On the In-context Generation of Language Models

Zhongtao Jiang[1,2], **Yuanzhe Zhang**[1,2], **Kun Luo**[1,2,3], **Xiaowei Yuan**[1,2,3]
Jun Zhao[1,2], **Kang Liu**[1,2,4*]
[1]The Key Laboratory of Cognition and Decision Intelligence for Complex Systems,
Institute of Automation, Chinese Academy of Sciences
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences
[3]Beijing Academy of Artificial Intelligence, and [4]Shanghai Artificial Intelligence Laboratory
{zhongtao.jiang, yzzhang, jzhao, kliu}@nlpr.ia.ac.cn
{luokun2024, yuanxiaowei2022}@ia.ac.cn

Abstract

Large language models (LLMs) are found to have the ability of in-context generation (ICG): when they are fed with an in-context prompt concatenating a few somehow similar examples, they can implicitly recognize the pattern of them and then complete the prompt in the same pattern. ICG is curious, since language models are usually not explicitly trained in the same way as the in-context prompt, and the distribution of examples in the prompt differs from that of sequences in the pretrained corpora. This paper provides a systematic study of the ICG ability of language models, covering discussions about its source and influential factors, in the view of both theory and empirical experiments. Concretely, we first propose a plausible latent variable model to model the distribution of the pretrained corpora, and then formalize ICG as a problem of next topic prediction. With this framework, we can prove that the repetition nature of a few topics ensures the ICG ability on them theoretically. Then, we use this controllable pretrained distribution to generate several medium-scale synthetic datasets (token scale: 2.1B~3.9B) and experiment with different settings of Transformer architectures (parameter scale: 4M~234M). Our experimental results further offer insights into how the data and model architectures influence ICG.

1 Introduction

As the data and parameter scale continue to increase, large language models (LLMs) have shown strikingly emergent abilities (Wei et al., 2022a), where one of the most exciting ones is in-context learning (ICL) (Brown et al., 2020). Given an *in-context prompt* that concatenates a few *in-context examples* and a query input, LLMs can somehow implicitly guess the "topic" of those examples and complete the query input in the desired way. LLMs can actually achieve more: they can imitate those examples using the topic learned in context (Meyerson et al., 2023) to generate new plausible examples, as shown in Figure 1. This in-context generation (ICG) ability forms the foundation of multiple few-shot prompting methods like ICL and other variants like Chain-of-thoughts (Wei et al., 2022b).

Intuitively, one might comment that LLMs learn the ICG ability from data in the *repetition mode*, which roughly refers to a type of text concatenated with sequences under the same topic. This is true to some extent. As known, typical pretrained corpora contain (e.g. CommonCrawl[1]) internet data which has an unneglectable portion of array-page data such as IMDB[2] review pages. After preprocessing, these pages are converted to repetition mode data, as shown in Figure 1a. However, this isn't enough to explain the ICG ability, since LLMs can also generate sequences of in-context learned topics that don't appear to repeat and even are unseen in the pretrained corpora. For example, Figure 1 shows sampled completions of Llama2-13B (Touvron et al., 2023) given in-context prompts of different types of topics:

1. The first one is a *repeated topic* called "movie review" (Figure 1a), where Llama2-13B naturally has the ICG ability on it since this topic appears to repeat in the pretrained corpora as mentioned.

2. The second type *nonrepeated topic* refers to those that appear in the pretrained corpora but never repeat, e.g., `forward` method in any class inherited from `nn.Module` of `Pytorch` (Paszke et al., 2019) code (Figure 1b). However, Llama2-13B can also generate plausible codes of `forward` method when prompting it with a few ones.

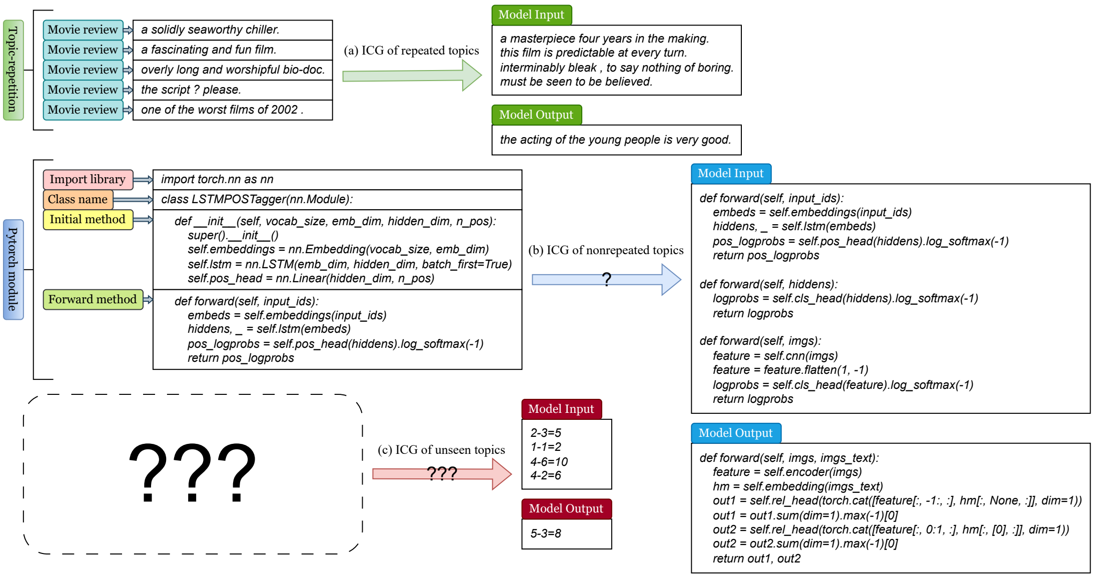
3. The last type *unseen topic* includes those that never appear in the pretrained corpora. For example, "unnatural addition" generates 2-digit arith-

[*] Corresponding Author

[1] https://commoncrawl.org

[2] https://www.imdb.com

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 10169–10187
November 12-16, 2024 ©2024 Association for Computational Linguistics



a solidly seaworthy chiller.
a fascinating and fun film.
overly long and worshipful bio-doc.
the script ? please.
one of the worst films of 2002 .

a masterpiece four years in the making.
this film is predictable at every turn.
interminably bleak , to say nothing of boring.
must be seen to be believed.

the acting of the young people is very good.

```
import torch.nn as nn
class LSTMPOSTagger(nn.Module):
    def __init__(self, vocab_size, emb_dim, hidden_dim, n_pos):
        super().__init__()
        self.embeddings = nn.Embedding(vocab_size, emb_dim)
        self.lstm = nn.LSTM(emb_dim, hidden_dim, batch_first=True)
        self.pos_head = nn.Linear(hidden_dim, n_pos)
    def forward(self, input_ids):
        embeds = self.embeddings(input_ids)
        hiddens, _ = self.lstm(embeds)
        pos_logprobs = self.pos_head(hiddens).log_softmax(-1)
        return pos_logprobs
```

```
def forward(self, input_ids):
    embeds = self.embeddings(input_ids)
    hiddens, _ = self.lstm(embeds)
    pos_logprobs = self.pos_head(hiddens).log_softmax(-1)
    return pos_logprobs

def forward(self, hiddens):
    logprobs = self.cls_head(hiddens).log_softmax(-1)
    return logprobs

def forward(self, imgs):
    feature = self.cnn(imgs)
    feature = feature.flatten(1, -1)
    logprobs = self.cls_head(feature).log_softmax(-1)
    return logprobs
```

```
def forward(self, imgs, imgs_text):
    feature = self.encoder(imgs)
    hm = self.embedding(imgs_text)
    out1 = self.rel_head(torch.cat([feature[:, -1:, :], hm[:, None, :]], dim=1))
    out1 = out1.sum(dim=1).max(-1)[0]
    out2 = self.rel_head(torch.cat([feature[:, 0:1, :], hm[:, [0], :]], dim=1))
    out2 = out2.sum(dim=1).max(-1)[0]
    return out1, out2
```

```
2-3=5
1-1=2
4-6=10
4-2=6
```

```
5-3=8
```

Figure 1: ICG examples (generated from Llama2-13B) of different kinds of topics.

metic expressions that input subtraction but expect addition (like "1-1=2"), which is intuitively believed to never be seen in the pretrained corpora (Rong, 2021). However, Llama2-13B can also recognize this topic and generate plausible sequences in context, as shown in Figure 1c.

The above results show that LLMs can generalize the repetition mode to nonrepeated and unseen topics. We term this phenomenon as the topic generalization of ICG, abbreviated as *ICG-generalization*. ICG-generalization is curious because LLMs are not explicitly trained in the way they test. The biggest challenge of studying ICG and its generalization is that the true pretrained distribution is not accessible. As a result, we don't know the topic of a sequence span or whether it appears to repeat, making it difficult to evaluate the ICG abilities of LLMs. To address this, we turn to synthetic data generated from a known and controlled pretrained distribution (Bowman et al., 2015; McCoy et al., 2018; White and Cotterell, 2021; Xie et al., 2021; Papadimitriou and Jurafsky, 2023; Jumelet and Zuidema, 2023). The distribution is a hierarchical latent variable model (LVM) as shown in Figure 2, where a document is guided by two kinds of latent variables. The distribution is not only plausible to explain the true pretrained data but also convenient for analysis since it decouples different levels of uncertainties.

Through the proposed pretrained distribution, we can naturally formalize ICG as a problem of next topic prediction, and then conduct mathematical analysis. We first theoretically prove that (Theorem 1), under some mild assumptions, if the language model fits the pretrained distribution well, then it's guaranteed to have the ICG ability on repeated topics in terms of convergence in probability. As a result, the ICG distribution (i.e., the generative distribution conditioned on the in-context prompt) converges to the true topic-paragraph distribution in probability. Next, we study ICG-generalization via exhaustive experiments, revealing that ICG-generalization is caused by factors of both data and models. Concretely, we use the controllable pretrained distribution to generate several synthetic datasets (token scale: 2.1B~3.9B), and train Transformer (Vaswani et al., 2017) language models with different settings (parameter scale: 4M~234M). Experiments show that data compositionality, proportion of repeated topics, Transformer's parameter scale, and window size play crucial roles in enabling ICG-generalization, while the data topic uncertainty and Transformer's attention head size have few influences[3]. Our study provides insights to better understanding the ICG ability and LLMs.

2 Settings

2.1 Pretrained Distribution

We assume the pretrained distribution is a hierarchical LVM as shown in Figure 2, where a document is generated via the following steps: 1) Draw a latent

[3] These results are consistent with previous works about attention head pruning (Michel et al., 2019; Voita et al., 2019) and the importance of large attention size (Ratner et al., 2023).

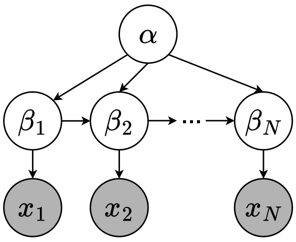


Figure 2: Bayesian network of the pretrained distribution, where the non-shaded nodes are latent variables.

mode $\alpha \in A$ from the mode prior $p(\alpha)$. 2) Draw a latent outline $\beta_{1:N} \in B^N$ containing topics of different paragraphs from the Markov mode-outline distribution $p(\beta_{1:N}|\alpha)$ parameterized by the mode α. 3) Sample each paragraph $x_i \in \Sigma^*$ (Σ is the vocabulary) individually from the topic-paragraph distribution $p(x|\beta_i)$, and concatenate them with delimiters. The joint distribution of this LVM is:

$$p(\alpha, \beta_{1:N}, x_{1:N}) = p(\alpha)p(\beta_{1:N}|\alpha)\prod_{i=1}^{N} p(x_i|\beta_i) \tag{1}$$

This distribution is plausible because: 1) It has a clear realistic interpretation of how humans write documents. Generally, humans would first determine the literature genre (e.g., narrative, letter, and so on), and then plan a specific structure of that genre before writing, as shown in Figure 1. Such a process is modeled via the mode prior $p(\alpha)$ and the mode-outline distribution $p(\beta_{1:N}|\alpha)$. 2) It is capable of describing any language marginal distribution via the marginalization over latent variables. Also, it is convenient to analyze because of disentanglement: two kinds of uncertainties, topic-transition and generation of paragraphs are handled by two separated models $p(\beta_n|\beta_{1:n-1}, \alpha)$ and $p(x_n|\beta_n)$, respectively, instead of one entangled marginal language distribution $p(x_{1:N})$.

