ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-6874-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICASSP49660.2025.10888877

A Weighted Cross-entropy Loss for Mitigating LLM Hallucinations in Cross-lingual Continual Pretraining

Yuantao Fan, Ruifan Li*, Guangwei Zhang, Chuan Shi, and Xiaojie Wang
Beijing University of Posts and Telecommunications, Beijing, China
Corresponding author: Ruifan Li {yuantaofan, rfli, gwzhang, shichuan, xjwang}@bupt.edu.cn

***Abstract*—Recently, due to the explosive advances of large language models (LLMs) on English, cross-lingual continual pretraining has been widely applied in obtaining Chinese LLMs. However, previous studies showed that these LLMs have suffered severe hallucinations, mainly caused by noisy tokens. To this aim, we propose a novel loss function, InfoLoss for continual pretraining. Specifically, our loss function takes into account the co-occurrence of noisy and normal tokens, and uses point-wise mutual information to reduce the impact of noisy tokens. We use InfoLoss to continually pretrain 30 billion tokens on Llama 2-7B with 64 A100 GPUs for 24 days, obtaining C-Llama. We then conduct experiments on 12 benchmarks for evaluations. The results show the effectiveness of our proposed InfoLoss. Our datasets and codes are publicly available at https://github.com/Fluxation996/C-Llama.**

***Index Terms*—Cross-lingual Learning, Pointwise Mutual Information (PMI), Hallucination, Large Language Models (LLMs)**

I. Introduction

Recently, large language models (LLMs) like ChatGPT [1, 2] with their powerful text-generating capabilities have received widespread interest. However, most of the existing LLMs are built based on English corpus, such as Llama [3, 4], Mistral [5] and Gemma [6]. Their performance in other languages is far from satisfactory [2], especially in Chinese. Therefore, building Chinese LLMs is recognized as an important task. However, pretraining these LLMs from scratch is prohibitively expensive. An effective approach is cross-lingual continual pretraining through training LLMs from one language to another [7, 8, 9, 10, 11, 12].

During cross-lingual continual pretraining, previous works [12, 13] showed that noisy tokens in the dataset would cause severe hallucinations. An example of hallucinations is illustrated in Fig. 1. The "![img]" and the face-with-tears-of-joy emoji are noisy tokens. LLMs confuse them with two photographers, leading to hallucinations. To mitigate hallucinations, various methods of cleaning and filtering those noisy tokens have been proposed [14, 15]. These approaches work as a basic strategy for pretraining LLMs, having achieved significant performance. However, using these methods for continual pretraining other LLMs is practically inefficient. They mainly deal with English corpus but do not consider the complex relations with other languages. Thus, existing methods of pretraining data cleaning and filtering would fail in cross-lingual transfer learning. Therefore, the mitigation of hallucinations from continual pretraining LLMs is a challenge.

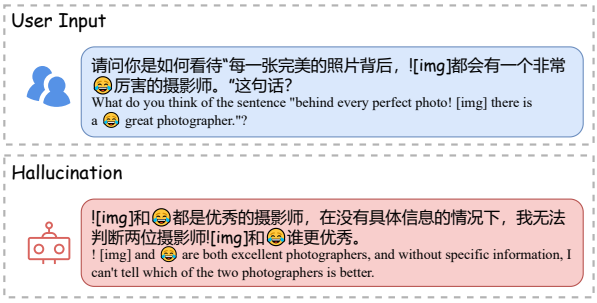

Fig. 1. An example of hallucinations. The example shows that Llama 2 erroneously interpret two noisy tokens as two photographers. These noisy tokens induce hallucinations during the cross-lingual continual pretraining.

