FAMICOM: Further Demystifying Prompts for Language Models with Task-Agnostic Performance Estimation

Anonymous ACL submission

Abstract

Language models have shown impressive in-context-learning capabilities, which allow them to benefit from input prompts and perform better on downstream end tasks. Existing works investigate the mechanisms behind this observation, and propose label-agnostic prompt metrics that can better estimate end-task performances. One popular approach is using perplexity as a way to measure models' *familiarity* with the prompt. While showing consistent improvements on in-domain tasks, we found that *familiarity* metrics such as perplexity cannot accurately estimate performance in complicated situations such as task or domain transferring scenarios. In this work, we propose a revised measure called FAMICOM, providing a more comprehensive measure for task-agnostic performance estimation. Specifically, FAMICOM combines *familiarity* with *complexity* – the inherent difficulty of end tasks, which is an important factor missing from current metrics. Experiments show that FAMICOM strongly correlates with end-task performances, producing a 0.85 Spearman's correlation, versus 0.43 of *familiarity*-only ones'. We further apply FAMICOM to automatic prompt and demonstration selection, and outperform existing methods and baselines by more than **7.0%** in accuracy.

1 Introduction

Recent works have shown that large language models (LLMs) can perform new NLP tasks by following simple instructions or seeing only a few in-context examples (Rubin et al., 2022; Brown et al., 2020; Liu et al., 2022; Wei et al., 2022).

Even though these models have demonstrated impressive capabilities to efficiently use the information in the input prompt to improve downstream task-specific performances, we still do not fully understand *why* textual prompts can contribute to end-task performances. This research direction is crucial because it will contribute to better automatic prompt design and few-shot example selection methods, in addition to helping us peek into the internal mechanisms of LLMs.

One promising way to investigate the relations between textual input and end-task performances is to propose label-agnostic measures regarding the input. If a measure strongly correlates with the end-task performance without seeing the task labels, it would suggest that such a measure is in the right direction to understand why the input prompt helps. Earlier works have proposed to use perplexity as a measure in this direction, which evaluates how familiar a model is to the prompt (Gonen et al., 2023; Kumar et al., 2023). While it is intuitive to estimate the model's expected performance indirectly by its *familiarity* to the prompts, we have found that such correlation diminishes in more complex reasoning tasks in §3. The correlation between perplexity and accuracy is no longer obvious on tasks with more reasoning steps. Therefore, we propose introducing another factor to better assess the model performance under different textual inputs.

Fig. 1 showcases examples where textual input to the LM can have similar *familiarity* yet diverse *complexity*, or diverse *familiarity* but similar *complexity*. Considering Q1 and Q2, both of which have nearly same values in model *familiarity*[1] but are intrinsically different in *complexity*. Q1 asks about the work of Walter Sinnott-Armstrong, a prominent philosopher who has extensively explored interpersonal relationships. With these knowledge stored in parameters, a LM can easily gives the correct answer of (B) families. However, Q2 asks about which movie is similar to the given five movies. It requires the model to first extract the features of each movie, find the common, and compare with each options. This process involves more reasoning steps than Q1, and questions like Q2 is therefore have lower expected accuracy. Q3 and Q4 demon-

[1] Mistral-7B, $\text{familiarity}_{\text{sim}}$ defined in §2.2.

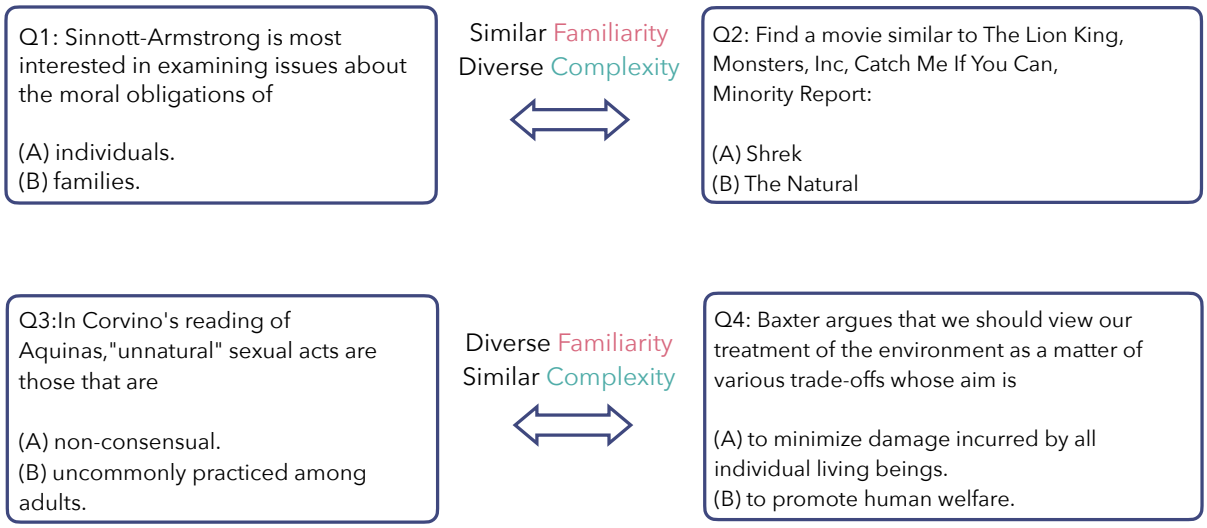

Figure 1: Prompts can have similar familiarity yet diverse complexity, or diverse familiarity but similar complexity. Estimating LM's performance merely on single factor is not enough.

strated a different situation. They are both reading comprehension questions from high school textbooks. To answer them, model can simply retrieve the parametric knowledge and select the option closest to the word distribution in training data. The reasoning steps are similar which yield close *complexity*. However, they have different *familiarities* which may results in different expectation of accuracy. These comparisons demonstrate that the *complexity* of a question in the prompt differs from the model's *familiarity* to the prompt.