2.1.1 Assumptions

The pretrained distribution has three additional assumptions. Firstly, as mentioned, typical pretrained distributions for LLMs include the repetition mode $\hat{\alpha} \in A$ that only generates repeated outlines β^N ($\beta \in B$) (β^N represents a N-length outline that each topic within is β). This formally raises the following:

Assumption 1. *There exists a mode $\hat{\alpha} \in A$ called repetition mode such that $p(\beta_{n+1}|\beta_n, \hat{\alpha}) = 1(\beta_{n+1} = \beta_n)$ for all timesteps n. Other modes $\alpha \in A/\hat{\alpha}$ are called continuous modes, since the outline under them seems to shift gradually and continuously.*

Secondly, we have to ensure that different modes and topics are different to get rid of redundancy. That is, they should be distinguished in terms of distance measure of distribution:

Assumption 2. *For two different modes $\alpha, \alpha' \in A$ and an arbitrary context $x_{1:n}$, define:*

$$\mathrm{KL}_n\left(\alpha\|\alpha'\right) := \sum_x p(x|x_{1:n}, \alpha) \log \frac{p(x|x_{1:n}, \alpha)}{p(x|x_{1:n}, \alpha')} \tag{2}$$

We assume that $\mathrm{KL}_n\left(\alpha\|\alpha'\right)$ is at least bigger than a constant for all α and α': $\mathrm{KL}_n\left(\alpha\|\alpha'\right) \geq \log c_1 > 0$. Likewise, for two different topics $\beta, \beta' \in B$, define:

$$\mathrm{KL}(\beta\|\beta') := \sum_x p(x|\beta) \log \frac{p(x|\beta)}{p(x|\beta')} \tag{3}$$

We assume that $\mathrm{KL}(\beta\|\beta') \geq \log c_2 > 0$.

Thirdly, for convenience and without loss of plausibility, we assume that:

Assumption 3. *For each paragraph $x \in \Sigma^*$, its support from any topic $\beta \in B$ is bounded by two constants: $0 < c_3 \leq p(x|\beta) \leq c_4 < 1$.*

2.1.2 Topic Types

With Assumption 1, the likelihood of any repeated outline β^N under the repetition mode $\hat{\alpha}$ only depends on the topic itself:

$$p(\beta^N|\hat{\alpha}) = p(\beta_1 = \beta|\hat{\alpha}) := p(\beta|\hat{\alpha}) \tag{4}$$

where $p(\beta|\hat{\alpha})$ is the *repetition prior* measuring how often the topic β is chosen to repeat under mode $\hat{\alpha}$. Analogously, let $p(\beta)$ be the *topic prior* assessing the frequency of the topic β:

$$p(\beta) := \sum_{\alpha \in A} p(\beta|\alpha)p(\alpha) \tag{5}$$

According to the appearance, we can formally group topics $\beta \in B$ into three mutually exclusive sets, as shown in Figure 1:

1. Repeated set R. $\forall \beta \in R$, $p(\beta|\hat{\alpha}) > 0$. That is, each topic within has a chance to appear in the repetition mode in the pretrained distribution. By intuition, repeated topics account for a very small proportion of all topics in realistic data, i.e., $r_R = |R|/|B|$ is small.

2. Nonrepeated set C. $\forall \beta \in C$, $p(\beta|\hat{\alpha}) = 0$, $p(\beta) > 0$. In other words, this set contains topics that don't repeat but appear in the pretrained corpora.

3. Unseen set U. $\forall \beta \in U$, $p(\beta) = 0$. Topics in this set are never seen in the pretrained corpora.

2.2 Problem Formalization

According to the above setting, The ICG ability could be formalized as the following:

Hypothesis 1. *Given a language model p_{LM} trained on the pretrained distribution p and an in-context prompt $x_{1:N}$, where each sample $x_n \sim p(x|\hat{\beta})$, the in-context topic-repetition rate (ICTR), i.e., the probability that the language model generates a paragraph belonging to topic $\hat{\beta}$ when prompting with $x_{1:N}$, is somehow close to 1:*

$$p_{\mathrm{LM}}(\hat{\beta}|x_{1:N}) := p_{\mathrm{LM}}(\beta_{N+1} = \hat{\beta}|x_{1:N}) \approx 1 \quad (6)$$

Accordingly, the model ICG distribution $p_{\mathrm{LM}}(x|x_{1:N})$ is somehow close to the true topic-paragraph distribution $p(x|\hat{\beta})$:

$$p_{\mathrm{LM}}(x|x_{1:N}) \approx p(x|\beta) \quad (7)$$

In this formalization, ICG is all about the next topic prediction, where language models seem to implicitly choose the topic of in-context examples as the next topic. Our goal is to find support for this hypothesis from the perspective of both theory and empirical experiments.

3 Theoretical Support

Intuitively, the pretrained distribution itself ensures the ICG ability for repeated topics R. This can be explicitly formalized by the following theorem:

Theorem 1. *Given an in-context prompt $x_{1:N}$, where each sample $x_n \sim p(x|\hat{\beta})$ and $\hat{\beta} \in R$, the pretrained distribution has the following properties:*

1. *The data ICTR*[4] *converges to 1 in probability (corollary 4):*

$$\operatorname*{plim}_{N\to\infty} p(\hat{\beta}|x_{1:N}) = 1 \quad (8)$$

where we denote $p(\beta_{N+1} = \beta|x_{1:N}) := p(\beta|x_{1:N})$.

2. *For any candidate paragraph $x \in \Sigma^*$, the data ICG distribution $p(x|x_{1:N})$ converges to the true topic-paragraph $p(x|\hat{\beta})$ in probability (corollary 5):*

$$\operatorname*{plim}_{N\to\infty} p(x|x_{1:N}) = p(x|\hat{\beta}) \quad (9)$$

If the language model p_{LM} is expressive enough, it would gradually approach the pretrained distribution p with the increase of the number of training examples[5]. As a result, p_{LM} would exhibit the same properties p as shown in Theorem 1. Therefore, the ICG ability for repeated topics directly originates from the pretrained corpora.

Detailed theoretical results are provided in Appendix B. Here, we only present a proof sketch.

Proof Sketch. According to Section 2.1, $\forall x \in \Sigma^*$, the data ICG distribution is:

$$p(x|x_{1:N}) = \sum_{\beta \in B} p(\beta|x_{1:N}) p(x|\beta) \quad (10)$$

Therefore, the data ICG distribution $p(x|x_{1:N})$ is dominated by the topic predictive distribution $p(\beta|x_{1:N})$, i.e., ICTR. $p(\beta|x_{1:N})$ can be further decomposed as the mixture of modes:

$$p(\beta|x_{1:N}) = \sum_{\alpha \in A} p(\alpha|x_{1:N}) p(\beta|x_{1:N}, \alpha) \quad (11)$$

Firstly, we can prove that if $\hat{\beta} \in R$, then $\operatorname{plim}_{N\to\infty} p(\hat{\alpha}|x_{1:N}) = 1$ (corollary 1). Therefore, the mixture in formula (11) focuses on the component of repetition mode $p(\beta|x_{1:N}, \hat{\alpha})$ when N is large:

$$\begin{aligned} p(\beta|x_{1:N}) &\approx p(\beta|x_{1:N}, \hat{\alpha}) \\ &= \frac{p(\beta|\hat{\alpha}) \prod_{n=1}^{N} p(x_n|\beta)}{p(x_{1:N}|\hat{\alpha})} \end{aligned} \quad (12)$$

This form is exactly the Bayesian posterior distribution, which is in accord with previous works connecting ICL and Bayesian statistics (Xie et al., 2021; Wang et al., 2023b; Hahn and Goyal, 2023). Likewise, it turns out that the if $\hat{\beta} \in R$, then $\operatorname{plim}_{N\to\infty} p(\hat{\beta}|x_{1:N}, \hat{\alpha}) = 1$ (corollary 3), thus establishing the first point of theorem 1. Since the data ICG distribution $p(x|x_{1:N})$ depends on the topic predictive distribution $p(\beta|x_{1:N})$, we can

[4] Note that we use the prefix "data" to distinguish values from pretrained distribution and language model distribution.

[5] Previous works (Xie et al., 2021; Hahn and Goyal, 2023) typically take this as the null hypothesis.

prove the second point of theorem 1 analogously[6]. In Appendix B and C, we also present a detailed formula of the convergence, in which the convergence speed depends on the distinguishment of different modes and topics.

4 Experiments

Theory 1 can't ensure the ICG ability for nonrepeated and unseen topics $\beta \in C \cup U$ because they have a zero repetition prior $p(\beta|\hat{\alpha}) = 0$ and so the posterior under repetition mode is also zero: $p(\beta|x_{1:N}, \hat{\alpha}) = 0$. Then, the correct component $p(x|\beta)$ would never be selected under the repetition mode, preventing the ICG/ICL ability as a consequence.

However, this is contrary to the real case, where LLMs have the ICG-generalization ability: they are able to generalize ICG/ICL abilities to nonrepeated and unseen topics $\beta \in C \cup U$. We speculate that this might be caused by factors in both data and model side:

- Data side: The compositionality of natural language (Grandy, 1990) and the proportion of repeated topics r_R. Compositionality considers the meaning of a linguistic unit is a result of individual meanings of its sub-parts and how they are combined (Anderson, 2018). In this view, nonrepeated and unseen topics might share the same "sub-topics" with repeated topics. The bigger the proportion of repeated topics, the more frequently those sub-topics are shared. Therefore, LLMs may be able to recombine those sub-topics to recognize out-of-distribution topics in the repetition mode and exhibit generalization.

- Model side: The Transformer (Vaswani et al., 2017) structure. As the mainstream architecture of NLP, the success of Transformer is believed to originate from its strong generalization ability (Hendrycks et al., 2020; Jiang and Bansal, 2021).

We conduct rich experiments to verify the above arguments.

4.1 Synthetic Data

We conduct the experiments on synthetic data generated via the controllable pretrained distribution. As mentioned, the distribution has three components:

1. Mode prior $p(\alpha)$. We set the mode prior to be uniform: $p(\alpha) = 1/|A|$.

2. Mode-outline distribution $p(\beta_{1:N}|\alpha)$. For continuous modes $\alpha \in A/\hat{\alpha}$, we don't exactly care the outline under them, so we set $p(\beta_{1:N}|\alpha) = \prod_{n=1}^{N} p(\beta_n|\alpha)$ for convenience, where $p(\beta_n|\alpha)$ is a categorical distribution and its parameter is initialized from a Dirichlet distribution. The Dirichlet parameters are 0 for unseen topics (so that $p(\beta) = 0$ for $\beta \in U$) and 5 for others. We set the repetition prior to be uniform: $p(\beta|\hat{\alpha}) = 1/|R| = 1/|B|r_R$ $(\beta \in R)$.

3. Topic-paragraph distribution $p(x|\beta)$. In order to simulate the compositionality, each topic $\beta \in B$ is a tuple containing M subtopics $\rho^{1:M}$, where $\rho^m \in B_*(m \in [M])$ and $B = B_*^M$. Accordingly, the paragraph x also contains M sub-paragraphs $s^{1:M}$, where each sub-paragraph is generated individually:

$$p(x|\beta) = \prod_{m=1}^{M} p(s^m|\rho^m) \tag{13}$$

The composition arity M controls the data compositionality. Given a fix number of topics $|B|$, the number of subtopics $|B_*| = \sqrt[M]{|B|}$ decreases when composition arity M increases, and different topics are more likely to share sub-structures as a result. Here, each sub-paragraph distribution $p(s^m|\rho^m)$ is a Markov model whose initial probability vector $\boldsymbol{\pi}_{\rho^m}$ and transition matrix $\mathbf{A}_{\rho^m}$ are both sampled from $\mathrm{Dir}(\gamma\mathbf{1})$, where $\mathbf{1}$ is an one vector. γ actually controls the uncertainty of different topics, where a lower value is expected to raise the KL divergence between different topic-paragraph models, making them easier to be distinguished, as shown in Appendix D.

Data Parameter Settings

We set the number of modes $|A| = 32$, the number of topics $|B| = 531441$[7], where 95% of topics are unseen ($|U| = 504868$). We set the vocab size $|\Sigma| = 324$, the length of sub-paragraph $|s^m| = 3$, and the number of paragraphs in a document $N = 30$. Thus, each document contains

[6] Based on of theorem 1, for regular in-context learning scenario where each example in the prompt is a tuple (x_n, y_n) consisting of an input x_n and an output y_n, we can also obtain similar theoretical conclusions about the ICL ability. Details are shown in proposition 5 and corollary 6.

[7] We choose this number because its square, cube and fourth root are all integers.