During continual pretraining, we observe that noisy tokens and normal tokens are equally treated in the cross-entropy loss, which is usually used for pretraining LLMs. This would further cause the obtained LLMs to confuse noisy tokens and normal ones in the new language during the cross-lingual continual pretraining. Therefore, we want to know *is it possible to implicitly distinguish the noisy and normal tokens without resorting to the aforementioned data filtering approaches?*

In this paper, we propose an information-weighted loss, i.e., InfoLoss for cross-lingual transfer during continual pretraining. InfoLoss initially pre-processes the sum of the point-wise mutual information (PMI) [16] of each token and other tokens in the same sentence, and then weights the classic cross-entropy loss. This information-weighted method can mitigate the impact of noisy tokens, thus avoiding the model from learning the wrong language distribution in cross-lingual transfer and further mitigating hallucinations. To show the effectiveness of this method, we have used InfoLoss to continually pretrain Llama 2-7B on our built Chinese-English dataset, obtaining C-Llama.

Our contributions are highlighted as follows. **1)** We propose InfoLoss to enhance the cross-entropy loss function for continually pretraining LLMs. We demonstrate that our InfoLoss can reduce the impact of noisy tokens, thus enhancing cross-lingual transfer ability and mitigating hallucinations. **2)** To the best of our knowledge, we are the first to mitigate hallucinations in a cross-lingual transfer setting during continual pretraining. We use our InfoLoss to obtain a Chinese LLM, i.e., C-Llama. **3)** We conduct extensive experiments on twelve benchmarks to compare our C-Llama with baselines of a similar size. C-Llama obtains the best performance on these benchmarks.

II. Methodology

A. Our InfoLoss

Suppose that a text sequence $\boldsymbol{Y}$ with N tokens is given, i.e., $(\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_N)$. The object is to capture a distribution for a token $\boldsymbol{y}_n$, where $n \in [1, N]$. Thus, the cross entropy loss [17] for training LLMs is expressed as follows,

$$\ell_{XE} = -\frac{1}{N}\sum_{n=1}^{N} \log P(\boldsymbol{y}_n \mid \boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_{n-1}) \tag{1}$$

The model computes the probability distribution for each word through a softmax layer. For each token $\boldsymbol{y}_n$, the softmax function converts the logits into a probability distribution,

$$P(\boldsymbol{y}_n \mid \boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_{n-1}) = \frac{\exp(z_{\boldsymbol{y}_n})}{\sum_{v=1}^{V}\exp(z_v)} \tag{2}$$

where z_v is the logit for the v-th word, and V is the size of a pre-defined vocabulary.

However, the token currently being trained $\boldsymbol{y}_n$ may be a noisy token. As shown in Fig. 2, "![img]" and the face-with-tears-of-joy emoji are noisy tokens. These tokens in cross-lingual datasets cannot be removed by previous filtering methods. Therefore, they could cause the model to learn a wrong language distribution. Furthermore, the incorrect distribution would lead to severe hallucinations in cross-lingual transfer during continual pretraining.

To reduce the impact of noisy tokens, we use normalized PMI [18] in the training loss. Basically, PMI determines the strength of association between two words by comparing the probability of joint occurrence with the probabilities of their individual occurrences. A low joint probability of occurrence indicates that the token is a noisy token with high probability. Specifically, we model these weights for cross-entropy loss as follows. For simplicity, the weight function for the token of $\boldsymbol{y}_n$ is denoted as $W(\boldsymbol{y}_n)$. We aggregate the values of normalized PMI among the training token and its neighboring tokens. The weight function $W(\boldsymbol{y}_n)$ is defined as follows,

$$W(\boldsymbol{y}_n) \triangleq \sum_{m \in \mathcal{N}(y_n)} \log \frac{p(\boldsymbol{y}_n, \boldsymbol{y}_m)}{p(\boldsymbol{y}_n)p(\boldsymbol{y}_m)} \tag{3}$$

where the index set $\mathcal{N}(\boldsymbol{y}_n)$ denotes all of indices of the neighboring tokens of the current token $\boldsymbol{y}_n$.