Motivated by such observations, we argue that the familiarity factor alone are not robust enough across tasks and it is insufficient to estimate LM's performance merely on a single factor. An important factor is missing, which we hypothesize to be the complexity of the end task. In other words, familiarity correlations only hold for tasks that have relatively similar complexity levels, and we should factor complexity in estimation if we want to build a label-agnostic metric that will transfer across domains and tasks. To be specific, if we consider different ways to prompt the LLM to address the same task, a better prediction performance by the LLM is more likely achieved with a more familiar and less complex textual input.

To demonstrate such intuition with quantitative analyses, we first investigate several strategies to measure the intrinsic **complexity** of any given prompt for the LLM. In addition to zero-shot or few-shot prompting the model's own complexity assessment on the task prompt, we also device a practical technique to translating a given task prompt to a Transformer programming language (Zhou et al., 2023) and assess its operation-level complexity. Inspired by Gonen et al. (2023), we involved perplexity into the quantification of model **familiarity**. Moreover, as the recent study (Li et al., 2024) suggests that the mutual similarities among key words in a reasoning question also have a noticeable correlation with model performance, we combined both of the perplexity and similarity to redefine the model's familiarity to the question in the textual input.

We conducted a cross-task prompting evaluation: Given a input question from a multiple choice task, we evaluate the performance of models with cross-task demonstration. The selection of evaluation tasks span across 28 different multiple choice question answering tasks yielding exponential demonstration-input question combinations which is enough to come to a statistical conclusion. We show empirically that the proposed task-agnostic measure, namely FAMICOM (Familarity and Complexity Based Performance Estimation), has a positive correlation with model performance across a diverse set of tasks and models, which provide a insight of how prompting affect the model performance. To further demonstrate the application of our measure, we device a method on prompt selection task. Our results revealed that FAMICOM has a better guidance than similarity search or perplexity ranking which proved the effectiveness of our method.

To further demonstrate the potential impact of FAMICOM, we propose a novel *indirect in-context*

learning setting where given any end task, the model is allowed to retrieve any available annotated demonstrations that are not originally labeled for the end task. Our experimental results also suggest that FAMICOM consistently outperforms similarity search and provide better demonstration benefit to the model, allowing for more effective retrieval of indirect supervision (Yin et al., 2023a) of in-context demonstrations from the wild.

To summarize, the contribution of this work is three-fold. First, we propose a task-agnostic prompt performance estimation measure FAMICOM describing the relationship between model performance, model familiarity to a prompt and the complexity of the input question. Second, we prove the effectiveness of our formula by evaluating it with a massive cross-task prompting experiment. Finally, to elaborate the potential application of our formula, we tested it on prompt selection and an innovative indirect ICL task.

2 What measures a good prompt?

In this section, we introduce FAMICOM– a task-agnostic measure for estimating the effectiveness of prompt for LLMs. The calculation of FAMICOM is based on the hypothesis that **the lower complexity of the textual input and higher familiarity of the model with regard to the prompt correlates with better performance**. We aim to clarify that the term **complexity** in our discussion pertains specifically to the instance-level complexity of an end-task question, while **familiarity** refers to the models' familiarity with the full input prompt, which includes the task instruction, the question, and any optional reasoning processes or few-shot demonstrations.

We start with the definition and calculation of two key factors of FAMICOM: **complexity**(§2.1) and **familiarity**(§2.2). We also discuss other controllable parameters and give the concluded calculation of FAMICOM in §2.3.

2.1 Complexity Estimation

As language models scale up and become more and more capable, benchmark questions and tasks also tend to become more "complex" to effectively evaluate the capabilities and limitations of these models. These evaluations have evolved from "simple" tasks such as sentiment analysis (Socher et al., 2013), natural language inference (Williams et al., 2018), to more "complex" multi-task understanding (Hendrycks et al., 2020) and open-domain question answering (Rajpurkar et al., 2016). The testing questions evolves from determining whether a comment is positive to answering high-school biology questions.

However, defining the **complexity** of task prompts remains challenging due to the inherently ambiguous nature of natural language prompts. In this work, we make some presumptions for the approximation of prompt complexity. First, the complexity is language-model-based. In other words, for fixed prompts of a task, the measured difficulty may vary according to the choice of language model. Second, the complexity of a question is proportional to how many steps (or sub-problems) it needs to solve the problem (Khot et al., 2022). The atomic steps in this presumption may differ according to the first presumption. Third, the complexity We propose three methods to approximate the prompt complexity, namely $\textbf{complexity}_{\textbf{direct}}$, $\textbf{complexity}_{\textbf{guided}}$ and $\textbf{complexity}_{\textbf{operational}}$.