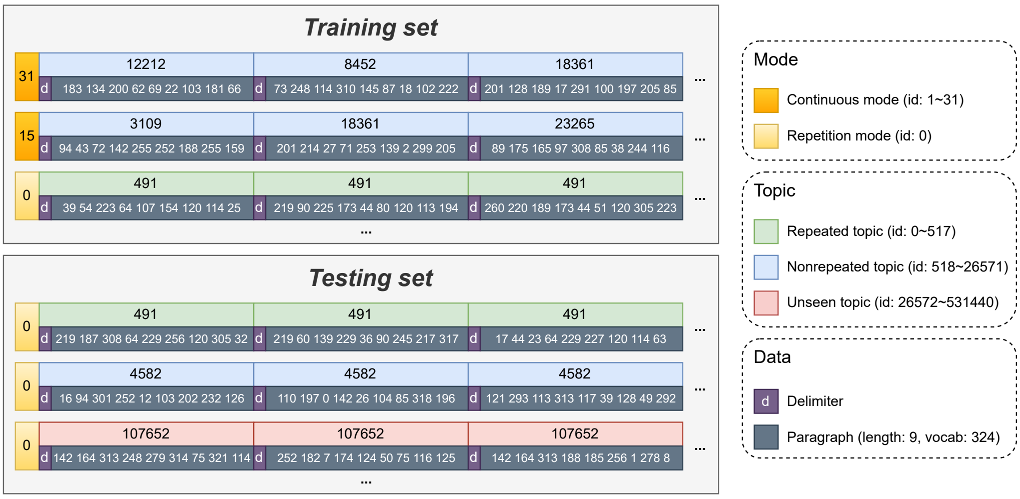

Training set
31
12212
d 183 134 200 62 69 22 103 181 66
8452
d 73 248 114 310 145 87 18 102 222
18361
d 201 128 189 17 291 100 197 205 85
...
15
3109
d 94 43 72 142 255 252 188 255 159
18361
d 201 214 27 71 253 139 2 299 205
23265
d 89 175 165 97 308 85 38 244 116
...
0
491
d 39 54 223 64 107 154 120 114 25
491
d 219 90 225 173 44 80 120 113 194
491
d 260 220 189 173 44 51 120 305 223
...
...
Testing set
0
491
d 219 187 308 64 229 256 120 305 32
491
d 219 60 139 229 36 90 245 217 317
491
d 17 44 23 64 229 227 120 114 63
...
0
4582
d 16 94 301 252 12 103 202 232 126
4582
d 110 197 0 142 26 104 85 318 196
4582
d 121 293 113 313 117 39 128 49 292
...
0
107652
d 142 164 313 248 279 314 75 321 114
107652
d 252 182 7 174 124 50 75 116 125
107652
d 142 164 313 188 185 256 1 278 8
...
...
Mode
Continuous mode (id: 1~31)
Repetition mode (id: 0)
Topic
Repeated topic (id: 0~517)
Nonrepeated topic (id: 518~26571)
Unseen topic (id: 26572~531440)
Data
d Delimiter
Paragraph (length: 9, vocab: 324)

Figure 3: Examples in the synthetic dataset, where we set $M = 3$, $r_R = 1/1024$ and $\gamma = 0.01$. We also visualize the latent variables mode α and outline $\beta_{1:N}$ for a better understanding.

Models	L	H	D	# params
X^{2}S	3	6	384	4M
XS	4	8	448	8M
S	5	8	448	9M
M	6	8	512	15M
L	9	12	768	48M
XL	12	16	1024	114M
X^{2}L	16	20	1280	234M

Table 1: Configurations of different models, where L is the number of layers, H is the number of attention heads, D is the hidden dimension. For parameter efficiency, we use grouped query attention (Ainslie et al., 2023) and set the number of key-value heads to be $H/2$.

$30(3M + 1)$ tokens. For other parameters of pretrained distribution including composition arity M, the ratio of repeated topics r_R, and topic uncertainty γ, we adjust their values to study the effects of data properties. In specific, we experiment with $M \in \{2, 3, 4\}$, $r_R \in \{2^{-d} | d = \{6, 7, \cdots, 13\}\}$, and $\gamma \in \{0.01, 0.02, \cdots, 0.05\}$.

For each configuration of the pretrained distribution, we generate 10M documents for training. Therefore, the number of tokens in the synthetic dataset ranges from 2.1B to 3.9B. Examples of the synthetic dataset are shown in Figure 3.

4.2 Models

We study the effect of model size, attention window size, and the number of attention heads of Transformer. Table 1 shows configurations of different experimental models, where the parameters scales from 4M to 237M. The models are based on the `Transformers` (Wolf et al., 2020) implementation of Mistral (Jiang et al., 2023a). We train each model for 1 epoch on one NVIDIA A100 (40GB).

4.3 Evaluation Metrics

We aim to evaluate the overall ICG performance and the ICG-generalization ability of models using ICTR. Firstly, we define topic-wise ICTR as the expectation of prompt-wise ICTR[8]:

$$\pi_N^\beta = \mathbb{E}_{p(x_{1:N}|\beta^N)} \left[p_{\text{LM}}(\beta | x_{1:N}) \right] \tag{14}$$

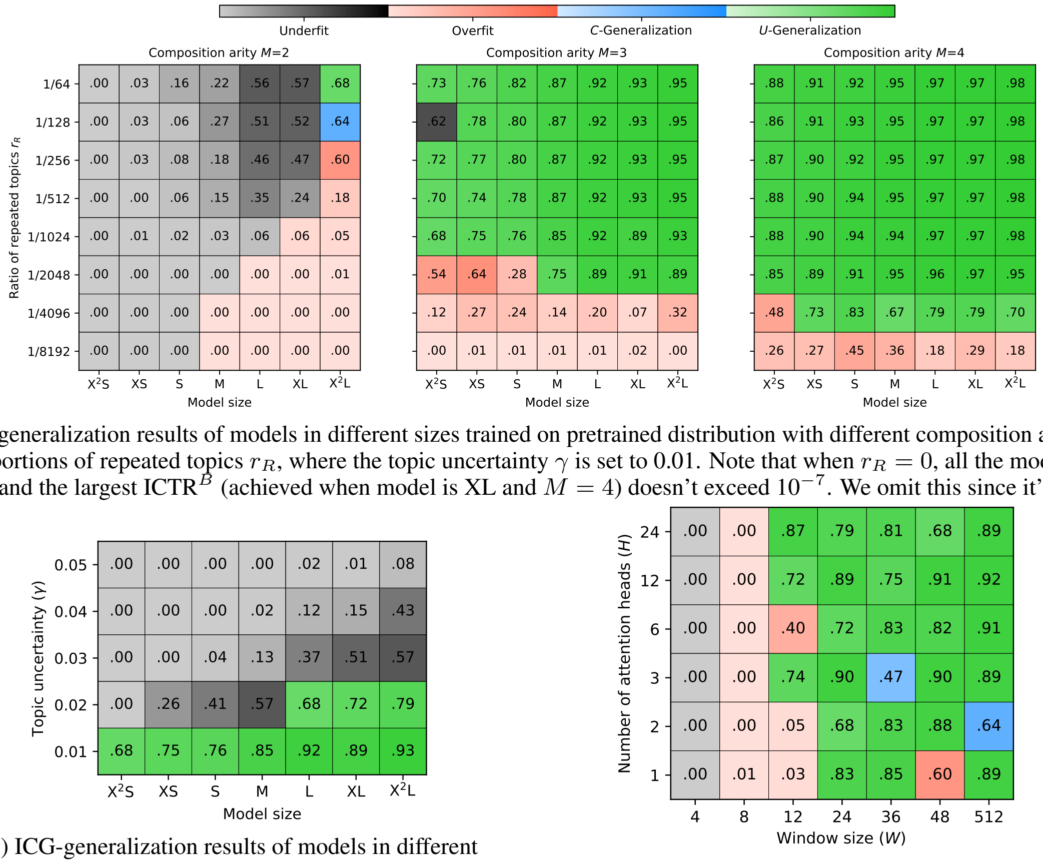
Then, we can obtain the average ICTR of different kinds of topics:

$$\text{ICTR}_N^B = \frac{1}{|B|} \sum_{\beta \in B} \pi_N^\beta, \quad \text{ICTR}_N^R = \frac{1}{|R|} \sum_{\beta \in R} \pi_N^\beta$$
$$\text{ICTR}_N^C = \frac{1}{|N|} \sum_{\beta \in C} \pi_N^\beta, \quad \text{ICTR}_N^U = \frac{1}{|U|} \sum_{\beta \in U} \pi_N^\beta \tag{15}$$

Here, ICTR_N^B measures the overall ICG ability, while ICTR_N^C and ICTR_N^U reflect the ICG-generalization ability, where higher values suggest better generalizations. In the experiments, since each pretrained document has 30 paragraphs, the trained model at most supports 29-shot in-context prompts. So by default, we reported $\text{ICTR}_{29}^{B/R/C/U}$, which is short of $\text{ICTR}^{B/R/C/U}$.

Following Liu et al. (2022), to get a compact and clear picture of the experimental results, we define four statuses of a trained model by thresholding the values of ICTRs as shown in Table 2.

[8] The computation of prompt and topic-wise ICTR is non-trivial, so we present it in Appendix E.



(a) ICG-generalization results of models in different sizes trained on pretrained distribution with different composition arities M and proportions of repeated topics r_R, where the topic uncertainty γ is set to 0.01. Note that when $r_R = 0$, all the models pose underfit and the largest ICTR^B (achieved when model is XL and $M = 4$) doesn't exceed 10^{-7}. We omit this since it's trival.

(b) ICG-generalization results of models in different sizes trained on pretrained distribution with different topic uncertainties γ, where we set $M = 3$ and $r_R = 1/1024$.

(c) ICG-generalization results of model L with different window sizes and numbers of attention heads, where we set $M = 3$, $r_R = 1/1024$, and $\gamma = 0.01$.

Figure 4: ICG-generalization results, where the color suggests the status of the corresponding model, and the number in the cell shows the corresponding ICTR^B_{29}.

ICTR^R	ICTR^C	ICTR^U	Status
< 0.65	-	-	Underfit
≥ 0.65	< 0.65	< 0.65	Overfit
≥ 0.65	≥ 0.65	< 0.65	C-Generalization
≥ 0.65	≥ 0.65	≥ 0.65	U-Generalization

Table 2: Definitions of different statuses.

4.4 Results & Discussions

Our experiments suggest the following arguments.

Data compositionality enables both ICG and ICG-generalization. Figure 4a shows the results of different composition arities. Clearly, we can see that data compositionality enables ICG and ICG-generalization, specifically: 1) As the composition arity M increases, the overall ICG performance consistently improves for models in any sizes trained on the pretrained distribution with different repeated topic proportions r_R. Notably, the improvement is especially significant when we increase M from 2 to 3. For example, for all r_R, the ICTR^B_{29} value nears 0 for many small models when $M = 2$, but is lifted to a considerable level when $M = 3$. 2) The models are easier to generalize on ICG when M is higher. When $M = 2$, most models are even hard to overfit on repeated topics, and only model X^2L can generalize ICG to both non-repeated and unseen topics only when $r_R = 1/64$. On the contrary, when $M = 3$ or $M = 4$, models in all sizes exhibit the ICG-generalization ability with much smaller r_R.

The model emerges the ICG-generalization as the proportion of repeated topics rises. As shown in Figure 4a, the model typically tends to overfit only on repeated topics when r_R is small, and then suddenly emerges the ICG-generalization ability when r_R hits the threshold. The threshold mainly corresponds to the data compositionality, where a higher composition arity M leads to a lower threshold and so makes the model easier to generalize. For example, for model X^2L, the generalization threshold of r_R is $1/64$ when $M = 2$, and

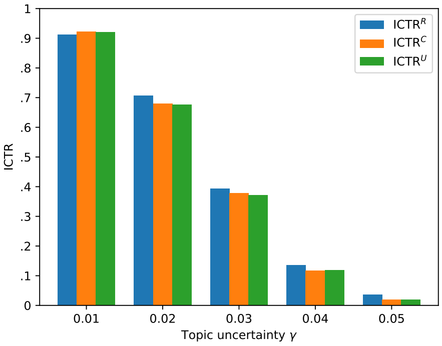


Figure 5: ICTR^*_{29} of different topics for model L trained on the pretrained distribution with different topic uncertainty γ, where the other parameters in the pretrained distribution are: $M = 3$, $r_R = 1/1024$.

decreases to $1/2048$ when $M = 3$. We speculate this is because the more compositionality of the data, the more likely that nonrepeated and unseen topics share sub-topics with repeated ones, therefore the less proportion of repeated topics is needed for generalization.

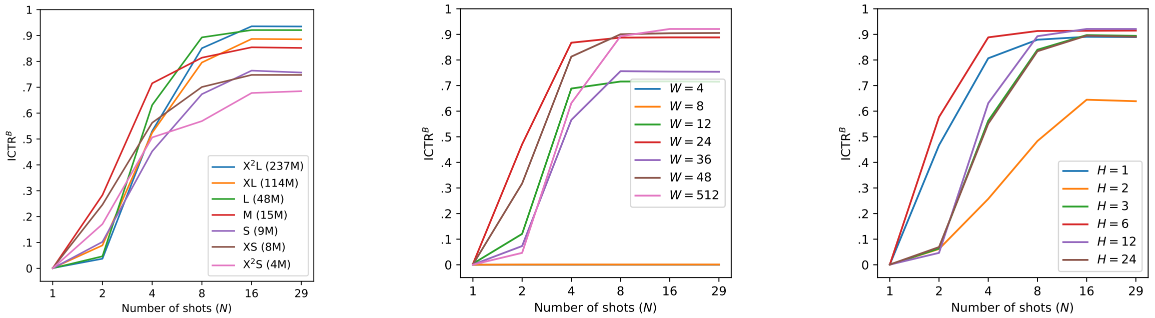
Topic uncertainty doesn't affect ICG-generalization. As shown in Figure 5, Topic uncertainty mainly affects the fitting difficulty of the data rather than the ICG-generalization difficulty: As the topic uncertainty γ increases, the ICTR_{29} of model L for all kinds of topics decreases consistently. Also, we don't observe apparent ICG performance gaps between those topics.