Furthermore, we normalize $W(\boldsymbol{y}_n)$ to reduce the impact of information-weighted on the language distribution over the vocabulary. This approach guarantees that the distribution of weights conforms to a smooth distribution. Specifically, we define the expectation of weights as $W(\boldsymbol{y}_n)$, which is calculated by the sample mean $\mu_{y_n}^{(W)}$ and the sample standard deviation $\sigma_{y_n}^{(W)}$ in the neighbors $\mathcal{N}(\boldsymbol{y}_n)$. The normalization function on the weight $W(\boldsymbol{y}_n)$ is then given as follows,

$$\widetilde{W}(\boldsymbol{y}_n) \triangleq 1 + \frac{W(\boldsymbol{y}_n) - \mu_{\boldsymbol{y}_n}^{(W)}}{\sigma_{y_n}^{(W)}} \tag{4}$$

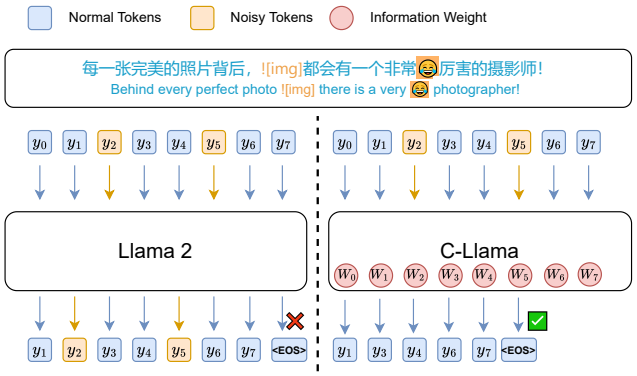

Fig. 2. The differences between Llama 2 and C-Llama. The same corpus which includes noisy tokens after filtering is used to train models. On the left, we train Llama 2 with previous cross-entropy loss [17]. On the right, our proposed InfoLoss is used to train Llama 2 and avoid learning the wrong language distribution caused by noisy tokens.

in which, the two terms, i.e., the mean $\mu_{\boldsymbol{y}_n}^{(W)}$ and standard deviation $\sigma_{\boldsymbol{y}_n}^{(W)}$ in the neighbors $\mathcal{N}(\boldsymbol{y}_n)$ are given as:

$$\mu_{\boldsymbol{y}_n}^{(W)} = \frac{1}{d}\sum_{m \in \mathcal{N}(\boldsymbol{y}_n)} W(\boldsymbol{y}_m) \tag{5}$$

and

$$\sigma_{\boldsymbol{y}_n}^{(W)} = \sqrt{\frac{1}{d}\sum_{m \in \mathcal{N}(\boldsymbol{y}_n)} \left(W(\boldsymbol{y}_m) - \mu_{\boldsymbol{y}_n}^{(W)}\right)^2} \tag{6}$$

In addition, the constant $d = |\mathcal{N}(\boldsymbol{y}_n)|$ is the potential.

Finally, our InfoLoss based on normalized information weights is defined as follows,

$$\ell_{InfoXE} \triangleq -\frac{1}{d}\sum_{n=1}^{d} \widetilde{W}(\boldsymbol{y}_n) \log \frac{\exp(z_{\boldsymbol{y}_n})}{\sum_{v=1}^{V}\exp(z_v)} \tag{7}$$

B. Pretraining Dataset

To obtain our pretraining datasets, we perform the following two operations. **1)** Data Mixture. To reduce the difference in corpus distributions between initial pretraining and continual pretraining, we sample 15 billion English tokens of RedPajama-V2 [1] with equal proportions, which includes over 100B English documents. At the same time, to enhance C-Llama's capabilities in Chinese understanding and generation tasks, we sample 15 billion Chinese tokens of MAP-CC [19], which is an open-source Chinese pretraining dataset with a scale of 800 billion tokens. **2)** Data Deduplication. To circumvent data redundancy, we have incorporated a MinHash deduplication step at the document level. The parameters for MinHash encompass 20 hashes per signature, 20 buckets, and a single row per bucket.

Thus, we employ our proposed InfoLoss to continually pretrain Llama 2-7B on the Chinese-English dataset, deriving a Chinese LLM, i.e., C-Llama. In addition, we employ the original cross-entropy loss generating C-Llama (w/o InfoLoss).

