} = \sum_{\boldsymbol{c}\neq\boldsymbol{c}^*} \frac{q_{\texttt{True}}(\boldsymbol{x}_0, \boldsymbol{Z}_{k=1}^N, \boldsymbol{c})}{q_{\texttt{True}}(\boldsymbol{Z}_{k=1}^N, \boldsymbol{x}_0, \boldsymbol{c}^*)}$$

Let us study each of the above expressions separately. Using the fact that chain-of-thoughts examples $\boldsymbol{Z}_{k=1}^N$ are collected i.i.d and independent from messages $(\boldsymbol{x}_r)_{0\leq r\leq m}$ $\boldsymbol{x}_0$ we have (for any context $\boldsymbol{c} \in \boldsymbol{C}$):

$$q_{\texttt{True}}((\boldsymbol{x}_r)_{0\leq r\leq m}, \boldsymbol{Z}_{k=1}^N, \boldsymbol{c}) = q_{\texttt{True}}((\boldsymbol{x}_r)_{1\leq r\leq m}|\boldsymbol{x}_0, \boldsymbol{c}) \times q_{\texttt{True}}(\boldsymbol{x}_0|\boldsymbol{c}) \prod_{k=1}^N q_{\texttt{True}}(\boldsymbol{Z}_k|\boldsymbol{c}) q_{\texttt{True}}(\boldsymbol{c})$$

and hence, we can write (using and definition of parameter γ):

$$\mathcal{A} = \sum_{\boldsymbol{c}\neq\boldsymbol{c}^*} q_{\texttt{True}}((\boldsymbol{x}_r)_{0\leq r\leq m}|\boldsymbol{c}, \boldsymbol{x}_0) \frac{q_{\texttt{True}}(\boldsymbol{x}_0|\boldsymbol{c}) q_{\texttt{True}}(\boldsymbol{c})}{q_{\texttt{True}}(\boldsymbol{x}_0|\boldsymbol{c}^*) q_{\texttt{True}}(\boldsymbol{c}^*)} \prod_{k=1}^N \left[\frac{q_{\texttt{True}}(\boldsymbol{Z}_k|\boldsymbol{c}) q_{\texttt{True}}(\boldsymbol{c})}{q_{\texttt{True}}(\boldsymbol{Z}_k|\boldsymbol{c}^*) q_{\texttt{True}}(\boldsymbol{c}^*)}\right] \left[\frac{q_{\texttt{True}}(\boldsymbol{c}^*)}{q_{\texttt{True}}(\boldsymbol{c})}\right]^N$$
$$\leq \gamma^N(\boldsymbol{c}^*) \sum_{\boldsymbol{c}\neq\boldsymbol{c}^*} \frac{q_{\texttt{True}}(\boldsymbol{x}_0|\boldsymbol{c}) q_{\texttt{True}}(\boldsymbol{c})}{q_{\texttt{True}}(\boldsymbol{x}_0|\boldsymbol{c}^*) q_{\texttt{True}}(\boldsymbol{c}^*)} \sum_{\boldsymbol{c}\neq\boldsymbol{c}^*} \prod_{k=1}^N \left[\frac{q_{\texttt{True}}(\boldsymbol{Z}_k|\boldsymbol{c}) q_{\texttt{True}}(\boldsymbol{c})}{q_{\texttt{True}}(\boldsymbol{Z}_k|\boldsymbol{c}^*) q_{\texttt{True}}(\boldsymbol{c}^*)}\right]$$
$$\leq \gamma^N(\boldsymbol{c}^*) \underbrace{\sum_{\boldsymbol{c}\neq\boldsymbol{c}^*} \frac{q_{\texttt{True}}(\boldsymbol{x}_0|\boldsymbol{c}) q_{\texttt{True}}(\boldsymbol{c})}{q_{\texttt{True}}(\boldsymbol{x}_0|\boldsymbol{c}^*) q_{\texttt{True}}(\boldsymbol{c}^*)}}_{\mathcal{A}_1} \prod_{k=1}^N \underbrace{\left[\frac{\sum_{\boldsymbol{c}\neq\boldsymbol{c}^*} q_{\texttt{True}}(\boldsymbol{Z}_k|\boldsymbol{c}) q_{\texttt{True}}(\boldsymbol{c})}{q_{\texttt{True}}(\boldsymbol{Z}_k|\boldsymbol{c}^*) q_{\texttt{True}}(\boldsymbol{c}^*)}\right]}_{\mathcal{A}_2}$$

where we use the probability point mass function $q_{\texttt{True}}((\boldsymbol{x}_r)_{0\le r\le m}|\boldsymbol{c},\boldsymbol{x}_0)\le 1$. Hence:

$$\begin{aligned}\mathcal{A}_1 = \sum_{\boldsymbol{c}\neq\boldsymbol{c}^*}\frac{q_{\texttt{True}}(\boldsymbol{x}_0,\boldsymbol{c})}{q_{\texttt{True}}(\boldsymbol{x}_0,\boldsymbol{c}^*)} = \frac{\sum_{\boldsymbol{c}\neq\boldsymbol{c}^*}\sum_{\boldsymbol{\theta}} q_{\texttt{True}}(\boldsymbol{x}_0,\boldsymbol{\theta},\boldsymbol{c})}{\sum_{\boldsymbol{\theta}} q_{\texttt{True}}(\boldsymbol{x}_0,\boldsymbol{\theta},\boldsymbol{c}^*)} &= \frac{q_{\texttt{True}}(\boldsymbol{x}_0)-\sum_{\boldsymbol{\theta}} q_{\texttt{True}}(\boldsymbol{\theta},\boldsymbol{c}^*|\boldsymbol{x}_0)q_{\texttt{True}}(\boldsymbol{x}_0)}{\sum_{\boldsymbol{\theta}} q_{\texttt{True}}(\boldsymbol{x}_0,\boldsymbol{\theta},\boldsymbol{c}^*)}\\
&\le \frac{q_{\texttt{True}}(\boldsymbol{x}_0)-q_{\texttt{True}}(\boldsymbol{\theta}_0,\boldsymbol{c}^*|\boldsymbol{x}_0)q_{\texttt{True}}(\boldsymbol{x}_0)}{q_{\texttt{True}}(\boldsymbol{x}_0,\boldsymbol{\theta}_0,\boldsymbol{c}^*)}\\
&= \frac{q_{\texttt{True}}(\boldsymbol{x}_0)-q_{\texttt{True}}(\boldsymbol{\theta}_0,\boldsymbol{c}^*|\boldsymbol{x}_0)q_{\texttt{True}}(\boldsymbol{x}_0)}{q_{\texttt{True}}(\boldsymbol{\theta}_0,\boldsymbol{c}^*|\boldsymbol{x}_0)q_{\texttt{True}}(\boldsymbol{x}_0)}\\
&= \frac{1-q_{\texttt{True}}(\boldsymbol{\theta}_0,\boldsymbol{c}^*|\boldsymbol{x}_0)}{q_{\texttt{True}}(\boldsymbol{\theta}_0,\boldsymbol{c}^*|\boldsymbol{x}_0)}\\
&= \frac{\boldsymbol{\epsilon}(\boldsymbol{x}_0)}{1-\boldsymbol{\epsilon}(\boldsymbol{x_0})}\end{aligned}$$

where in the last step we use the definition of ambiguity measure $\epsilon(\boldsymbol{x}_0)=q_{\texttt{True}}(\boldsymbol{\theta}_0,\boldsymbol{c}^*|\boldsymbol{x}_0)$. Next,

$$\begin{aligned}\mathcal{A}_2 = \frac{\sum_{\boldsymbol{c}\neq\boldsymbol{c}^*} q_{\texttt{True}}(\boldsymbol{Z}_k|\boldsymbol{c})q_{\texttt{True}}(\boldsymbol{c})}{q_{\texttt{True}}(\boldsymbol{Z}_k|\boldsymbol{c}^*)q_{\texttt{True}}(\boldsymbol{c}^*)} &= \frac{\sum_{\boldsymbol{c}\neq\boldsymbol{c}^*}\sum\limits_{(\boldsymbol{\theta}_{\boldsymbol{z}_{k,r}})_{0\le r\le m_k}} q_{\texttt{True}}(\boldsymbol{Z}_k,(\boldsymbol{\theta}_{\boldsymbol{z}_{k,r}})_{0\le r\le m_k}|\boldsymbol{c})q_{\texttt{True}}(\boldsymbol{c})}{\sum\limits_{(\boldsymbol{\theta}_{\boldsymbol{z}_{k,r}})_{0\le r\le m_k}} q_{\texttt{True}}(\boldsymbol{Z}_k,(\boldsymbol{\theta}_{\boldsymbol{z}_{k,r}})_{0\le r\le m_k}|\boldsymbol{c}^*)q_{\texttt{True}}(\boldsymbol{c}^*)}\\
&\le \frac{\sum\limits_{\substack{\boldsymbol{c}\neq\boldsymbol{c}^*\\(\boldsymbol{\theta}_{\boldsymbol{z}_{k,r}})_{0\le r\le m_k}}} q_{\texttt{True}}(\boldsymbol{Z}_k,(\boldsymbol{\theta}_{\boldsymbol{z}_{k,r}})_{0\le r\le m_k}|\boldsymbol{c})q_{\texttt{True}}(\boldsymbol{c})}{q_{\texttt{True}}(\boldsymbol{Z}_k,(\boldsymbol{\theta}^*_{\boldsymbol{z}_{k,r}})_{0\le r\le m_k}|\boldsymbol{c}^*)q_{\texttt{True}}(\boldsymbol{c}^*)}\end{aligned}$$