Larger models do better on ICG and ICG-generalization. Model size is considered to be a great factor impacting the ability of language models (Wei et al., 2022a). This is also verified in our experiments, which we find: 1) As shown in Figure 4a, obviously, larger models not only achieve better ICTR^B_{29}, but also require less repeated topics to generalize to nonrepeated and unseen topics. 2) As shown in Figure 4b, larger models are able to deal with topics with more uncertainties, i.e., bigger γ, where models larger than model M are capable of ICG-generalization when $\gamma = 0.02$ but smaller models pose underfit (Especially for model $\mathrm{X}^2\mathrm{S}$, whose ICTR^B_{29} is 0). 3) As shown in Figure 6a, in most cases, larger models achieve better ICTR^B given fewer demonstrations. However, curiously, this does not hold when the number of shots N is too small. For example, ICTR^B_2 of model $\mathrm{X}^2\mathrm{S}$, XS, S, and M are typically greater than that of model L, XL, and $\mathrm{X}^2\mathrm{L}$. We speculate this might be because when N is small, larger models are more cautious in identifying the repetition mode.

Big window size is necessary for ICG and ICG–generalization. Recently, Wang et al. (2023a) show that LLMs conduct ICL by collecting information of demonstrations in the prompt from previous label words. Specifically, the hidden states of previous label words are good summarizations of corresponding demonstrations. Thus, the model needs to attend to all those previous "anchors" to conduct ICL, which hints that a small window size might harm the ICL performance. For example, in the experimental results of Jiang et al. (2023b), we can find that the ICL performance of RWKV (Peng et al., 2023) series is generally inferior to that of classic Transformer structures. Our experiments also support this argument. As shown in Figure 4c and 6b, when the number of attention heads is fixed, a low window size would cause underfit. In most cases, when we increase the window size, the model is shifted to overfit and finally U-Generalization, the overall ICTR^B_{29} also rises at the same time. Note that there also exists the emergent phenomenon, where the model suddenly learns ICG and ICG-generalization when its window size hits a threshold.

Big number of attention heads is not necessary for ICG and ICG-generalization. Multi-head/group attention is always believed to be one of the most important components of state-of-the-art Transformer models. By intuition, different heads can potentially attend onto different parts of the text, making the model more expressive. However, our experiments show this mechanism is not very important for ICG and ICG-generalization. As shown in Figure 4c, reducing the number of attention heads H for XL model hardly change the model status. Also, we can find that when the model size is fixed, the model with the highest overall ICG performance does not necessarily have the most attention heads. We speculate this is because the attention pattern for ICG is relatively simple, so different heads are actually functional equivalent. This is consistent with Michel et al. (2019), which finds that the performance of many tasks including machine translation and natural language inference is insensitive to the number of attention heads.

Generalizations towards nonrepeated and unseen topics are almost the same. As shown in Figure 4, in most cases, no matter how pretrained distributions and models are configured, the mod-



(a) ICTR$_*^B$ of different model sizes.

(b) ICTR$_*^B$ of different window sizes for model L.

(c) ICTR$_*^B$ of different numbers of attention heads for Model L.

Figure 6: ICTR$_*^B$ of different model configurations, where we set $M = 3$, $\gamma = 0.01$, and $r_R = 1/1024$.

els generally end with either underfit, overfit, or U-Generalization, but hardly in the status of C-Generalization. This suggests that nonrepeated topics, though appear in the pretrained distribution, are not easier for models to generalize than unseen ones.

5 Related Works

As one of the most exciting emergent abilities (Wei et al., 2022a), the mechanism of ICL (Brown et al., 2020) has been widely studied in previous works. Empirically, researchers usually perturb the in-context prompt (e.g. label, order or choice of demonstrations) in order to observe effects of different aspects of both prompt and model architectures (Min et al., 2022; Yoo et al., 2022; Lu et al., 2022; Wei et al., 2023; Jiang et al., 2023b). A few works try to interpret ICL theoretically and conduct experiments on controlled synthetic datasets (Xie et al., 2021; Han et al., 2023; Garg et al., 2022; Jiang, 2023; Panwar et al., 2023; Bai et al., 2024). As a typical example, Garg et al. (2022) show that Transformers can learn some function classes (e.g. linear functions) in-context when explicitly training them to do so. However, these works only show appearance of ICL with synthetic data of repetition mode[9]. Comparing to them, Our synthetic data originates from a more plausible pretrained distribution. This distribution is much more plausible, since realistic data also has many continuous modes. So our results are more applicable for the realistic case. Furthermore, out experiments are to our knowledge the first to investigate the ICG ability (which beyond ICL) of language models.

6 Conclusions

This paper provides a systematic study of ICG ability of language models. Firstly, we propose a plausible latent variable pretrained distribution, formalizing ICG as a problem of next topic prediction. Then, we prove that the repetition nature of a few topics ensures the ICG ability on them theoretically. We also conduct rich experiments to study the effects of different factors of data and model architectures on ICG and ICG-generalization. We believe this paper is beneficial to a better understanding of the ICG ability, as well as large language models.

7 Limitations

The major limitation of this work is that we don't provide a theoretical support for ICG-generalization, while doing so is non-trivial. Now we can only speculate the ICG-generalization results from the smoothing effects of neural probability approximator (e.g. Transformer), where unseen inputs would have non-zero probabilities (Xie et al., 2017). Therefore, nonrepeated and unseen topics might have a non-zero repetition prior, making them possible to be chosen as the topic of the next paragraph. This phenomenon might be especially obvious when these topics are similar to repeated ones according to our experimental results. Further work on the theoretical understanding of ICG-generalization might take similarities between topics into account.

Acknowledgements

This work was supported by the National Key R&D Program of China (No. 2022ZD0160503) and the National Natural Science Foundation of China (No.62276264). This work was also sponsored by CCF-BaiChuan-Ebtech Foundation Model Fund.

[9] One might think that the synthetic data of Xie et al. (2021) is not in the repetition mode. However, the assumption 3 of their work actually implies that the repetition mode dominates the synthetic corpora. For example, if we take $\Delta = 0$, the synthetic examples are all in the repetition mode.

References

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. 2023. Gqa: Training generalized multi-query transformer models from multi-head checkpoints. *arXiv preprint arXiv:2305.13245*.

Catherine Anderson. 2018. *Essentials of linguistics*. McMaster University.

Kazuoki Azuma. 1967. Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal, Second Series*, 19(3):357–367.

Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. 2024. Transformers as statisticians: Provable in-context learning with in-context algorithm selection. *Advances in neural information processing systems*, 36.

Christopher M Bishop and Nasser M Nasrabadi. 2006. *Pattern recognition and machine learning*, volume 4. Springer.

Samuel R Bowman, Christopher D Manning, and Christopher Potts. 2015. Tree-structured composition in neural networks without tree-structured architectures. *arXiv preprint arXiv:1506.04834*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Shivam Garg, Dimitris Tsipras, Percy S Liang, and Gregory Valiant. 2022. What can transformers learn in-context? a case study of simple function classes. *Advances in Neural Information Processing Systems*, 35:30583–30598.

Richard E Grandy. 1990. Understanding and the principle of compositionality. *Philosophical Perspectives*, 4:557–572.

Michael Hahn and Navin Goyal. 2023. A theory of emergent in-context learning as implicit structure induction. *arXiv preprint arXiv:2303.07971*.

Chi Han, Ziqi Wang, Han Zhao, and Heng Ji. 2023. In-context learning of large language models explained as kernel regression. *arXiv preprint arXiv:2305.12766*, page 3.

Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic, Rishabh Krishnan, and Dawn Song. 2020. Pretrained transformers improve out-of-distribution robustness. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2744–2751.

Wassily Hoeffding. 1994. Probability inequalities for sums of bounded random variables. *The collected works of Wassily Hoeffding*, pages 409–426.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023a. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Hui Jiang. 2023. A latent space theory for emergent abilities in large language models. *arXiv preprint arXiv:2304.09960*.

Yichen Jiang and Mohit Bansal. 2021. Inducing transformer's compositional generalization ability via auxiliary sequence prediction tasks. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6253–6265.

Zhongtao Jiang, Yuanzhe Zhang, Cao Liu, Jun Zhao, and Kang Liu. 2023b. Generative calibration for in-context learning. *arXiv preprint arXiv:2310.10266*.

Jaap Jumelet and Willem Zuidema. 2023. Transparency at the source: Evaluating and interpreting language models with access to the true distribution. *arXiv preprint arXiv:2310.14840*.

Ziming Liu, Ouail Kitouni, Niklas Nolte, Eric J. Michaud, Max Tegmark, and Mike Williams. 2022. Towards understanding grokking: An effective theory of representation learning. *ArXiv*, abs/2205.10343.

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.

R Thomas McCoy, Robert Frank, and Tal Linzen. 2018. Revisiting the poverty of the stimulus: Hierarchical generalization without a hierarchical bias in recurrent neural networks. *arXiv preprint arXiv:1802.09091*.

Elliot Meyerson, Mark J Nelson, Herbie Bradley, Arash Moradi, Amy K Hoover, and Joel Lehman. 2023. Language model crossover: Variation through few-shot prompting. *arXiv preprint arXiv:2302.12170*.

Paul Michel, Omer Levy, and Graham Neubig. 2019. Are sixteen heads really better than one? *Advances in neural information processing systems*, 32.

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.

Madhur Panwar, Kabir Ahuja, and Navin Goyal. 2023. In-context learning through the bayesian prism. *arXiv preprint arXiv:2306.04891*.

Isabel Papadimitriou and Dan Jurafsky. 2023. Injecting structural hints: Using language models to study inductive biases in language learning. In *Findings* *of the Association for Computational Linguistics: EMNLP 2023*, pages 8402–8413.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, Huanqi Cao, Xin Cheng, Michael Chung, Matteo Grella, Kranthi Kiran GV, et al. 2023. Rwkv: Reinventing rnns for the transformer era. *arXiv preprint arXiv:2305.13048*.

Nir Ratner, Yoav Levine, Yonatan Belinkov, Ori Ram, Inbal Magar, Omri Abend, Ehud Karpas, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. Parallel context windows for large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6383–6402.

Frieda Rong. 2021. Extrapolating to unnatural language processing with gpt-3's in-context learning: The good, the bad, and the mysterious.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

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

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808.

Lean Wang, Lei Li, Damai Dai, Deli Chen, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. 2023a. Label words are anchors: An information flow perspective for understanding in-context learning. *arXiv preprint arXiv:2305.14160*.

Xinyi Wang, Wanrong Zhu, and William Yang Wang. 2023b. Large language models are implicitly topic models: Explaining and finding good demonstrations for in-context learning. *arXiv preprint arXiv:2301.11916*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022a. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

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

Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, et al. 2023. Larger language models do in-context learning differently. *arXiv preprint arXiv:2303.03846*.

Jennifer C White and Ryan Cotterell. 2021. Examining the inductive bias of neural language models with artificial languages. *arXiv preprint arXiv:2106.01044*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2021. An explanation of in-context learning as implicit bayesian inference. *arXiv preprint arXiv:2111.02080*.

Ziang Xie, Sida I Wang, Jiwei Li, Daniel Lévy, Aiming Nie, Dan Jurafsky, and Andrew Y Ng. 2017. Data noising as smoothing in neural network language models. *arXiv preprint arXiv:1703.02573*.

Kang Min Yoo, Junyeob Kim, Hyuhng Joon Kim, Hyunsoo Cho, Hwiyeol Jo, Sang-Woo Lee, Sang-goo Lee, and Taeuk Kim. 2022. Ground-truth labels matter: A deeper look into input-label demonstrations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2422–2437.

A Lemmas

To access the theoretical results in Appendix B, the following lemmas are useful.

Lemma 1. *For an arbitrary continuous mode $\alpha \in A/\hat{\alpha}$, let*

$$s_n = \sum_{i=1}^{n} \log \frac{p(x_i|x_{1:i-1}, \alpha)}{p(x_i|x_{1:i-1}, \hat{\alpha})} + \mathrm{KL}_{i-1}(\hat{\alpha}\|\alpha) \tag{16}$$

where

$$\mathrm{KL}_{i-1}(\hat{\alpha}\|\alpha) = \mathbb{E}_{p(x|x_{1:i-1},\hat{\alpha})}\left[\log \frac{p(x|x_{1:i-1}, \hat{\alpha})}{p(x|x_{1:i-1}, \alpha)}\right] \tag{17}$$

Then, s_n is a martingale about $x_{1:n}$.