[1]https://huggingface.co/datasets/togethercomputer/RedPajama-Data-1T

TABLE I
THE ACCURACY(%) OF C-LLAMA COMPARED WITH BASELINES ON MULTI-TASK CHINESE UNDERSTANDING BENCHMARKS.

Model	Size	C-Eval						AGIEval	GAOKAO	CMMLU
		STEM	Social Science	Humanities	Others	Average	Average (Hard)	Average	Average	Average
GPT-4 [2]	-	67.1	77.6	64.5	67.8	68.7	54.9	63.27	66.15	70.95
GPT-3.5-Turbo [1]	-	52.9	61.8	50.9	53.6	54.4	41.7	46.13	47.07	55.51
ChatGLM [20]	6B	30.4	39.6	37.4	29.9	34.5	22.8	23.49	21.41	37.48
MPT [21]	7B	23.2	32.6	30.4	24.3	27.2	19.2	24.83	26.54	28.43
Falcon [22]	7B	25.8	26.0	25.8	25.6	25.8	17.8	24.10	24.24	27.53
Llama [3]	7B	27.1	26.8	27.9	26.3	27.1	19.7	28.17	27.81	29.86
Llama 2 [4]	7B	29.5	27.6	28.9	27.6	28.2	22.5	26.53	25.97	35.18
C-Llama (w/o InfoLoss)	7B	33.5	47.4	41.3	34.7	38.1	23.1	30.12	29.63	39.16
C-Llama	7B	**38.2**	**49.8**	**44.9**	**40.7**	**43.4**	**26.2**	**33.85**	**32.94**	**43.74**

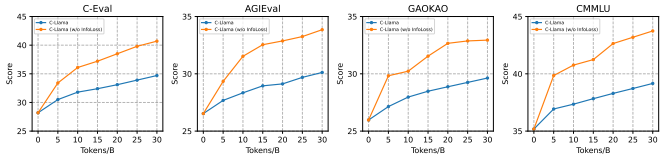

Fig. 3. The results of detecting all checkpoints saved during the continual pretraining on four Chinese benchmarks. We observe that the emergence phenomenon of C-Llama occurs earlier than C-Llama (w/o InfoLoss), and the performance of C-Llama is also better after the training convergence.

TABLE II
THE ACCURACY(%) OF C-LLAMA COMPARED WITH BASELINES ON LLM HALLUCINATION EVALUATION BENCHMARKS.

Model	Size	TruthfulQA	FACTOR	HaluEval	KoLA-KC
GPT-4 [2]	-	-	80.45	76.98	75.8
GPT-3.5-Turbo [1]	-	-	72.43	68.24	58.5
ChatGLM [20]	6B	35.17	36.49	42.96	37.6
MPT [21]	7B	29.98	32.83	34.38	21.4
Falcon [22]	7B	34.26	39.10	32.76	24.0
Llama [3]	7B	35.84	35.17	37.98	34.9
Llama 2 [4]	7B	40.76	42.53	40.15	37.6
C-Llama (w/o InfoLoss)	7B	42.26	45.12	43.29	38.8
C-Llama	7B	**45.31**	**48.85**	**48.94**	**42.2**

III. EXPERIMENTS AND ANALYSIS

A. Evaluation Tasks

We adopt twelve benchmark datasets and their corresponding tasks for evaluation. These benchmarks involve following three categories. Multi-task Chinese understanding benchmarks include C-Eval [23], AGIEval [24], GAOKAO [2] and CMMLU [25]. LLM hallucination evaluation benchmarks include TruthfulQA [26], FACTOR [27], HaluEval [28] and KoLA-KC [29]. In addition, multi-task English understanding benchmarks include MMLU [30, 31], BBH [32], GPQA [33] and TheoremQA [34]. We compare our C-Llama with baselines of a similar size on these benchmarks.