The $\textbf{complexity}_{\textbf{direct}}$ is measured by querying the model with a simple prompt $p_{complex}$: *"How many steps does it takes to solve the problem."*. Formally, for an input prompt q, the complexity of the prompt for language model L is defined as $L(p, q)$.

The $\textbf{complexity}_{\textbf{guided}}$ is similar to the $\textbf{complexity}_{\textbf{direct}}$ with extra guidance of demonstrations. We add detailed human written examples to the prompt $p_{complex}$, in which each example contains a sample question, a list of steps of the question and the final result for the steps. The examples involves questions of different steps to have a broader guidance.

The $\textbf{complexity}_{\textbf{operational}}$ is based on the pseudo programming language of RASP-L (Zhou et al., 2023). It is a human-readable programming language that can be compiled into Transformer weights. Each atomic operation in RASP-L is a calculation in a Transformers block. In this way, the steps are equivalent to the quantity of operations needed to write a RASP-L program to answer the prompt q. One key challenge of this measurement is that each prompt needs a individually composed RASP-L program which makes it impossible to scale up. To address such issue, we append human written demonstrations containing other prompts and their RASP-L programs to the evaluated prompt $p_{complex}$ and query the language model to compose the RASP-L program for complexity calculation.

For each prompt, we query the model five times and calculate the averaged complexity score as the final estimation.

2.2 Familiarity Estimation

We consider two ways of estimating the familiarity of the language model to a prompt.

The concept of **familiarity** is intended to approximate how a model is familiar with the input text. Gonen et al. (2023) proposed to use the perplexity of a prompt as a proxy for its occurrences in the data. It is based on the intuition that "a sequence that is more expected by the model is more likely to aid the model to extract the relevant information.". The experimental results showed a positive correlation between model performances and the perplexity of the prompt. We adopt perplexity as the method of $\textbf{familiarity}_{\textbf{ppl}}$ measurement.

In addition, Li et al. (2024) also found that the similarities among key words in a question, such as key entities, have a correlation with model accuracy in chained reasoning tasks. This inspired of our second method to measure familiarity. Here we focus on the "key" tokens and their mutual similarities. We first extract the "key" tokens by selecting Top-K tokens that have the highest perplexities in a prompt. Inspired by the counterfactual theories (Lewis, 2013) and prior studies on probing token salience (Kaushik et al., 2019; Wang et al., 2021; Qi et al., 2021), we adopt the intuition that any word that leads to high entropy given the previous contexts are likely important words, such as preposition or pronounce. Those words with lower probability can be conveying salient information of the prompt. Among the selected tokens, we compute the mutual cosine similarities of the embedding of the tokens in the input text and use the averaged similarities as $\textbf{familiarity}_{\textbf{sim}}$.

2.3 Combining Familiarity and Complexity

Based on the previous analysis, the measured factors of FAMICOM can be incorporated in the following polynomial approximation which reflect our hypothesis that the lower complexity of the textual input and higher familiarity of the model with regard to the prompt correlates with better performance:

$$\text{FAMICOM} = f^a \cdot c^{-b} \quad (1)$$

where f, c are familiarity and complexity respectively. a, b are positive hyper-parameters that we will investigate in §3 where we also show that both properly chosen a, b values that has a positive correlation with model performance.

3 Validation Analysis

This section presents the analytical experiment for validating the FAMICOM measure. We first introduce the task settings and the data used for analysis in §3.1, and subsequently elaborate the model configurations and evaluation protocol in §3.2[2]. Finally, §3.3 presents the experimental results and the analysis.

3.1 Task Description

Since FAMICOM is a task-agnostic performance estimation, we design a **cross-task prompting** scenario for the validation. Given a prompt from a specific task, we randomly pair it with the demonstrations of other tasks to evaluate the model performance. The FAMICOM measure is computed for each of the input and we conduct cross-task prompting evaluation in a scale to expose the statistical relationship between the measure and the true model performance.

The 28 evaluation task pool are sampled from MMLU, BigBench together with StrategyQA. For each of these tasks, we retrieve its Chain-of-Thought prompts from the CoT Hub (Fu et al., 2023), which is an open-source project for measuring LLMs' reasoning capabilities with resourceful prompts for various benchmarks. During inference, a question is paired with three randomly selected CoT demonstrations for other tasks in the pool. We randomly sample 200 instances for each task, along with three chain-of-thought demonstrations per task. This preparation method yields approximately 100,000 question-demonstration pairings for our experiment, thereby guaranteeing its statistical significance.

Every sampled task is multiple choice question answering[3] and we follow the prompt format in the CoT Hub where each question is given as "[Question] Options:[list of options]". Each option is labeled with an alphabetical letter, for example "(A) Jupiter", which helps locating the choice from the generative LMs' response.

It is noteworthy that these questions differ in the number of options which may introduce biases in

[2]The configurations discussed here also apply to experiments in §4 and §5

[3]StrategyQA has a "Yes or No" question type, which can be easily formatted as "Options: (A) Yes (B) No"

Correlation		
Measure	**Spearman Corr.**	**P-value**
Familiarity	0.426	2e-3
Complexity	0.695	2e-8
FAMICOM	0.848	7e-9

Table 1: Enhanced Table of Spearman Correlations and P-values for Model Performance against Proposed Factors. As we can see, FAMICOM surpasses single usage of familiarity or complexity.