where $\boldsymbol{\theta}^*_{\boldsymbol{z}_{k,r}}$ is the intention behind message $\boldsymbol{z}_{k,r}$. Let us investigate the denominator term more carefully:

$$\begin{aligned}\sum_{\substack{\boldsymbol{c}\neq\boldsymbol{c}^*\\(\boldsymbol{\theta}_{\boldsymbol{z}_{k,r}})_{0\le r\le m_k}}} q_{\texttt{True}}(\boldsymbol{Z}_k,(\boldsymbol{\theta}_{\boldsymbol{z}_{k,r}})_{0\le r\le m_k}|\boldsymbol{c})q_{\texttt{True}}(\boldsymbol{c}) \le{} & \sum_{\substack{\boldsymbol{c}\neq\boldsymbol{c}^*\\(\boldsymbol{\theta}_{\boldsymbol{z}_{k,r}})_{0\le r\le m_k}}} q_{\texttt{True}}(\boldsymbol{Z}_k,(\boldsymbol{\theta}_{\boldsymbol{z}_{k,r}})_{0\le r\le m_k}|\boldsymbol{c})q_{\texttt{True}}(\boldsymbol{c})+\\
& \sum_{\substack{(\boldsymbol{\theta}_{\boldsymbol{z}_{k,r}})_{0\le r\le m_k}\\\neq(\boldsymbol{\theta}^*_{\boldsymbol{z}_{k,r}})_{0\le r\le m_k}}} q_{\texttt{True}}(\boldsymbol{Z}_k,(\boldsymbol{\theta}_{\boldsymbol{z}_{k,r}})_{0\le r\le m_k}|\boldsymbol{c}^*)q_{\texttt{True}}(\boldsymbol{c}^*)\\
\le{} & \sum_{\substack{[\boldsymbol{c},(\boldsymbol{\theta}_{\boldsymbol{z}_{k,r}})_{0\le r\le m_k}]\\\neq[\boldsymbol{c}^*,(\boldsymbol{\theta}^*_{\boldsymbol{z}_{k,r}})_{0\le r\le m_k}]}} q_{\texttt{True}}(\boldsymbol{Z}_k,(\boldsymbol{\theta}_{\boldsymbol{z}_{k,r}})_{0\le r\le m_k}|\boldsymbol{c})q_{\texttt{True}}(\boldsymbol{c})\\
={} & \sum_{\substack{[\boldsymbol{c},(\boldsymbol{\theta}_{\boldsymbol{z}_{k,r}})_{0\le r\le m_k}]\\\neq[\boldsymbol{c}^*,(\boldsymbol{\theta}^*_{\boldsymbol{z}_{k,r}})_{0\le r\le m_k}]}} q_{\texttt{True}}(\boldsymbol{Z}_k,(\boldsymbol{\theta}_{\boldsymbol{z}_{k,r}})_{0\le r\le m_k},\boldsymbol{c})\\
={} & q_{\texttt{True}}(\boldsymbol{Z}_k)-q_{\texttt{True}}(\boldsymbol{Z}_k,(\boldsymbol{\theta}^*_{\boldsymbol{z}_{k,r}})_{0\le r\le m_k},\boldsymbol{c}^*)\\
={} & q_{\texttt{True}}(\boldsymbol{Z}_k)-q_{\texttt{True}}((\boldsymbol{\theta}^*_{\boldsymbol{z}_{k,r}})_{0\le r\le m_k},\boldsymbol{c}^*|\boldsymbol{Z}_k)q_{\texttt{True}}(\boldsymbol{Z}_k)\end{aligned}$$

Hence, we have:

$$\begin{aligned}\mathcal{A}_2 &\le \frac{q_{\texttt{True}}(\boldsymbol{Z}_k)-q_{\texttt{True}}((\boldsymbol{\theta}^*_{\boldsymbol{z}_{k,r}})_{0\le r\le m_k},\boldsymbol{c}^*|\boldsymbol{Z}_k)q_{\texttt{True}}(\boldsymbol{Z}_k)}{q_{\texttt{True}}(\boldsymbol{Z}_k,(\boldsymbol{\theta}^*_{\boldsymbol{z}_{k,r}})_{0\le r\le m_k}|\boldsymbol{c}^*)q_{\texttt{True}}(\boldsymbol{c}^*)}\\
&= \frac{q_{\texttt{True}}(\boldsymbol{Z}_k)\left[1-q_{\texttt{True}}((\boldsymbol{\theta}^*_{\boldsymbol{z}_{k,r}})_{0\le r\le m_k},\boldsymbol{c}^*|\boldsymbol{Z}_k)\right]}{q_{\texttt{True}}(\boldsymbol{Z}_k)q_{\texttt{True}}((\boldsymbol{\theta}_{\boldsymbol{z}_{k,r}})_{0\le r\le m_k},\boldsymbol{c}^*|\boldsymbol{Z}_k)}\\
&= \frac{1-q_{\texttt{True}}((\boldsymbol{\theta}_{\boldsymbol{z}_{k,r}})_{0\le r\le m_k},\boldsymbol{c}^*|\boldsymbol{Z}_k)}{q_{\texttt{True}}((\boldsymbol{\theta}_{\boldsymbol{z}_{k,r}})_{0\le r\le m_k},\boldsymbol{c}^*|\boldsymbol{Z}_k)}\\
&= \frac{\boldsymbol{\epsilon}(\boldsymbol{Z}_k)}{1-\boldsymbol{\epsilon}(\boldsymbol{Z}_k)}\end{aligned}$$

where in the last step we use the definition of ambiguity measure for chain-of-thoughts $\boldsymbol{Z}_k$. Combining results for $\mathcal{A}_1$ and $\mathcal{A}_2$ gives:

$$\mathcal{A} \leq \frac{\gamma^N(\boldsymbol{c}^*)\boldsymbol{\epsilon}(\boldsymbol{x}_0)}{1-\boldsymbol{\epsilon(x_0)}} \prod_{k=1}^{N} \frac{\boldsymbol{\epsilon}(\boldsymbol{Z}_k)}{1-\boldsymbol{\epsilon}(\boldsymbol{Z}_k)}$$

Similarly, we establish the bound

$$\mathcal{B} \leq \frac{\gamma^N(\boldsymbol{c}^*)\boldsymbol{\epsilon}(\boldsymbol{x}_0)}{1-\boldsymbol{\epsilon(x_0)}} \prod_{k=1}^{N} \frac{\boldsymbol{\epsilon}(\boldsymbol{Z}_k)}{1-\boldsymbol{\epsilon}(\boldsymbol{Z}_k)}$$

Hence, using the above results we have:

$$\begin{aligned}
|p_{\texttt{LLM}}((\boldsymbol{x}_r)_{1\leq r\leq m}|\boldsymbol{x}_0, \boldsymbol{Z}_{k=1}^N) - q_{\texttt{True}}((\boldsymbol{x}_r)_{1\leq r\leq m}|\boldsymbol{x}_0, \boldsymbol{c}^*)| &= \frac{|\mathcal{A} - \mathcal{B}q_{\texttt{True}}((\boldsymbol{x}_r)_{1\leq r\leq m}|\boldsymbol{x}_0, \boldsymbol{c}^*)|}{1+\mathcal{B}} \\
&\leq |\mathcal{A} + \mathcal{B}q_{\texttt{True}}((\boldsymbol{x}_r)_{1\leq r\leq m}|\boldsymbol{x}_0, \boldsymbol{c}^*)| \\
&\leq \mathcal{A} + \mathcal{B} \\
&\leq a \prod_{k=1}^{N} \frac{\boldsymbol{\epsilon}(\boldsymbol{Z}_k)}{1-\boldsymbol{\epsilon}(\boldsymbol{Z}_k)}
\end{aligned}$$

where $a = 2\frac{\gamma^N(\boldsymbol{c}^*)\boldsymbol{\epsilon}(\boldsymbol{x}_0)}{1-\boldsymbol{\epsilon}(\boldsymbol{x}_0)}$. Hence, in case Assumption 3.1 holds we have $\gamma(\boldsymbol{c}^*) = 1$ and $a = 2\frac{\boldsymbol{\epsilon}(\boldsymbol{x}_0)}{1-\boldsymbol{\epsilon}(\boldsymbol{x}_0)}$

□

B. Proof of Lemma 4.3

Proof. Let us fix $\delta \in [0, \frac{1}{2})$, then for CoT $\boldsymbol{Z}_k = (\boldsymbol{z}_{k,r})_{0\leq r\leq m_k}$ satisfying Assumption 4.2 we have:

$$\lim_{m_k\to\infty} \epsilon(\boldsymbol{Z}_k) = 0$$

then there exists $m^*_{k,\delta} \in \mathbb{N}$ such that for any $m_k \geq m^*_{k,\delta}$ we have:

$$\epsilon(\boldsymbol{Z}_k) \leq \delta$$

which finishes the proof of the Lemma. □