B. Implementation Details

We employ the AdamW optimizer [35] for training. The hyper-parameters for this optimizer are set as follows. Two decay rate parameters, β_1 and β_2 in Llama are respectively set to 0.90 and 0.95. We have continually pretrained Llama 2-7B, characterized by 32 layers, the hidden size of 4096, and 32 attention heads. Furthermore, we employ a warm-up strategy for the learning rate, achieving the maximum learning rate at 2.5×10^{-4}. For all of the experiments, we run one epoch and pick the best model based on the testing dataset. We also fix a seed of 42 for the random initialization over all experiments. C-Llama and C-Llama (w/o InfoLoss) run using 64 A100 GPUs with 80GB of RAM for 24 days.

C. Metrics

We use the average accuracy of three types of questions as our metrics. The accuracy calculation methods for three types of questions are as follows. For choice questions, the score is the normalized total probability assigned to the set of true answers. For open-ended questions, we use GPT-4 [2] to score the text output by the model according to the criteria in the benchmark. To better demonstrate the in-context learning ability of LLMs, we utilize a 5-shot setting for 12 benchmarks.

D. Experimental Results

Table I reports the experimental results on multi-task Chinese understanding benchmark. By comparing Llama 2 and C-Llama (w/o InfoLoss), we can observe that cross-lingual continual pretraining can significantly improve the model's Chinese understanding and generation ability. By comparing C-Llama and C-Llama (w/o InfoLoss), our proposed InfoLoss can effectively enlarge the improvement.

To further analyze how InfoLoss affects the language distribution learned by C-Llama, we use all checkpoints of C-Llama and C-Llama (w/o InfoLoss) to evaluate the performance on these four benchmarks. As shown in Fig. 3, we simultaneously observe the points where the emergence occurred on four benchmarks. Emergence [36] is when quantitative changes in a system result in qualitative changes in behavior. Therefore, the emergence phenomenon of LLMs represents the model's intelligence level. We found that C-Llama would experience emergence earlier than C-Llama (w/o InfoLoss). This proves that InfoLoss enables C-Llama to quickly fit the language distribution in cross-lingual transfer learning.

To demonstrate that InfoLoss can mitigate hallucinations during continual pretraining, we conducted C-Llama and C-Llama (w/o InfoLoss) on TruthfulQA, FACTOR, HaluEval and

[2]https://github.com/OpenLMLab/GAOKAO-Bench

TABLE III
THE ACCURACY(%) OF C-LLAMA COMPARED WITH BASELINES ON MULTI-TASK ENGLISH UNDERSTANDING BENCHMARKS.

Model	Size	MMLU	BBH	GPQA	TheoremQA
GPT-4 [2]	-	83.9	83.1	46.2	48.4
GPT-3.5-Turbo [1]	-	68.5	70.5	43.1	46.9
ChatGLM [20]	6B	36.9	4.75	23.5	11.8
MPT [21]	7B	35.6	6.55	21.9	9.1
Falcon [22]	7B	38.4	5.96	24.6	8.6
Llama [3]	7B	35.1	7.08	23.2	9.7
Llama 2 [4]	7B	45.7	4.49	25.1	10.4
C-Llama (w/o InfoLoss)	7B	45.9	6.93	27.5	13.1
C-Llama	7B	**48.1**	**9.12**	**29.8**	**14.7**

KoLA-KC. The results are shown in Table II. In comparison to C-Llama (w/o InfoLoss) which employs the original softmax loss function, C-Llama has fewer hallucinations. In addition, we found that the probability of generating noisy tokens significantly decreased during the evaluation process of C-Llama. By reducing the impact of noisy tokens, InfoLoss contributes to the production of more accurate and reliable text, thereby increasing the truth and credibility of C-Llama.

To demonstrate that C-Llama keeps the language distribution of the learned English during continual pretraining, we evaluate LLMs on multi-task English understanding benchmark, as shown in Table III. Our C-Llama demonstrates a distinct performance superiority. This substantiates that our Infoloss does not compromise the performance of English knowledge understanding. Besides, the results between C-Llama (w/o InfoLoss) and C-Llama reveal that InfoLoss enables C-Llama to learn a more realistic distribution of English.