Proof. This lemma is easy to prove according to the definition of martingale so we omit it. □

Lemma 2. *Let z_n $(n \in [N])$ be a series of positive random variables, $\forall t \geq 0$,*

$$\mathrm{P}\left(\sum_{n=1}^{N} z_n \geq t\right) \leq \sum_{n=1}^{N} \mathrm{P}\left(z_n \geq \frac{t}{N}\right) \tag{18}$$

Proof. Firstly, we have:

$$\begin{aligned}
\mathrm{P}\left(\sum_{n=1}^{N} z_n \geq t\right) &= \mathrm{P}\left(\sum_{n=1}^{N} z_n \geq t, z_N \geq \frac{t}{N}\right) \\
&+ \mathrm{P}\left(\sum_{n=1}^{N-1} z_n \geq \frac{N-1}{N}t, z_N \geq \frac{t}{N}\right) \\
&+ \mathrm{P}\left(\sum_{n=1}^{N} z_n \geq t, \sum_{n=1}^{N-1} z_n \geq \frac{N-1}{N}t\right) \\
&\leq \mathrm{P}\left(\sum_{n=1}^{N-1} z_n \leq \frac{N-1}{N}t, z_N \geq \frac{t}{N}\right) \\
&+ \mathrm{P}\left(\sum_{n=1}^{N-1} z_n \geq \frac{N-1}{N}t, z_N \leq \frac{t}{N}\right) \\
&+ 2\mathrm{P}\left(\sum_{n=1}^{N-1} z_n \geq \frac{N-1}{N}t, z_N \geq \frac{t}{N}\right) \\
&= \mathrm{P}\left(\sum_{n=1}^{N-1} z_n \geq \frac{N-1}{N}t\right) + \mathrm{P}\left(z_N \geq \frac{t}{N}\right)
\end{aligned} \tag{19}$$

Then, according to this recursion,

$$\begin{aligned}
&\mathrm{P}\left(\sum_{n=1}^{N} z_n \geq t\right) \\
&\leq \mathrm{P}\left(\sum_{n=1}^{N-1} z_n \geq \frac{N-1}{N}t\right) + \mathrm{P}\left(z_N \geq \frac{t}{N}\right) \\
&\leq \mathrm{P}\left(\sum_{n=1}^{N-2} z_n \geq \frac{N-2}{N}t\right) + \mathrm{P}\left(z_{N-1} \geq \frac{t}{N}\right) \\
&\qquad + \mathrm{P}\left(z_N \geq \frac{t}{N}\right) \\
&\dots \\
&\leq \sum_{n=1}^{N} \mathrm{P}\left(z_N \geq \frac{t}{N}\right)
\end{aligned} \tag{20}$$

So the result is proved. □

B Complete Theoretical Results

We analyze the data ICG distribution $p(x|x_{1:N})$, where $x_{1:N}$ are independent and identical distributed with PDF $p(x|\hat{\beta})$ and x is an arbitrary value in the domain of paragraph. As shown in Section 2.1, x depends on its topic:

$$p(x|x_{1:N}) = \sum_{\beta \in B} p(\beta|x_{1:N}) p(x|\beta) \tag{21}$$

where the topic predictive distribution $p(\beta|x_{1:N}) := p(\beta_{1:N} = \beta|x_{1:N})$ controls the strength of each topic for the $N+1$-th paragraph. We then study the property of this distribution.

Note that the topic predictive distribution can also analogously be factorized as the mixture of modes:

$$p(\beta|x_{1:N}) = \sum_{\alpha \in A} p(\alpha|x_{1:N}) p(\beta|x_{1:N}, \alpha) \tag{22}$$

where the mode posterior $p(\alpha|x_{1:N})$ controls the strength of each mode.

B.1 Property of mode posterior

Firstly, we study the property of the mode posterior $p(\alpha|x_{1:N})$.

Proposition 1. *Let:*

$$p_{\max}(\hat{\alpha}) = \max_{\alpha \in A/\hat{\alpha}} p(\alpha) \tag{23}$$

If t satisfies:

$$\frac{|A| p_{\max}(\hat{\alpha}) c_1^{-N}}{p(\hat{\alpha}) + |A| p_{\max}(\hat{\alpha}) c_1^{-N}} \leq t < 1 \tag{24}$$

and $\hat{\beta} \in R$, for repetition mode $\hat{\alpha}$, we have:

$$\begin{aligned}
&\mathrm{P}(1 - p(\hat{\alpha}|x_{1:N}) \geq t) \\
&\leq |A| e^{-\frac{\left(N \log c_1 + \log \frac{t p(\hat{\alpha})}{|A|(1-t) p_{\max}(\hat{\alpha})}\right)^2}{8N \log^2(c_4/c_3)}}
\end{aligned} \tag{25}$$

For any continuous mode $\alpha \in A/\hat{\alpha}$, we also have:

$$\begin{aligned}
&\mathrm{P}(p(\alpha|x_{1:N}) \geq t) \\
&\leq |A| e^{-\frac{\left(N \log c_1 + \log \frac{t p(\hat{\alpha})}{|A|(1-t) p_{\max}(\hat{\alpha})}\right)^2}{8N \log^2(c_4/c_3)}}
\end{aligned} \tag{26}$$

Proof. Firstly, note that the absolute martingale residual difference of s_n in formula (17) is bounded:

$$\begin{aligned}
&|s_n - s_{n-1}| \\
&= \left|\log \frac{p(x_n|x_{1:n-1}, \alpha)}{p(x_n|x_{1:n-1}, \hat{\alpha})} + \mathrm{KL}_{n-1}(\hat{\alpha}\|\alpha)\right| \\
&\leq \left|\log \frac{p(x_n|x_{1:n-1}, \alpha)}{p(x_n|x_{1:n-1}, \hat{\alpha})}\right| + |\mathrm{KL}_{n-1}(\hat{\alpha}\|\alpha)| \\
&\leq 2 \log \frac{c_4}{c_3}
\end{aligned} \tag{27}$$

Then, according to Azuma's inequity (Azuma, 1967), $\forall \epsilon > 0$, we have:

$$\begin{aligned} \mathrm{P}\left(\sum_{n=1}^{N} \log \frac{p(x_n|x_{1:n-1},\alpha)}{p(x_n|x_{1:n-1},\hat{\alpha})} + \mathrm{KL}_{n-1}(\hat{\alpha}\|\alpha) \geq \epsilon\right) \\ \leq e^{-\frac{\epsilon^2}{8N\log^2(c_4/c_3)}} \end{aligned} \tag{28}$$

Since $\mathrm{KL}_{i-1}(\hat{\alpha}\|\alpha) \geq \log c_1$, we can rewrite formula (28) as:

$$\begin{aligned} \mathrm{P}\left(\sum_{i=1}^{N} \log \frac{p(x_n|x_{1:n-1},\alpha)}{p(x_n|x_{1:n-1},\hat{\alpha})} \geq \epsilon - N\log c_1\right) \\ \leq e^{-\frac{\epsilon^2}{8N\log^2(c_4/c_3)}} \end{aligned} \tag{29}$$

Let $t = e^{\epsilon - N\log c_1} \in [c_1^{-N}, 1)$ and rearrange the formula, we can obtain the following inequality about the ratio of mode likelihoods:

$$\mathrm{P}\left(\frac{p(x_{1:N}|\alpha)}{p(x_{1:N}|\hat{\alpha})} \geq t\right) \leq e^{-\frac{(N\log c_1 + \log t)^2}{8N\log^2(c_4/c_3)}} \tag{30}$$

The ratio of mode likelihoods has a direct impact to the mode posterior. First, for repetiton mode $\hat{\alpha}$, $\forall 0 < t < 1$, we have:

$$\begin{aligned} &\mathrm{P}(1 - p(\hat{\alpha}|x_{1:N}) \geq t) = \mathrm{P}\left(\frac{1}{p(\hat{\alpha}|x_{1:N})} \geq \frac{1}{1-t}\right) \\ &= \mathrm{P}\left(\sum_{\alpha\in A/\hat{\alpha}} \frac{p(\alpha)}{p(\hat{\alpha})}\frac{p(x_{1:N}|\alpha)}{p(x_{1:N}|\hat{\alpha})} \geq \frac{t}{1-t}\right) \\ &\leq \sum_{\alpha\in A/\hat{\alpha}} \mathrm{P}\left(\frac{p(x_{1:N}|\alpha)}{p(x_{1:N}|\hat{\alpha})} \geq \frac{tp(\hat{\alpha})}{(|A|-1)(1-t)p(\alpha)}\right) \\ &\leq \sum_{\alpha\in A/\hat{\alpha}} \mathrm{P}\left(\frac{p(x_{1:N}|\alpha)}{p(x_{1:N}|\hat{\alpha})} \geq \frac{tp(\hat{\alpha})}{|A|(1-t)p_{\max}(\hat{\alpha})}\right) \end{aligned} \tag{31}$$

where we unpack the probability in the third line using lemma 2. Now, if

$$\begin{aligned} &\frac{tp(\hat{\alpha})}{|A|(1-t)p_{\max}(\hat{\alpha})} \geq c_1^{-N} \\ &\Rightarrow t \geq \frac{|A|p_{\max}(\hat{\alpha})c_1^{-N}}{p(\hat{\alpha}) + |A|p_{\max}(\hat{\alpha})c_1^{-N}} \end{aligned} \tag{32}$$

then we can apply formula (30):

$$\begin{aligned} &\mathrm{P}(1 - p(\hat{\alpha}|x_{1:N}) \geq t) \\ &\leq |A|e^{-\frac{\left(N\log c_1 + \log\frac{tp(\hat{\alpha})}{|A|(1-t)p_{\max}(\hat{\alpha})}\right)^2}{8N\log^2(c_4/c_3)}} \end{aligned} \tag{33}$$

As for continuous modes $\alpha \in A/\hat{\alpha}$, note that:

$$\begin{aligned} \mathrm{P}(p(\alpha|x_{1:N}) \geq t) &\leq \mathrm{P}\left(\sum_{\alpha\in A/\hat{\alpha}} p(\alpha|x_{1:N}) \geq t\right) \\ &= \mathrm{P}(1 - p(\hat{\alpha}|x_{1:N}) \geq t) \\ &\leq |A|e^{-\frac{\left(N\log c_1 + \log\frac{tp(\hat{\alpha})}{|A|(1-t)p_{\max}(\hat{\alpha})}\right)^2}{8N\log^2(c_4/c_3)}} \end{aligned} \tag{34}$$

□

Based on proposition 1, we can immediately obtain the following two corollaries:

Corollary 1. *If* $\hat{\beta} \in R$, $\mathrm{plim}_{N\to\infty}\, p(\hat{\alpha}|x_{1:N}) = 1$

Proof. To prove the results, we need to prove that, $\forall \epsilon > 0$, $\delta > 0$, there exists N_0 such that when $N \geq N_0$,

$$\mathrm{P}(1 - p(\hat{\alpha}|x_{1:N}) \geq \epsilon) < \delta \tag{35}$$

Firstly, note that when $\epsilon > 1$ or $\delta \geq 1$, the above formula holds trivially. When $0 < \epsilon \leq 1$, define:

$$\hat{N}(\epsilon) = \log_{c_1} \frac{|A|(1-\epsilon)p_{\max}(\hat{\alpha})}{tp(\hat{\alpha})} \tag{36}$$

If $N \geq \hat{N}(\epsilon)$, then

$$\epsilon \geq \frac{|A|p_{\max}(\hat{\alpha})c_1^{-N}}{p(\hat{\alpha}) + |A|p_{\max}(\hat{\alpha})c_1^{-N}} \tag{37}$$

Therefore, according to proposition 1, we have:

$$\mathrm{P}(1 - p(\hat{\alpha}|x_{1:N}) \geq \epsilon) \leq f(N) \tag{38}$$

where

$$f(N) = |A|e^{-\frac{\left(N\log c_1 + \log\frac{tp(\hat{\alpha})}{|A|(1-\epsilon)p_{\max}(\hat{\alpha})}\right)^2}{8N\log^2(c_4/c_3)}} \tag{39}$$

Since $f(N) \in (0, |A|^2]$ is a monotonic decreasing function in the domain of $[\hat{N}(\epsilon), \infty]$, $\forall \delta \in (0,1)$ there must exists $N' \geq \hat{N}(\epsilon)$ such that $\delta = f(N')$, or equivalently, $N' = f^{-1}(\delta)$. Let's set $N_0 = \lceil f^{-1}(\delta) \rceil + 1$. If $N \geq N_0$,

$$\mathrm{P}(1 - p(\hat{\alpha}|x_{1:N}) \geq \epsilon) \leq f(\lceil f^{-1}(\delta)\rceil + 1) < \delta \tag{40}$$