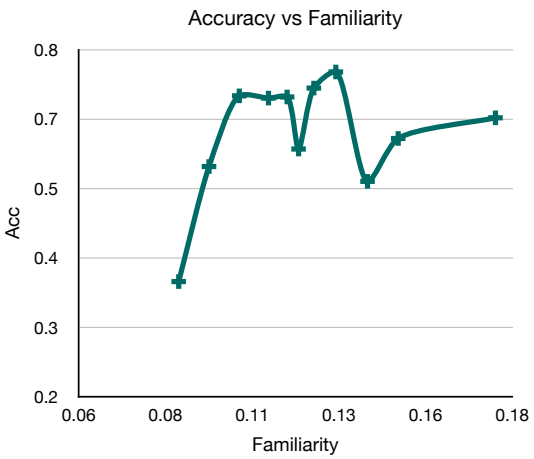

Figure 2: Familiarity itself doe not show a clear correlation with performance.

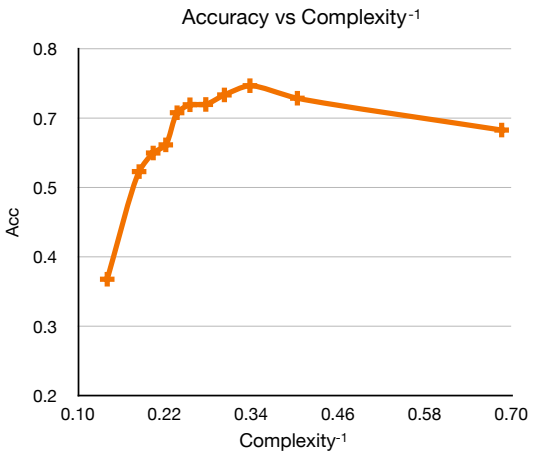

Figure 3: Complexity demonstrates a better correlation with performance but needs further calibration.

evaluation, especially when computing the complexity. A question with more options may be regarded as more complex by the model. To alleviate this bias, each of the sampled questions is reduced to a binary option question. To do so, the correct option is retained while the other is a randomly picked false option. The sequence of options is shuffled to mitigate the potential biases.

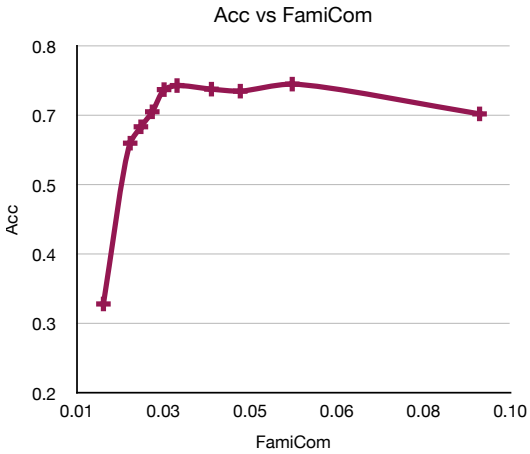

Figure 4: Compared to Figs. 2 and 3, the trend line is much more consistent, and the improvement is more steady.

Instructional prompt selection					
Method	AG News	Imdb	Cola	Emotion	Offensive
Ori	67.7	84.1	39.8	44.3	67.0
SPELL	78.0	86.0	**43.4**	46.8	65.1
FAMICOM	**78.9**	**87.6**	41.0	**51.4**	**67.1**

Table 2: Prompt selection results given in accuracy. The best result for each task is in **bold**. As we can see, FAMICOM consistently performs better than Ori and SPELL on almost all of the datasets.

3.2 Model configuration and Evaluation Protocol

We evaluate on open-source LLMs in different sizes due to the fact that this scale of evaluation is too costly on closed-source LLMs and the token distributions or perplexities are not accessible in their APIs. The tested models include Phi-3-mini-128k, Mistral-7B-instruct-v0.2, and Llama-2-13b-chat to validate our hypothesis across different model sizes. All models run with a temperature $\tau = 0.8$ and all experiments are run on a machine with 8*Nvidia ADA 6000 GPUs.

We evaluate the **accuracy** of LLMs towards the correct answer. Since all the prepared data are multiple-choice QA with labeled options, a response is considered correct if it contains the correct label. We also applied the self-consistency strategy with majority voting in five runs for all experiments (Wang et al., 2022a).

3.3 Validation Analysis Results

We show that our hypothesis hold and FAMICOM measure has a correlation with the model performance. Fig. 2, Fig. 3, Fig. 4 provide visualizations

between the model performance and familiarity, complexity and FAMICOM measures respectively. Each dot in graph represents a cluster of instances with similar measured values indicated by the x-axis. We calculate the averaged accuracy of the model given instances in each cluster and the y-axis indicates the accuracy. These figures adopts **familiarity**$_{\textbf{sim}}$ and the **guided complexity** which are tested to have better correlation. The token similarities are computed over tokens with top 20 perplexities. The coefficients are fixed as $f^1 \cdot c^{-1}$ after the brute-force hyper-parameter search[4]. Both a and b range from -2 to +2 with a step size of 0.5, leading to 81 combinations. Tab. 1 depicts the Spearman correlation of model performance against these measures.