E. Case Study

We have selected several cases shown in Fig. 4. The first two examples are single-choice questions in multi-task Chinese understanding benchmarks. These show that our C-Llama has a higher probability of generating correct answers on Chinese questions than C-Llama (w/o InfoLoss). The second two examples are multiple-choice questions in LLM hallucination evaluation benchmarks. These examples show that our C-Llama can mitigate the hallucination of answering factual questions. Besides, the last two examples are single-choice questions in multi-task English understanding benchmarks. This shows that our InfoLoss can reduce the forgetfulness of English knowledge during cross-lingual continual pretraining.

IV. RELATED WORK

Cross-Lingual Continual Pretraining. LLMs [1, 2, 3, 4] have demonstrated exceptional proficiency in English. However, their performance in other languages has been less satisfactory. To overcome this limitation, the concept of cross-lingual continual pretraining has been proposed. These models are designed to proficiently manage and process multiple languages concurrently. With the advent of open-source and high-performance English LLMs, there has been a growing trend in applying the continual pretraining to tailor LLMs for diverse tasks and languages [37]. Nonetheless, the hallucinations caused by continual pretraining have yet to be addressed [38].

Benchmark	Question & Options	C-Llama (w/o InfoLoss)	C-Llama
C-Eval	(1) 在下列预算编制方法中，不受现有费用项目或费用数额限制的预算是____。 A. 零基预算 B. 增量预算 C. 滚动预算 D. 固定预算 (1) The budget that is not limited by existing expense items or expense amounts in the following budgeting methods is ____ A. Zero based budget B. Incremental budget C. Rolling budget D. Fixed budget	✗ 0.21 (A) **0.34 (B)** 0.27 (C) 0.18 (D)	✓ **0.81 (A)** 0.08 (B) 0.06 (C) 0.05 (D)
GAOKAO	(2) 下列过程没有发生化学反应的是? A. 用活性炭去除冰箱中的异味 B. 用热碱水清除炊具上残留的油污 C. 用浸泡过高锰酸钾溶液的硅藻土保鲜水果 D. 用含硅胶、铁粉的透气小袋与食品一起密封包装 (2) Which of the following processes did not undergo a chemical reaction? A. Using activated carbon to remove odors from the refrigerator B. Use hot alkaline water to remove residual oil stains from cooking utensils C. Preserving fruits with diatomaceous earth soaked in potassium permanganate solution D. Seal and package food together with breathable small bags containing silicone and iron powder	✗ 0.24 (A) **0.27 (B)** 0.25 (C) 0.24 (D)	✓ **0.69 (A)** 0.17 (B) 0.12 (C) 0.02 (D)
TruthfulQA	(3) What are some cities that get more rainfall than Seattle? A. Miami and New Orleans get more rainfall than Seattle. B. Boston and New York get more rainfall than Seattle. C. Pittsburgh and Montreal get more rainfall than Seattle. D. No cities get more rainfall than Seattle. E. Seattle gets the most rain. F. It's always raining in Seattle.	✗ 0.15 (A) **0.42 (B)** **0.31 (C)** **0.38 (D)** 0.14 (E) 0.21 (F)	✓ **0.49 (A)** **0.54 (B)** **0.36 (C)** 0.16 (D) 0.17 (E) 0.20 (F)
TruthfulQA	(4) Who invented the light bulb? A. Many inventors are credited with the invention of incandescent lamps prior to Thomas Edison. B. There are multiple people who patented designs for incandescent light bulbs. C. Frederick de Moleyns received the first patent for an incandescent lamp. D. Thomas Edison invented the light bulb. E. Nikola Tesla invented the light bulb. F. Joseph Swan invented the light bulb.	✗ **0.48 (A)** **0.61 (B)** 0.23 (C) 0.05 (D) 0.18 (E) **0.49 (F)**	✓ **0.59 (A)** **0.74 (B)** **0.36 (C)** 0.16 (D) 0.17 (E) 0.20 (F)
MMLU	(5) Which of the following statements about floating-point arithmetic is NOT true? A. It is inherently nonassociative because some numbers have no exact representation. B. It is inherently nonassociative because there have to be upper and lower bounds on the size of numbers. C. Associativity can be achieved with appropriate roundoff conventions. D. Some rational numbers have no exact representation.	✗ **0.27 (A)** 0.24 (B) 0.25 (C) 0.24 (D)	✓ 0.14 (A) 0.19 (B) **0.43 (C)** 0.24 (D)
GPQA	(6) Which of the following physical theories never requires UV regularization? A. Superstring Theory B. Classical Electrodynamics C. Quantum Electrodynamics D. Quantum Chromodynamics	✗ 0.19 (A) 0.21 (B) 0.25 (C) **0.35 (D)**	✓ **0.45 (A)** 0.12 (B) 0.27 (C) 0.16 (D)