Therefore, the result is proven. □

Corollary 2. *If t satisfies:*

$$\frac{|A|^{5/2}p_{\max}(\hat{\alpha})c_1^{-N}}{p(\hat{\alpha}) + |A|p_{\max}(\hat{\alpha})c_1^{-N}} \leq t < 1 \tag{41}$$

and $\hat{\beta} \in R$, we have:

$$
\begin{aligned}
&\mathrm{P}(|p(\beta|x_{1:N}) - p(\beta|x_{1:N}, \hat{\alpha})| \geq t) \\
&\leq |A|^2 e^{-\frac{\left(N \log c_1 + \log \frac{tp(\hat{\alpha})}{|A|(|A|^{\frac{3}{2}} - t)p_{\max}(\hat{\alpha})}\right)^2}{8N \log^2(c_4/c_3)}}
\end{aligned} \tag{42}
$$

Proof. Let $\mathbf{p}_N^\alpha \in \Delta^{|A|}$ be the topic posterior vector:

$$
\mathbf{p}_N^\alpha = \begin{bmatrix} \cdots \\ p(\alpha|x_{1:N}) \\ \cdots \end{bmatrix} \in \Delta^{|A|} \tag{43}
$$

and $\boldsymbol{\delta}^{\hat{\alpha}}$ be the one-hot vector peaking at $\hat{\alpha}$. $\forall 0 < t < 1$, Obviously:

$$
\begin{aligned}
&\mathrm{P}\left(\|\mathbf{p}_N^\alpha - \boldsymbol{\delta}^{\hat{\alpha}}\|_2 \geq t\right) \\
&\leq \mathrm{P}\left(\sum_{\alpha \in A/\hat{\alpha}} p(\alpha|x_{1:N}) + 1 - p(\hat{\alpha}|x_{1:N}) \geq t\right) \\
&\leq \sum_{\alpha \in A/\hat{\alpha}} \mathrm{P}\left(p(\alpha|x_{1:N}) \geq \frac{t}{|A|}\right) \\
&\qquad + \mathrm{P}\left(1 - p(\hat{\alpha}|x_{1:N}) \geq \frac{t}{|A|}\right)
\end{aligned} \tag{44}
$$

If

$$
\begin{aligned}
\frac{t}{|A|} &\geq \frac{|A| p_{\max}(\hat{\alpha}) c_1^{-N}}{p(\hat{\alpha}) + |A| p_{\max}(\hat{\alpha}) c_1^{-N}} \\
\Rightarrow t &\geq \frac{|A|^2 p_{\max}(\hat{\alpha}) c_1^{-N}}{p(\hat{\alpha}) + |A| p_{\max}(\hat{\alpha}) c_1^{-N}}
\end{aligned} \tag{45}
$$

then we can apply formula (25) and (26) to get the following:

$$
\begin{aligned}
&\mathrm{P}\left(\|\mathbf{p}_N^\alpha - \boldsymbol{\delta}^{\hat{\alpha}}\|_2 \geq t\right) \\
&\leq |A|^2 e^{-\frac{\left(N \log c_1 + \log \frac{tp(\hat{\alpha})}{|A|(|A| - t)p_{\max}(\hat{\alpha})}\right)^2}{8N \log^2(c_4/c_3)}}
\end{aligned} \tag{46}
$$

Now, denote:

$$
\mathbf{p}_{\cdot|N,\alpha}^\beta = \begin{bmatrix} \cdots \\ p(\beta|x_{1:N}, \alpha) \\ \cdots \end{bmatrix} \in [0,1]^{|A|} \tag{47}
$$

Then, $\forall 0 < t < 1$, we have:

$$
\begin{aligned}
&\mathrm{P}(|p(\beta|x_{1:N}) - p(\beta|x_{1:N}, \hat{\alpha})| \geq t) \\
&= \mathrm{P}\left(\left|\left(\mathbf{p}_N^\alpha - \boldsymbol{\delta}^{\hat{\alpha}}\right)^T \mathbf{p}_{\cdot|N,\alpha}^\beta\right| \geq t\right) \\
&\leq \mathrm{P}\left(\left\|\mathbf{p}_N^\alpha - \boldsymbol{\delta}^{\hat{\alpha}}\right\|_2 \left\|\mathbf{p}_{\cdot|N,\alpha}^\beta\right\|_2 \geq t\right) \\
&\leq \mathrm{P}\left(\left|\mathbf{p}_N^\alpha - \boldsymbol{\delta}^{\hat{\alpha}}\right| \geq \frac{t}{\sqrt{|A|}}\right)
\end{aligned} \tag{48}
$$

If $t \geq \frac{|A|^{5/2} p_{\max}(\hat{\alpha}) c_1^{-N}}{p(\hat{\alpha}) + |A| p_{\max}(\hat{\alpha}) c_1^{-N}}$, we can then apply formula (46) to obtain the result. □

B.2 Property of topic posterior under repetition mode

Secondly, we study the property of the topic posterior under the repetition mode $p(\beta|x_{1:N}, \hat{\alpha})$.

Proposition 2. *Let*

$$
p_{\max}(\hat{\beta}) = \max_{\beta \in B/\hat{\beta}} p(\beta|\hat{\alpha}) \tag{49}
$$

If t satisfies:

$$
\frac{|B| p_{\max}(\hat{\beta}|\hat{\alpha}) c_2{}^{-N}}{p(\hat{\beta}|\hat{\alpha}) + |B| p_{\max}(\hat{\beta}|\hat{\alpha}) c_2{}^{-N}} \leq t < 1 \tag{50}
$$

Then, for the ground-truth topic $\hat{\beta}$, if $\hat{\beta} \in R$, we have:

$$
\begin{aligned}
&\mathrm{P}(1 - p(\hat{\beta}|x_{1:N}, \hat{\alpha}) \geq t) \leq \sum_{\beta \in B/\hat{\beta}} \\
&\leq |B| e^{-\frac{2\left(N \log c_2 + \log \frac{tp(\hat{\beta}|\hat{\alpha})}{|B|(1-t)p_{\max}(\hat{\beta}|\hat{\alpha})}\right)^2}{N \log^2(c_4/c_3)}}
\end{aligned} \tag{51}
$$

For any other topic $\beta \in R/\hat{\beta}$, we also have:

$$
\begin{aligned}
&\mathrm{P}(p(\beta|x_{1:N}, \hat{\alpha}) \geq t) \\
&\leq |B| e^{-\frac{2\left(N \log c_2 + \log \frac{tp(\hat{\beta}|\hat{\alpha})}{|B|(1-t)p_{\max}(\hat{\beta}|\hat{\alpha})}\right)^2}{N \log^2(c_4/c_3)}}
\end{aligned} \tag{52}
$$

Proof. For any topic $\beta \in B/\hat{\beta}$, let

$$
s_n = \sum_{i=1}^n \log \frac{p(x_i|\beta)}{p(x_i|\hat{\beta})} \tag{53}
$$

Since each demonstration x_n is independently sampled from $p(x|\hat{\beta})$, all the addends in the above formula are independent. Also, note that:

$$
\begin{aligned}
\mathbb{E}[s_n] = \sum_{i=1}^n \mathbb{E}\left[\log \frac{p(x_i|\beta)}{p(x_i|\hat{\beta})}\right] &= n\mathrm{KL}(\hat{\beta}\|\beta) \\
&\geq n \log c_2 \\
\left|\log \frac{p(x_i|\beta)}{p(x_i|\hat{\beta})}\right| \leq \log \frac{c_4}{c_3}&
\end{aligned} \tag{54}
$$

Then, according to Hoeffding's inequity (Hoeffding, 1994), $\forall \epsilon > 0$,

$$\begin{aligned}
&\mathrm{P}\left(\sum_{i=1}^{N} \log \frac{p(x_i|\beta)}{p(x_i|\hat{\beta})} \geq \epsilon - N \log c_2\right) \\
&\leq \mathrm{P}\left(\sum_{i=1}^{N} \log \frac{p(x_i|\beta)}{p(x_i|\hat{\beta})} \geq \epsilon - N\mathrm{KL}(\hat{\beta}\|\beta)\right) \\
&= \mathrm{P}\left(\prod_{i=1}^{N} \frac{p(x_i|\beta)}{p(x_i|\hat{\beta})} \geq e^{\epsilon - N\mathrm{KL}(\hat{\beta}\|\beta)}\right) \\
&\qquad\qquad \leq e^{-\frac{2\epsilon^2}{N \log^2(c_4/c_3)}}
\end{aligned} \tag{55}$$

Let $t = e^{\epsilon - N \log c_2} \geq {c_2}^{-N}$, we have:

$$\mathrm{P}\left(\prod_{n=1}^{N} \frac{p(x_n|\beta)}{p(x_n|\hat{\beta})} \geq t\right) \leq e^{-\frac{2(N \log c_2 + \log t)^2}{N \log^2(c_4/c_3)}} \tag{56}$$

The rest of proof of is very similar to that of proposition 1, $\forall t \geq \frac{|B| p_{\max}(\hat{\beta}|\hat{\alpha}) {c_2}^{-N}}{p(\hat{\beta}|\hat{\alpha}) + |B| p_{\max}(\hat{\beta}|\hat{\alpha}) {c_2}^{-N}}$,

$$\begin{aligned}
&\mathrm{P}(1 - p(\hat{\beta}|x_{1:N}, \hat{\alpha}) \geq t) \leq \sum_{\beta \in B/\hat{\beta}} \\
&\quad \mathrm{P}\left(\prod_{n=1}^{N} \frac{p(x_n|\beta)}{p(x_n|\hat{\beta})} \geq \frac{tp(\hat{\beta}|\hat{\alpha})}{|B|(1-t) p_{\max}(\hat{\beta}|\hat{\alpha})}\right) \\
&\quad \leq |B| e^{-\frac{2\left(N \log c_2 + \log \frac{tp(\hat{\beta}|\hat{\alpha})}{|B|(1-t) p_{\max}(\hat{\beta}|\hat{\alpha})}\right)^2}{N \log^2(c_4/c_3)}}
\end{aligned} \tag{57}$$

And $\forall \beta \in R/\hat{\beta}$,

$$\begin{aligned}
&\mathrm{P}(p(\beta|x_{1:N}, \hat{\alpha}) \geq t) \\
&\quad \leq |B| e^{-\frac{2\left(N \log c_2 + \log \frac{tp(\hat{\beta}|\hat{\alpha})}{|B|(1-t) p_{\max}(\hat{\beta}|\hat{\alpha})}\right)^2}{N \log^2(c_4/c_3)}}
\end{aligned} \tag{58}$$

□

Likewise, we can also obtain the following corollary:

Corollary 3. *If* $\hat{\beta} \in R$, $\mathrm{plim}_{N \to \infty}\, p(\hat{\beta}|x_{1:N}, \hat{\alpha}) = 1$.

Proof. The proof is identical to the proof of corollary 4 so we omit it. □

B.3 Property of topic predictive distribution

Based on the above results, we are able to investigate the property of the topic predictive distribution $p(\beta|x_{1:N})$.