It is shown in Fig. 2 that **familiarity** itself does not highly correlates with the model performance. This conclusion can also be drawn from its Spearman Correlation $\rho = 0.426 < 0.5$ with $p = 0.002$. The **complexity** factor, on the other hand, has a positive correlation with model performance with $\rho = 0.695$ and $p << 0.00625$[5]. This meets the intuition that model perform worse on more complex questions and also validate the effectiveness of our approximation of **complexity**. The FAMICOM measure has a better monotonic correlation with model performances, whose Spearman correlation value reaches 0.848 with $p << 0.00625$. It can then be concluded that our hypothesis hold in a large scale of experiments and FAMICOM can effectively estimate the performance of prompts.

4 The Prompt Selection Task

In this section, we discuss a prompt selection task to further explore the effectiveness and application of FAMICOM. §4.1 introduces the task, §4.2 discusses the evaluation datasets and the following §4.4 analyzes the experimental results. In this task, we only examine the Mistral model[6] and using the same set of evaluation metrics as §3.2.

4.1 Task Description

To inspect if FAMICOM can properly estimate model performance on different prompts as observed, we conduct evaluation on SUPER-NATURAL INSTRUCTIONS(SUP-NATINST for short, Wang et al. (2022b)). This benchmark contains expert-contributed instructional prompts for 1,616 NLP tasks for evaluating the zeo-shot performance of LLMs. In SUP-NATINST, every task is provided with a instruction that includes the task's definition for transforming an input text into a specified output, along with multiple examples to illustrate both the desired and the undesired results. We only use the task definition as the instructional prompt and use GPT-4 to generate four more task descriptions in SUP-NATINST style for each task involved.

To calculate the FAMICOM measure for each candidate prompt, we pair it with the query question and compute their combined **familiarity**$_{\textbf{sim}}$. The salient words in both of the question and the instructional prompt can therefore be involved in calculation.

4.2 Datasets

We choose five discriminative tasks tested in (Gonen et al., 2023) from Huggingface Dataset[7] for a broad evaluation. These includes: (i)**AG News** (Zhang et al., 2015) for news topic classification (ii) **IMDB** (Maas et al., 2011) for sentiment analysis on movie reviews (iii) **GLUE-Cola** (Warstadt et al., 2019) for grammatical acceptability discrimination; (iv) **Emotion** (Saravia et al., 2018) for emotion classification on tweets; (v) **Tweet Offensive** (Barbieri et al., 2020) for offensive tweet discrimination. Each dataset is sampled 1,000 test data for a balanced evaluation of prompt selection.

4.3 Baselines

We compare our method with two baseline: **SPELL** (Gonen et al., 2023) and the original prompt(**Ori**) in SUP-NATINST. **SPELL**(**S**electing **P**rompts by **E**stimating **LM** **L**ikelihood) selects the prompts with the lowest perplexity for a given task after manually creating a set of candidate prompts and expanding them to hundred-scale using automatic paraphrasing. The original SUP-NATINST prompt is expert-created for each task.

4.4 Prompt Selection Results

Tab. 2 shows the performances of different methods. FAMICOM consistently achieves the best performance on most of the tasks except for Cola[8].

[4]We therefore adopt these coefficients with the metric combination in the following experiments

[5]Bonferroni Test following Gonen et al. (2023)

[6]Following Liu et al. (2024)

[7]https://huggingface.co/docs/datasets/index

[8]This can be attributed to the questions in Cola focusing on grammatical acceptability, which can be directly measured by perplexity while **familiarity**$_{\textbf{sim}}$ captures the token similarities.

Indirect in-context learning							
	#In-context examples	K=3		K=5		K=7	
Source	Task	FAMICOM	KNN	FAMICOM	KNN	FAMICOM	KNN
MMLU	medical-genetics	82.6	76.8	79.7	69.6	82.6	79.7
	professional-psychology	78.3	63.6	76.9	71.3	79.7	73.4
	formal-logic	66.7	54.9	71.6	62.7	68.6	63.7
	moral-disputes	81.9	70.8	81.9	77.8	83.3	75.7
	public-relations	78.8	71.8	76.5	76.5	81.2	70.6
	computer-security	85.9	76.9	79.5	82.1	83.3	82.1
	astronomy	87.2	68.8	88.1	78.0	85.3	83.5
	abstract-algebra	63.7	50.0	63.7	50.0	65.0	57.5
	nutrition	85.2	81.2	87.2	85.9	89.3	89.9
	high-school-biology	82.9	76.0	85.6	80.8	82.9	88.4
	business-ethics	76.7	75.3	79.5	78.1	71.2	78.1
	StrategyQA	57.0	56.0	62.5	58.5	60.5	55.5
BIG	tracking-shuffled-objects-seven-objects	52.5	54.5	47.5	56.0	47.5	57.0
	formal-fallacies	25.5	16.5	30.5	14.5	29.0	17.5
	hyperbaton	70.0	69.5	75.0	74.5	76.5	74.0
	tracking-shuffled-objects-three-objects	47.0	45.5	50.0	50.5	48.5	54.5
	logical-deduction-five-objects	83.0	77.0	78.5	83.0	80.5	78.0
	Macro-Avg	70.9	63.8	71.4	67.6	71.5	69.4

Table 3: Indirect ICL Results of FAMICOM and KNN with K=3,5,7 examples. FAMICOM outperforms KNN on the majority of tasks across various K values, indicating that FAMICOM outperforms similarity-based familiarity for prompt performance estimation. Only Mistral results are reported, similar findings can be drawn from the results of the other two models in Appx. §A

.