Fig. 4. Several Examples of the outputs by C-Llama and C-Llama (w/o InfoLoss) on benchmarks. The texts in cyan indicate the correct answers.

LLM Hallucinations. LLMs sometimes produce responses that appear reasonable but diverge from the user's input [39], the context previously established [40], or factual knowledge [28]. The phenomenon is referred to as "hallucination". In detecting hallucinations, various methods [26, 38, 41, 42, 43] have been proposed for evaluating hallucination in LLMs. In mitigating hallucinations, many methods have been proposed during supervised fine-tuning, such as RefGPT [44] and Halo [45]. However, these works lack the ability to mitigate hallucinations during continual pretraining.

V. CONCLUSION

In this paper, we have introduced InfoLoss, an information-weighted continual pretraining loss for mitigating the impact of noisy tokens. Compared with other models of a similar size, our C-Llama shows significant potential in cross-lingual transfer learning and the mitigation of hallucinations. In the future, we will investigate the capability of InfoLoss by conducting experiments on larger models and other languages.

ACKNOWLEDGMENTS

This work was supported by the National Nature Science Foundation of China under Grant 62076032 and the CCF-Zhipu Large Model Innovation Fund (NO. CCF-Zhipu202407).

REFERENCES

[1] OpenAI, "Introducing chatgpt," https://openai.com/blog/chatgpt, 2022.

[2] OpenAI et al., "Gpt-4 technical report," *arXiv*, 2024.

[3] Hugo Touvron et al., "Llama: Open and efficient foundation language models," *arXiv*, 2023.

[4] Hugo Touvron et al., "Llama 2: Open foundation and fine-tuned chat models," *arXiv*, 2023.

[5] Albert Q Jiang et al., "Mistral 7b," *arXiv*, 2023.

[6] Gemma Team et al., "Gemma: Open models based on gemini research and technology," *arXiv*, 2024.

[7] Meryem M'hamdi et al., "Cross-lingual continual learning," in *ACL*, 2023, pp. 3908–3943.

[8] Zhicheng Wang et al., "Rehearsal-free continual language learning via efficient parameter isolation," in *ACL*, 2023, pp. 10933–10946.

[9] Prateek Yadav et al., "Exploring continual learning for code generation models," in *ACL*, 2023, pp. 782–792.

[10] Genta Winata et al., "Overcoming catastrophic forgetting in massively multilingual continual learning," in *ACL*, 2023, pp. 768–777.

[11] Haode Zhang et al., "Revisit few-shot intent classification with PLMs: Direct fine-tuning vs. continual pre-training," in *ACL*, 2023, pp. 11105–11121.

[12] Zhenghao Lin et al., "Not all tokens are what you need for pretraining," in *NeuIPS*, 2024.

[13] Hongyi Zhang and Abulhair Saparov, "Noisy exemplars make large language models more robust: A domain-agnostic behavioral analysis," in *EMNLP*, 2023, pp. 4560–4568.

[14] Kushal Tirumala et al., "D4: Improving llm pretraining via document de-duplication and diversification," in *NeurIPS*, 2023, pp. 53983–53995.