Proposition 3. *If t satisfies:*

$$1 > t \geq \max \left\{ \begin{array}{c} \frac{2|A|^{5/2} p_{\max}(\hat{\alpha}) c_1^{-N}}{p(\hat{\alpha}) + |A| p_{\max}(\hat{\alpha}) c_1^{-N}} \\ \frac{2|B| p_{\max}(\hat{\beta}|\hat{\alpha}) {c_2}^{-N}}{p(\hat{\beta}|\hat{\alpha}) + |B| p_{\max}(\hat{\beta}|\hat{\alpha}) {c_2}^{-N}} \end{array} \right. \tag{59}$$

Then, for the ground-truth topic $\hat{\beta}$, if $\hat{\beta} \in R$, we have:

$$\begin{aligned}
&\mathrm{P}(1 - p(\hat{\beta}|x_{1:N}) \geq t) \\
&\leq |A|^2 e^{-\frac{\left(N \log c_1 + \log \frac{tp(\hat{\alpha})}{|A|(2|A|^{\frac{3}{2}} - t) p_{\max}(\hat{\alpha})}\right)^2}{8N \log^2(c_4/c_3)}} \\
&+ |B| e^{-\frac{2\left(N \log c_2 + \log \frac{tp(\hat{\beta}|\hat{\alpha})}{|B|(2-t) p_{\max}(\hat{\beta}|\hat{\alpha})}\right)^2}{N \log^2(c_4/c_3)}}
\end{aligned} \tag{60}$$

For other topics $\beta \in B/\hat{\beta}$, we also have:

$$\begin{aligned}
&\mathrm{P}(p(\beta|x_{1:N}) \geq t) \\
&\leq |A|^2 e^{-\frac{\left(N \log c_1 + \log \frac{tp(\hat{\alpha})}{|A|(2|A|^{\frac{3}{2}} - t) p_{\max}(\hat{\alpha})}\right)^2}{8N \log^2(c_4/c_3)}} \\
&+ |B| e^{-\frac{2\left(N \log c_2 + \log \frac{tp(\hat{\beta}|\hat{\alpha})}{|B|(2-t) p_{\max}(\hat{\beta}|\hat{\alpha})}\right)^2}{N \log^2(c_4/c_3)}}
\end{aligned} \tag{61}$$

Proof. For the ground-truth topic $\hat{\beta}$ and any $0 < t < 1$, we have:

$$\begin{aligned}
&\mathrm{P}(1 - p(\hat{\beta}|x_{1:N}) \geq t) \\
&= \mathrm{P}(p(\hat{\beta}|x_{1:N}, \hat{\alpha}) - p(\hat{\beta}|x_{1:N}) + \\
&\qquad 1 - p(\hat{\beta}|x_{1:N}, \hat{\alpha}) \geq t) \\
&\leq \mathrm{P}(|p(\hat{\beta}|x_{1:N}, \hat{\alpha}) - p(\hat{\beta}|x_{1:N})| + \\
&\qquad 1 - p(\hat{\beta}|x_{1:N}, \hat{\alpha}) \geq t) \\
&\leq \mathrm{P}\left(|p(\hat{\beta}|x_{1:N}, \hat{\alpha}) - p(\hat{\beta}|x_{1:N})| \geq \frac{t}{2}\right) \\
&\qquad \mathrm{P}\left(1 - p(\hat{\beta}|x_{1:N}, \hat{\alpha}) \geq \frac{t}{2}\right)
\end{aligned} \tag{62}$$

Therefore, if

$$1 > t \geq \max \left\{ \begin{array}{c} \frac{2|A|^{5/2} p_{\max}(\hat{\alpha}) c_1^{-N}}{p(\hat{\alpha}) + |A| p_{\max}(\hat{\alpha}) c_1^{-N}} \\ \frac{2|B| p_{\max}(\hat{\beta}|\hat{\alpha}) {c_2}^{-N}}{p(\hat{\beta}|\hat{\alpha}) + |B| p_{\max}(\hat{\beta}|\hat{\alpha}) {c_2}^{-N}} \end{array} \right. \tag{63}$$

we can then apply corollary 2 and proposition 2 to prove formula (60). Meanwhile, for other topics $\beta \in B/\hat{\beta}$, we have:

$$\begin{aligned}
&\mathrm{P}(p(\beta|x_{1:N}) \geq t) \leq \mathrm{P}\left(\sum_{\beta \in B/\hat{\beta}} p(\beta|x_{1:N}) \geq t)\right) \\
&\qquad\qquad = \mathrm{P}(1 - p(\hat{\beta}|x_{1:N}) \geq t))
\end{aligned} \tag{64}$$

Then, if t satisfies formula (63), we can obtain formula (61). □

The property of the topic predictive distribution can be summarized more compactly via the following corollary:

Corollary 4. *If $\hat{\beta} \in R$, $\operatorname{plim}_{N\to\infty} p(\hat{\beta}|x_{1:N}) = 1$.*

Proof. The proof is identical to the proof of corollary 4 so we omit it. □

B.4 Property of in-context generative distribution

According the property of the topic predictive distribution, we can finally study the property of the in-context generative distribution.

Proposition 4. *If t satisfies:*

$$1 > t \geq \max \begin{cases} \frac{2c_4|A|^{5/2}|B|^{3/2}p_{\max}(\hat{\alpha})c_1^{-N}}{p(\hat{\alpha})+|A|p_{\max}(\hat{\alpha})c_1^{-N}} \\ \frac{2c_4|B|^{3/2}p_{\max}(\hat{\beta}|\hat{\alpha})c_2{}^{-N}}{p(\hat{\beta}|\hat{\alpha})+|B|p_{\max}(\hat{\beta}|\hat{\alpha})c_2{}^{-N}} \end{cases} \tag{65}$$

and $\hat{\beta} \in R$, for any candidate paragraph $x \in \Sigma^*$, we have:

$$\begin{aligned} &\mathrm{P}(|p(x|x_{1:N}) - p(x|\hat{\beta})| \geq t) \\ &\leq |A|^2|B|e^{-\frac{\left(N\log c_1 + \log \frac{tp(\hat{\alpha})}{|A|(2|A|^{\frac{3}{2}}|B|^{\frac{3}{2}}c_4 - t)p_{\max}(\hat{\alpha})}\right)^2}{8N\log^2(c_4/c_3)}} \\ &+ |B|^2e^{-\frac{2\left(N\log c_2 + \log \frac{tp(\hat{\beta}|\hat{\alpha})}{|B|(2|B|^{\frac{3}{2}}c_4 - t)p_{\max}(\hat{\beta}|\hat{\alpha})}\right)^2}{N\log^2(c_4/c_3)}} \end{aligned} \tag{66}$$

Proof. Let $\mathbf{p}_N^{\beta} \in \Delta^{|B|}$ be the topic predictive vector:

$$\mathbf{p}_N^{\beta} = \begin{bmatrix} \cdots \\ p(\beta|x_{1:N}) \\ \cdots \end{bmatrix} \in \Delta^{|B|} \tag{67}$$

and $\boldsymbol{\delta}^{\hat{\beta}}$ be the one-hot vector peaking at $\hat{\beta}$. For all $0 < t < 1$, we have:

$$\begin{aligned} &\mathrm{P}\left(\|\mathbf{p}_N^{\beta} - \boldsymbol{\delta}^{\hat{\beta}}\|_2 \geq t\right) \\ &\leq \mathrm{P}\left(\sum_{\beta\in B/\hat{\beta}} p(\beta|x_{1:N}) + 1 - p(\hat{\beta}|x_{1:N}) \geq t\right) \\ &\leq \sum_{\beta\in B/\hat{\beta}} \mathrm{P}\left(p(\beta|x_{1:N}) \geq \frac{t}{|B|}\right) \\ &\qquad + \mathrm{P}\left(1 - p(\hat{\beta}|x_{1:N}) \geq \frac{t}{|B|}\right) \end{aligned} \tag{68}$$

If

$$\begin{aligned} \frac{t}{|B|} &\geq \max \begin{cases} \frac{2|A|^{5/2}p_{\max}(\hat{\alpha})c_1^{-N}}{p(\hat{\alpha})+|A|p_{\max}(\hat{\alpha})c_1^{-N}} \\ \frac{2|B|p_{\max}(\hat{\beta}|\hat{\alpha})c_2{}^{-N}}{p(\hat{\beta}|\hat{\alpha})+|B|p_{\max}(\hat{\beta}|\hat{\alpha})c_2{}^{-N}} \end{cases} \\ \Rightarrow t &\geq \max \begin{cases} \frac{2|A|^{5/2}|B|p_{\max}(\hat{\alpha})c_1^{-N}}{p(\hat{\alpha})+|A|p_{\max}(\hat{\alpha})c_1^{-N}} \\ \frac{2|B|^2p_{\max}(\hat{\beta}|\hat{\alpha})c_2{}^{-N}}{p(\hat{\beta}|\hat{\alpha})+|B|p_{\max}(\hat{\beta}|\hat{\alpha})c_2{}^{-N}} \end{cases} \end{aligned} \tag{69}$$

Then we can apply results from proposition 3 to get the following:

$$\begin{aligned} &\mathrm{P}\left(\|\mathbf{p}_N^{\beta} - \boldsymbol{\delta}^{\hat{\beta}}\|_2 \geq t\right) \\ &\leq |A|^2|B|e^{-\frac{\left(N\log c_1 + \log \frac{tp(\hat{\alpha})}{|A|(2|A|^{\frac{3}{2}}|B| - t)p_{\max}(\hat{\alpha})}\right)^2}{8N\log^2(c_4/c_3)}} \\ &+ |B|^2e^{-\frac{2\left(N\log c_2 + \log \frac{tp(\hat{\beta}|\hat{\alpha})}{|B|(2|B| - t)p_{\max}(\hat{\beta}|\hat{\alpha})}\right)^2}{N\log^2(c_4/c_3)}} \end{aligned} \tag{70}$$

Now, denote:

$$\mathbf{p}_{\cdot|\beta}^{x} = \begin{bmatrix} \cdots \\ p(x|\beta) \\ \cdots \end{bmatrix} \in [c_3, c_4]^{|B|} \tag{71}$$

Therefore, For all $0 < t < 1$,

$$\begin{aligned} &\mathrm{P}(|p(x|x_{1:N}) - p(x|\hat{\beta})| \geq t) \\ &= \mathrm{P}\left(\left|\left(\mathbf{p}_N^{\beta} - \boldsymbol{\delta}^{\hat{\beta}}\right)^T \mathbf{p}_{\cdot|\beta}^{x}\right| \geq t\right) \\ &\leq \mathrm{P}\left(\left\|\mathbf{p}_N^{\beta} - \boldsymbol{\delta}^{\hat{\beta}}\right\|_2 \left\|\mathbf{p}_{\cdot|\beta}^{x}\right\|_2 \geq t\right) \\ &\leq \mathrm{P}\left(\left\|\mathbf{p}_N^{\beta} - \boldsymbol{\delta}^{\hat{\beta}}\right\|_2 \geq \frac{t}{\sqrt{|B|}c_4}\right) \end{aligned} \tag{72}$$

Therefore, if t satisfies formula (65), we can then apply formula (66) to prove the result. □

Proposition 4 directly supports the following corollary:

Corollary 5. *If $\hat{\beta} \in R$, $\operatorname{plim}_{N\to\infty} p(x|x_{1:N}) = p(x|\hat{\beta})$.*

Proof. The proof is identical to the proof of corollary 4 so we omit it. □

B.5 Property of in-context predictive distribution

We can generalize the property of ICG distribution to the in-context predictive distribution as well, which forms the theoretical foundation of ICL.

Proposition 5. *If t satisfies:*

$$1 > t \geq \max \begin{cases} \frac{4c_3^2 c_4^2 |A|^{5/2} |B|^{3/2} p_{\max}(\hat{\alpha}) c_1^{-N}}{p(\hat{\alpha}) + |A| p_{\max}(\hat{\alpha}) c_1^{-N}} \\ \frac{4c_3^2 c_4^2 |B|^{3/2} p_{\max}(\hat{\beta}|\hat{\alpha}) c_2{}^{-N}}{p(\hat{\beta}|\hat{\alpha}) + |B| p_{\max}(\hat{\beta}|\hat{\alpha}) c_2{}^{-N}} \end{cases} \tag{73}$$

and $\hat{\beta} \in R$, we have

$$\begin{aligned} &\mathrm{P}\left(\left|p(y|(x,y)_{1:N}, x) - p(y|x,\hat{\beta})\right| \geq t\right) \\ &\leq |A|^2 |B| e^{-\frac{\left(N \log c_1 + \log \frac{tp(\hat{\alpha})}{|A|(4|A|^{\frac{3}{2}} |B|^{\frac{3}{2}} c_3^2 c_4^2 - t) p_{\max}(\hat{\alpha})}\right)^2}{8N \log^2 (c_4/c_3)}} \\ &+ |B|^2 e^{-\frac{2\left(N \log c_2 + \log \frac{tp(\hat{\beta}|\hat{\alpha})}{|B|(4|B|^{\frac{3}{2}} c_3^2 c_4^2 - t) p_{\max}(\hat{\beta}|\hat{\alpha})}\right)^2}{N \log^2 (c_4/c_3)}} \end{aligned} \tag{74}$$

Proof. $\forall 0 < t < 1$, we have

$$\begin{aligned} &\mathrm{P}\left(\left|p(y|(x,y)_{1:N}, x) - p(y|x,\hat{\beta})\right| \geq t\right) \\ &= \mathrm{P}\left(\left|\frac{p(x,y|(x,y)_{1:N})}{p(x|(x,y)_{1:N})} - \frac{p(x,y|\hat{\beta})}{p(x|\hat{\beta})}\right| \geq t\right) \\ &= \mathrm{P}\left(\left|\frac{p(x|\hat{\beta})p(x,y|(x,y)_{1:N})}{p(x|(x,y)_{1:N})p(x|\hat{\beta})}\right.\right. \\ &\qquad \left.\left.\frac{-p(x,y|\hat{\beta})p(x|(x,y)_{1:N})}{}\right| \geq t\right) \\ &\leq \mathrm{P}\left(\left|p(x|\hat{\beta})p(x,y|(x,y)_{1:N})\right.\right. \\ &\qquad \left.\left.-p(x,y|\hat{\beta})p(x|(x,y)_{1:N})\right| \geq \frac{t}{c_3^2}\right) \\ &= \mathrm{P}\left(\left|p(x|\hat{\beta})\left(p(x,y|(x,y)_{1:N}) - p(x,y|\hat{\beta})\right)\right.\right. \\ &+ \left.\left. p(x,y|\hat{\beta})\left(p(x|\hat{\beta}) - p(x|(x,y)_{1:N})\right)\right| \geq \frac{t}{c_3^2}\right) \\ &\leq \mathrm{P}\left(\left|p(x|(x,y)_{1:N}) - p(x|\hat{\beta})\right| \geq \frac{t}{2c_3^2 c_4}\right) \\ &+ \mathrm{P}\left(\left|p(x,y|(x,y)_{1:N}) - p(x,y|\hat{\beta})\right| \geq \frac{t}{2c_3^2 c_4}\right) \end{aligned} \tag{75}$$