Specifically, FAMICOM improves the accuracy of Mistral-7B by 4.6% on average across all tasks, surpassing the improvement of 3.3% offered by SPELL. Although the **complexity** factor is normalized under this experiment setting, these results suggest that **familiarity**$_{\textbf{sim}}$ yields better performance estimation than perplexity alone.

5 Indirect In-context Learning

In this section, we demonstrate a novel Indirect In-context Learning (ICL) task to further showcases the practicality of FAMICOM. We start by explaining the task (§5.1), discuss the data preparation (§5.2), and demonstrate the baseline(§5.3). The model configurations follow §3.2 and the results are shown in §5.4.

5.1 Task Description

The term "indirect" in the name suggests that the focus of this task setting is not on task-specific demonstration selection. Instead, indirect ICL considers a generalized pool of examples from different tasks together, and the goal is to identify most contributive examples for each specific input. Each example comprises a question and its corresponding chain-of-thought response. Indirect ICL generalizes regular ICL to allow for any available labeled examples that are not necessarily dedicated to the end-task, including those that may provide helpful incidental supervision (Yin et al., 2023b) to aid in-context learning. This task also addresses a practical situation where searching for the most beneficial demonstration when the user query or the end task is not known beforehand.

5.2 Data Preparation

In this task, we collect examples from each of 28 tasks in a pool. Each task contributes three examples and each example consists of a question and its chain-of-thought response. The pool finally gathers 84 CoT examples for later inference. We evaluate the model on 17 different tasks from MMLU, BigBench, and StrategyQA. For each task, 200 examples at most are randomly sampled for evaluation. We retrieved the CoT demonstration from CoT Hub (Fu et al., 2023) and formatted the CoT examples as well as the testing questions in the same way described in §3.1.

5.3 Baselines

We compare FAMICOM with K-Nearest Neighbors(KNN), which ranks and selects the candidate examples based on their similarities with the input prompt. To be specific, we extract the embeddings of the last tokens of an example and the input prompt, calculate the cosine similarity between them as the distance. Same as FAMICOM, we select top $K = 3, 5, 7$ in this experiment for a broader inspection.

5.4 Indirect ICL Results

Tab. 3 provides detailed experimental results for Indirect ICL. FAMICOM outperforms KNN on the majority of tasks across various K values. Specifically, the macro-average score of FAMICOM surpasses that of KNN by 7.1%, 3.8%, and 2.1% for K=3, 5, and 7, respectively. The improvement is more pronounced as the number of examples decreases, indicating that FAMICOM is particularly beneficial in few-shot settings. The results showed that FAMICOM provides better guidance to the model than similarity search, proving its effectiveness in correlating the input features to the output performance.

6 Related Work

Zero-shot prompt tuning. There are numerous studies trying to refine the prompt for better outcomes from LMs in recent years. They can be roughly classified in two categories: prompt selection and prompt rewriting. These tasks focus on retrieving or creating prompts that will improve the zero-shot performance of LMs. Based on gradient-guided search, Shin et al. (2020) leverages an automated method to project a prompt onto discrete phrases to improve masked LM performances. Deng et al. (2022) applied reinforcement leaning to optimize discrete prompts.

To adapt to general-purpose LLMs, where the user query may be unknown, researches have been conducted to exploit the intrinsic word distribution of LLMs obtained from the training data and propose familiarity-based prompt refinement methods (Gonen et al., 2023; Wang et al., 2023; Lee et al., 2024). These works are based on a phenomenon that lower perplexities of prompts are preferred by LMs to perform better across a wide range of tasks.

Few-shot demonstration selection. Furthermore, some researches focus on in-context-learning (ICL) example selection to improve model performance with few-shot learning (Iter et al., 2023; Lee et al., 2024; Lu et al., 2022). Works have also been done to further investigate what are the driving factors that contributes to the different performances in ICL (Min et al., 2022), which is closer to the scope of this work. While the last sections of this work lie in zero-shot and few-shot prompt tuning, the main target on this work is stepping forward to study the essence of prompting: the heuristics of word distributions in LMs are not enough to estimate the prompt performance, the complexity of the query shall also be considered. Our paper proposed a measure that will better describe the effectiveness of prompt, without any training effort.

Analytical studies on model performance. Several studies have examined how the distribution of training data affects model performance on specific tasks. Gonen et al. (2023) suggested utilizing perplexity to estimate the distribution of input queries within the training dataset. Razeghi et al. (2022) explored the capability of LMs to process numerical tasks involving terms infrequently encountered during pre-training. They found that models perform better on instances where the terms are more common in the training data. While these studies suggested using word frequency to link pre-training data with task performance, FAMICOM emphasized that the **complexity** of the input question is another contributor to the estimation of performance.

7 Conclusion

In this work, we propose FAMICOM, a label-agnostic metric that predicts whether an input prompt can lead to high downstream end-task performance or not, without the need to understand what the task is about. Different from previous methods (Min et al., 2023; Gonen et al., 2023) that only involve perplexity as familiarity, FAMICOM considers two dimensions: familiarity and complexity, which are shown both intuitively and empirically to be essential when building such metrics and solving cross-domain tasks. On the one hand, FAMICOM provides insights into the internal mechanisms of the in-context-learning capabilities of large language models; on the other hand, it can be used as an automatic prompt selection method because it predicts whether a prompt can produce high end-task performance. Our work motivates future works on building more fine-grained metrics using familiarity and complexity.