Therefore, if t satisfies:

$$1 > t \geq \max \begin{cases} \frac{4c_3^2 c_4^2 |A|^{5/2} |B|^{3/2} p_{\max}(\hat{\alpha}) c_1^{-N}}{p(\hat{\alpha}) + |A| p_{\max}(\hat{\alpha}) c_1^{-N}} \\ \frac{4c_3^2 c_4^2 |B|^{3/2} p_{\max}(\hat{\beta}|\hat{\alpha}) c_2{}^{-N}}{p(\hat{\beta}|\hat{\alpha}) + |B| p_{\max}(\hat{\beta}|\hat{\alpha}) c_2{}^{-N}} \end{cases} \tag{76}$$

we can use the results of proposition 4 to obtain the results. □

We can also obtain the following convergence corollary from proposition 5:

Corollary 6. *If $\hat{\beta} \in R$, $\operatorname{plim}_{N \to \infty} p(y|x_{1:N}, x) = p(y|x, \hat{\beta})$.*

Proof. The proof is identical to the proof of corollary 4 so we omit it. □

C Convergence Speed

We can also observe the convergence speed from $p(\hat{\beta}|x_{1:N})$ to 1 from proposition 3. Specifically, take the derivative of the upper-bound to N in formula (60), we can see that the convergence speed is around

$$O\left(-\left(e^{\frac{\log^2 c_1}{8 \log^2 (c_4/c_3)}}\right)^{-N} - \left(e^{\frac{2 \log^2 c_2}{\log^2 (c_4/c_3)}}\right)^{-N}\right) \tag{77}$$

Therefore, the higher the distinction between different modes and topics, i.e, the higher $\log c_1$ and $\log c_2$, the faster the convergence of the data ICTR.

D Expectation of $\mathrm{KL}(\hat{\beta}\|\beta)$

According to the settings, each topic $\beta \in B$ contains a few sub-topics, then the expectation of $\mathrm{KL}(\hat{\beta}\|\beta)$ depends on KL divergences of those sub-topics:

$$\begin{aligned} \mathbb{E}\left[\mathrm{KL}(\hat{\beta}\|\beta)\right] &= \sum_{m=1}^{M} \mathbb{E}_{\hat{\rho}_m, \rho_m}\left[\mathrm{KL}(\hat{\rho}_m\|\rho_m)\right] \\ &= \sum_{m=1}^{M} \mathbb{E}_{\hat{\rho}_m, \rho_m}\left[\sum_s p(s|\hat{\rho}_m) \log \frac{p(s|\hat{\rho}_m)}{p(s|\rho_m)}\right] \end{aligned} \tag{78}$$

Given that $\hat{\beta}$ and β are different, there at least exists one subtopic is different between them, so:

$$\mathbb{E}\left[\mathrm{KL}(\hat{\beta}\|\beta)\right] \geq \mathbb{E}_{\hat{\rho}, \rho}\left[\mathrm{KL}(\hat{\rho}\|\rho)\right] \tag{79}$$

Note that for each $\rho \in B_*$, the sub-paragraph distribution $p(s|\rho) = p(s|\tilde{\mathbf{A}}_\rho)$ is Markovian, where $\tilde{\mathbf{A}}_\rho = [\boldsymbol{\pi}_\rho, \mathbf{A}_\rho]$ is a row concatenation of the initial probability vector $\boldsymbol{\pi}_\rho$ and transition matrix $\mathbf{A}_\rho$ sampled from $\mathrm{Dir}([\gamma]^{|\Sigma|})$. Let T be the length of s.

Expand the KL divergence, we have

$$\begin{aligned}
&\mathbb{E}_{\hat{\rho},\rho}\left[\mathrm{KL}(\hat{\rho}\|\rho)\right] = \mathbb{E}^{T}_{\hat{\rho},\rho}\left[\mathrm{KL}(\hat{\rho}\|\rho)\right] \\
&= \mathbb{E}_{\tilde{\mathbf{A}}_{\hat{\rho}},\tilde{\mathbf{A}}_{\rho}}\left[\mathrm{KL}\left(p(\cdot|\tilde{\mathbf{A}}_{\hat{\rho}})\|p(\cdot|\tilde{\mathbf{A}}_{\rho}\right)\right] \\
&= \mathbb{E}_{\tilde{\mathbf{A}}_{\hat{\rho}},\tilde{\mathbf{A}}_{\rho}}\Bigg[\sum_{s_{1:T-1}}\sum_{s_T} \\
&p(s_{1:T-1}|\tilde{\mathbf{A}}_{\hat{\rho}})\tilde{\mathbf{A}}_{\hat{\rho}}^{s_{T-1},s_T}\log\frac{p(s_{1:T-1}|\tilde{\mathbf{A}}_{\hat{\rho}})\tilde{\mathbf{A}}_{\hat{\rho}}^{s_{T-1},s_T}}{p(s_{1:T-1}|\tilde{\mathbf{A}}_{\rho})\tilde{\mathbf{A}}_{\rho}^{s_{T-1},s_T}}\Bigg] \\
&= \mathbb{E}^{T-1}_{\hat{\rho},\rho}\left[\mathrm{KL}(\hat{\rho}\|\rho)\right] + \mathbb{E}_{\tilde{\mathbf{A}}_{\hat{\rho}},\tilde{\mathbf{A}}_{\rho}}\Bigg[\sum_{s_{T-1},s_T} \\
&\qquad p(s_{T-1}|\tilde{\mathbf{A}}_{\hat{\rho}})\tilde{\mathbf{A}}_{\hat{\rho}}^{s_{T-1},s_T}\log\frac{\tilde{\mathbf{A}}_{\hat{\rho}}^{s_{T-1},s_T}}{\tilde{\mathbf{A}}_{\rho}^{s_{T-1},s_T}}\Bigg]
\end{aligned} \tag{80}$$

Note that Assumption 3 actually implicit that $p(s_T|\tilde{\mathbf{A}}_\rho)$ is bounded for all T and $\rho \in B_*$. We assume the lower bound is c_5. Then, the second term of the above formula has the following lower bound:

$$\begin{aligned}
&\mathbb{E}_{\tilde{\mathbf{A}}_{\hat{\rho}},\tilde{\mathbf{A}}_{\rho}}\Bigg[\sum_{s_{T-1},s_T} \\
&\qquad p(s_{T-1}|\tilde{\mathbf{A}}_{\hat{\rho}})\tilde{\mathbf{A}}_{\hat{\rho}}^{s_{T-1},s_T}\log\frac{\tilde{\mathbf{A}}_{\hat{\rho}}^{s_{T-1},s_T}}{\tilde{\mathbf{A}}_{\rho}^{s_{T-1},s_T}}\Bigg] \\
&\geq c_5\mathbb{E}_{\tilde{\mathbf{A}}_{\hat{\rho}},\tilde{\mathbf{A}}_{\rho}}\left[\sum_{s_{T-1},s_T}\tilde{\mathbf{A}}_{\hat{\rho}}^{s_{T-1},s_T}\log\frac{\tilde{\mathbf{A}}_{\hat{\rho}}^{s_{T-1},s_T}}{\tilde{\mathbf{A}}_{\rho}^{s_{T-1},s_T}}\right] \\
&= c_5\mathbb{E}_{\tilde{\mathbf{A}}_{\hat{\rho}}}\left[\sum_{s_{T-1},s_T}\tilde{\mathbf{A}}_{\hat{\rho}}^{x_{T-1},x_T}\log\tilde{\mathbf{A}}_{\hat{\rho}}^{x_{T-1},x_T}\right] \\
&- c_5\mathbb{E}_{\tilde{\mathbf{A}}_{\hat{\rho}},\tilde{\mathbf{A}}_{\rho}}\left[\sum_{s_{T-1},s_T}\tilde{\mathbf{A}}_{\hat{\rho}}^{x_{T-1},x_T}\log\tilde{\mathbf{A}}_{\rho}^{x_{T-1},x_T}\right] \\
&= c_5|\Sigma|\left[\psi(\gamma+1)-\psi(|\Sigma|\gamma+1)\right] \\
&\qquad\qquad - c_5|\Sigma|\left[\psi(\gamma)-\psi(|\sigma|\gamma)\right] \\
&= \frac{c_5(|\Sigma|-1)}{\gamma}
\end{aligned} \tag{81}$$

where $\psi(x)$ is the digamma function, and we use the property $\psi(x+1) = \psi(x) + 1/x$ to simplify the above formula. Therefore, we have:

$$\begin{aligned}
\mathbb{E}^{T}_{\hat{\rho},\rho}\left[\mathrm{KL}(\hat{\rho}\|\rho)\right] \geq{}& \mathbb{E}^{T-1}_{\hat{\rho},\rho}\left[\mathrm{KL}(\hat{\rho}\|\rho)\right] + \frac{c_5(|\Sigma|-1)}{\gamma} \\
\geq{}& \mathbb{E}^{T-2}_{\hat{\rho},\rho}\left[\mathrm{KL}(\hat{\rho}\|\rho)\right] + \frac{2c_5(|\Sigma|-1)}{\gamma} \\
&\cdots \\
\geq{}& \frac{Tc_5(|\Sigma|-1)}{\gamma}
\end{aligned} \tag{82}$$

Therefore, the expectation of $\mathrm{KL}(\hat{\beta}\|\beta)$ is bounded:

$$\mathbb{E}\left[\mathrm{KL}(\hat{\beta}\|\beta)\right] \geq \frac{Tc_5(|\Sigma|-1)}{\gamma} \tag{83}$$

We can see that the lower the value of γ, the larger the expected topic-wise KL divergence, and the more significant the topic distinction is.

E Computation of Prompt and Topic-wise ICTR

According to the definition, given an in-context prompt $x_{1:N}$, where each sample $x_n \sim p(x|\hat{\beta})$, ICTR is the probability that the language model generates a paragraph also belonging to topic $\hat{\beta}$. Thus, to measure the belongness of the generated paragraph, we use the mixture of topic-paragraph models $\sum_{\beta\in B}\pi^{\beta}_{x_{1:N}}p(x|\beta)$ to fit the ICG distribution of the target language model $p_{\mathrm{LM}}(x|x_{1:N})$. Here, $p(x|\beta)$ is fixed, and we sample L_1 paragraphs from $p_{\mathrm{LM}}(x|x_{1:N})$ to fit $\pi^{\beta}_{x_{1:N}}$ using EM algorithm (Bishop and Nasrabadi, 2006) as shown in Algorithm 1. As a result, the estimated $\pi^{\hat{\beta}}_{x_{1:N}}$ can represent the ICTR given the in-context prompt $x_{1:N}$.

We further compute the topic-wise ICTR to summarize the ICG ability of a specific topic. Topic-wise ICTR is the expectation of prompt-wise ICTR:

$$\pi^{\beta}_{N} = \mathbb{E}_{p(x_{1:N}|\beta^N)}\left[\pi^{\beta}_{x_{1:N}}\right] \simeq \frac{1}{L_2}\sum_{l=1}^{L_2}\pi^{\beta}_{x^l_{1:N}} \tag{84}$$

Here, we use Monte-Carlo sampling to estimate the expectation, where $x^l_{1:N}$ is the l-th sample of $\prod_{n=1}^{N}p(x_n|\hat{\beta})$. Due to the large number of the topics (531441) in the pretrained distribution, for simplicity, L_1 and L_2 are both set to 1. Thus, the evaluation of a model just requires 531441 forward passes, where the time consumption is acceptable. In-context prompts for evaluation is shown in Figure 3.

Algorithm 1 Prompt-wise ICTR computation

```
Randomly initialize $\pi_{x_{1:N}}^{\beta}$.
for $l = 1, \cdots, L_1$ do
    $x^l \sim p_{\mathrm{LM}}(x|x_{1:N})$
end for
while not convergence do
    for $l = 1, \cdots, L_1$ do
        $\omega_{x_{1:N}}^{\beta,l} = \frac{\pi_{x_{1:N}}^{\beta} p(x^l|\beta)}{\sum_{\beta' \in B} \pi_{x_{1:N}}^{\beta'} p(x^l|\beta)}$
    end for
    $\pi_{x_{1:N}}^{\beta} = \frac{\sum_{l=1}^{L_1} \omega_{x_{1:N}}^{\beta,l}}{L_1}$
end while
$p_{\mathrm{LM}}(\beta|x_{1:N}) \leftarrow \pi_{x_{1:N}}^{\beta}$
return $p_{\mathrm{LM}}(\beta|x_{1:N})$
```