Ethical Considerations

Innovations in technology often encounter the moral challenge of dual-use: the same development can bring both benefits and risks. With the probing method and benchmark presented in this paper, the line between beneficial and harmful usage largely depends on data. Proper utilization of the technology necessitates the legal and ethical acquisition of input text corpora and other modalities of inputs. Legal frameworks and standards are crucial for ensuring proper data use and for granting individuals the right to remove their data. In the absence of such regulation, the ethical use of data depends on the responsibility of technology users. Additionally, the generated and analysis data may exhibit biases that systematically affect accuracy for less represented groups or in new areas, potentially resulting in performance disparities among sub-populations based on ethnicity, race, gender, and other factors. Moreover, the effectiveness of trained systems diminishes when applied to new data that deviates from their training set. Therefore, issues of generalizability and fairness must be thoroughly examined when implementing the methodologies discussed in this paper. It is crucial to embed ethical considerations as fundamental principles at each stage of system development, ensure high levels of transparency and clarity in data, algorithms, models, and functions within the system, release software under open-source licenses to facilitate public scrutiny and investigate strategies to safeguard at-risk groups.

Limitations

Our work proposes that he inherent perceived complexity of the end tasks should be included in the estimation of prompt performance, together with the familiarity of LMs to the prompt. To this end, we identify the following limitations.

Limited validation experiments We only conducted the validation experiments on Mistral-7B as the massive data volume to process in this experiment. With more effort in the future, we can extend to more families of large language models.

Limited demonstrations The current Indirect ICL is still a pilot study and does not include a large scale of demonstrations? Future works may benefit from wider range of tasks involved in the indirect ICL experiments and bring more insights in the prompting studies.

References

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Mingkai Deng, Jianyu Wang, Cheng-Ping Hsieh, Yihan Wang, Han Guo, Tianmin Shu, Meng Song, Eric Xing, and Zhiting Hu. 2022. RLPrompt: Optimizing discrete text prompts with reinforcement learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3369–3391, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yao Fu, Litu Ou, Mingyu Chen, Yuhao Wan, Hao Peng, and Tushar Khot. 2023. Chain-of-thought hub: A continuous effort to measure large language models' reasoning performance. *Preprint*, arXiv:2305.17306.

Hila Gonen, Srini Iyer, Terra Blevins, Noah Smith, and Luke Zettlemoyer. 2023. Demystifying prompts in language models via perplexity estimation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10136–10148, Singapore. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *Preprint*, arXiv:2009.03300.

Dan Iter, Reid Pryzant, Ruochen Xu, Shuohang Wang, Yang Liu, Yichong Xu, and Chenguang Zhu. 2023. In-context demonstration selection with cross entropy difference. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1150–1162, Singapore. Association for Computational Linguistics.

Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. 2019. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*.

Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. 2022. Decomposed prompting: A modular

approach for solving complex tasks. In *The Eleventh International Conference on Learning Representations*.

Aswanth Kumar, Ratish Puduppully, Raj Dabre, and Anoop Kunchukuttan. 2023. CTQScorer: Combining multiple features for in-context example selection for machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7736–7752, Singapore. Association for Computational Linguistics.

Yoonsang Lee, Pranav Atreya, Xi Ye, and Eunsol Choi. 2024. Crafting in-context examples according to lms' parametric knowledge. *Preprint*, arXiv:2311.09579.

David Lewis. 2013. *Counterfactuals*. John Wiley & Sons.

Bangzheng Li, Ben Zhou, Fei Wang, Xingyu Fu, Dan Roth, and Muhao Chen. 2024. Deceptive semantic shortcuts on reasoning chains: How far can models go without hallucination? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.

Qin Liu, Fei Wang, Nan Xu, Tianyi Yan, Tao Meng, and Muhao Chen. 2024. Monotonic paraphrasing improves generalization of language model prompting. *Preprint*, arXiv:2403.16038.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

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

Sewon Min, Weijia Shi, Mike Lewis, Xilun Chen, Wen-tau Yih, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2023. Nonparametric masked language modeling. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2097–2118, Toronto, Canada. Association for Computational Linguistics.

Fanchao Qi, Yangyi Chen, Mukai Li, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2021. ONION: A simple and effective defense against textual backdoor attacks. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9558–9566, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Yasaman Razeghi, Robert L Logan IV, Matt Gardner, and Sameer Singh. 2022. Impact of pretraining term frequencies on few-shot numerical reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 840–854, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2022. Learning to retrieve prompts for in-context learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2655–2671, Seattle, United States. Association for Computational Linguistics.

Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. 2018. CARER: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3687–3697, Brussels, Belgium. Association for Computational Linguistics.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, Online. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Fei Wang, Kexuan Sun, Jay Pujara, Pedro Szekely, and Muhao Chen. 2021. Table-based fact verification

with salience-aware learning. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4025–4036, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022a. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Kuntal Kumar Pal, Maitreya Patel, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Savan Doshi, Shailaja Keyur Sampat, Siddhartha Mishra, Sujan Reddy A, Sumanta Patro, Tanay Dixit, and Xudong Shen. 2022b. Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zezhong Wang, Luyao Ye, Hongru Wang, Wai-Chung Kwan, David Ho, and Kam-Fai Wong. 2023. ReadPrompt: A readable prompting method for reliable knowledge probing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7468–7479, Singapore. Association for Computational Linguistics.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Wenpeng Yin, Muhao Chen, Ben Zhou, Qiang Ning, Kai-Wei Chang, and Dan Roth. 2023a. Indirectly supervised natural language processing. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 6: Tutorial Abstracts)*.

Wenpeng Yin, Muhao Chen, Ben Zhou, Qiang Ning, Kai-Wei Chang, and Dan Roth. 2023b. Indirectly supervised natural language processing. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 6: Tutorial Abstracts)*, pages 32–40, Toronto, Canada. Association for Computational Linguistics.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.

Hattie Zhou, Arwen Bradley, Etai Littwin, Noam Razin, Omid Saremi, Josh Susskind, Samy Bengio, and Preetum Nakkiran. 2023. What algorithms can transformers learn? a study in length generalization. *Preprint*, arXiv:2310.16028.

A Appendix

Here we present experimental results of Phi-3 and Llama-2 on Indirect ICL task introduced in §5 in Tab. 4 and Tab. 5 respectively. For Phi-3, FAMICOM's averaged accuracy of all the tasks surpasses KNN's when K>3. For LLama-2, FAMICOM yields better performances on all the settings.

Indirect in-context learning: Phi-3							
	#In-context examples	K=3		K=5		K=7	
Source	Task	FAMICOM	KNN	FAMICOM	KNN	FAMICOM	KNN
MMLU	medical-genetics	89.9	88.4	94.2	91.3	92.8	88.4
	professional-psychology	88.1	86	84.6	83.9	88.1	86
	formal-logic	69.6	74.5	75.5	75.5	78.4	67.6
	moral-disputes	83.3	83.3	86.1	84	84.7	81.9
	public-relations	77.6	85.9	82.4	78.8	87.1	77.6
	computer-security	94.9	87.2	92.3	92.3	89.7	89.7
	astronomy	91.7	90.8	90.8	92.7	92.7	90.8
	abstract-algebra	68.8	71.2	67.5	72.5	71.2	66.2
	nutrition	87.2	88.6	91.9	87.9	91.3	88.6
	high-school-biology	92.5	91.8	91.8	93.2	94.5	93.8
	business-ethics	87.7	84.9	89.0	91.8	87.7	87.7
	StrategyQA	60.0	61.5	64.0	56.0	66.5	55.0
BIG	tracking-shuffled-objects-seven-objects	84.0	89.5	84.5	87.0	87.5	83.0
	formal-fallacies	62.5	70.5	63.0	70.5	63.0	63.5
	hyperbaton	98.0	97.0	97.0	97.5	95.5	97.0
	tracking-shuffled-objects-three-objects	83.5	90	87.5	84.0	86.0	89.0
	logical-deduction-five-objects	82.0	77.0	83.0	75.0	82.5	77.0
	Macro-Avg	82.4	83.4	83.8	83.2	84.7	81.3

Table 4: Indirect ICL Results on phi-3-mini-128k-instruct by FAMICOM and KNN with K=3,5,7 examples. FAMICOM outperforms KNN on the majority of tasks when K>3.

Indirect in-context learning: Llama-2							
	#In-context examples	K=3		K=5		K=7	
Source	Task	FAMICOM	KNN	FAMICOM	KNN	FAMICOM	KNN
MMLU	medical-genetics	92.8	81.2	88.4	84.1	89.9	85.5
	professional-psychology	86.7	81.8	88.8	75.5	91.6	79.0
	formal-logic	82.4	72.5	85.3	79.4	81.4	74.5
	moral-disputes	95.8	82.6	96.5	76.4	95.1	78.5
	public-relations	90.6	83.5	95.3	80.0	94.1	77.6
	computer-security	92.3	84.6	94.9	88.5	94.9	74.4
	astronomy	93.6	84.4	94.5	77.1	92.7	78.9
	abstract-algebra	87.5	67.5	92.5	72.5	90.0	76.2
	nutrition	92.6	85.2	96.0	81.9	93.3	81.2
	high-school-biology	97.3	78.1	96.6	74.0	96.6	80.8
	business-ethics	90.4	87.7	87.7	83.6	89.0	74.0
	StrategyQA	59.5	69.5	58.5	62.0	61.5	67.0
BIG	tracking-shuffled-objects-seven-objects	58.5	63.0	69.5	65.0	63.0	56.0
	formal-fallacies	72.0	71.5	65.5	79.0	83.0	78.0
	hyperbaton	56.0	96.5	58.5	91.5	51.0	91.5
	tracking-shuffled-objects-three-objects	75.5	70.5	76.5	78.5	85.0	78.5
	logical-deduction-five-objects	79.0	61.0	85.0	76.5	82.5	73.5
	Macro-Avg	82.5	77.7	84.1	77.8	84.4	76.8

Table 5: Indirect ICL Results on Llama-2-13b-chat by FAMICOM and KNN with K=3,5,7 examples. FAMICOM outperforms KNN on the majority of tasks across various K.