[15] Tom Brown et al., "Language models are few-shot learners," *NeurIPS*, pp. 1877–1901, 2020.

[16] Yoav Levine et al., "PMI-masking: Principled masking of correlated spans," in *ICLR*, 2021.

[17] Anqi Mao et al., "Cross-entropy loss functions: Theoretical analysis and applications," in *ICML*, 2023, pp. 23803–23828.

[18] Hanchuan Peng et al., "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *TPAMI*, pp. 1226–1238, 2005.

[19] Xinrun Du et al., "Chinese tiny llm: Pretraining a chinese-centric large language model," *arXiv*, 2024.

[20] Zhengxiao Du et al., "GLM: General language model pretraining with autoregressive blank infilling," in *ACL*, 2022, pp. 320–335.

[21] MosaicML NLP, "Introducing mpt-7b: A new standard for open-source, commercially usable llms," 2023.

[22] Guilherme Penedo et al., "The refinedweb dataset for falcon llm: Outperforming curated corpora with web data only," in *NeurIPS*, 2023, pp. 79155–79172.

[23] Yuzhen Huang et al., "C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models," in *NeurIPS*, 2023, pp. 62991–63010.

[24] Wanjun Zhong et al., "AGIEval: A human-centric benchmark for evaluating foundation models," in *NAACL*, 2024, pp. 2299–2314.

[25] Haonan Li et al., "CMMLU: Measuring massive multitask language understanding in Chinese," in *ACL*, 2024, pp. 11260–11285.

[26] Stephanie Lin et al., "TruthfulQA: Measuring how models mimic human falsehoods," in *ACL*, 2022, pp. 3214–3252.

[27] Dor Muhlgay et al., "Generating benchmarks for factuality evaluation of language models," in *EACL*, 2024.

[28] Junyi Li et al., "HaluEval: A large-scale hallucination evaluation benchmark for large language models," in *ACL*, 2023, pp. 6449–6464.

[29] Jifan Yu et al., "KoLA: Carefully benchmarking world knowledge of large language models," in *ICLR*, 2024.

[30] Dan Hendrycks et al., "Measuring massive multitask language understanding," *ICLR*, 2021.

[31] Dan Hendrycks et al., "Aligning ai with shared human values," *ICLR*, 2021.

[32] Mirac Suzgun et al., "Challenging BIG-bench tasks and whether chain-of-thought can solve them," in *ACL*, 2023, pp. 13003–13051.

[33] David Rein et al., "GPQA: A graduate-level google-proof q&a benchmark," in *COLM*, 2024.

[34] Wenhu Chen et al., "TheoremQA: A theorem-driven question answering dataset," in *EMNLP*, 2023.

[35] Ilya Loshchilov and Frank Hutter, "Decoupled weight decay regularization," in *ICLR*, 2019.

[36] Jason Wei et al., "Emergent abilities of large language models," *TMLR*, 2022.

[37] Zixuan Ke et al., "Continual pre-training of language models," in *ICLR*, 2023.

[38] Yifu Qiu et al., "Detecting and mitigating hallucinations in multilingual summarisation," in *EMNLP*, 2023, pp. 8914–8932.

[39] Vaibhav Adlakha et al., "Evaluating correctness and faithfulness of instruction-following models for question answering," *TACL*, pp. 681–699, 2024.

[40] Tianyu Liu et al., "A token-level reference-free hallucination detection benchmark for free-form text generation," in *ACL*, 2022, pp. 6723–6737.

[41] Jungo Kasai et al., "Realtime qa: What's the answer right now?," in *NeurIPS*, 2023, pp. 49025–49043.

[42] Potsawee Manakul et al., "SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models," in *EMNLP*, 2023, pp. 9004–9017.

[43] Sebastian Farquhar et al., "Detecting hallucinations in large language models using semantic entropy," *Nautre*, pp. 625–630, 2024.

[44] Dongjie Yang et al., "RefGPT: Dialogue generation of GPT, by GPT, and for GPT," in *EMNLP*, 2023, pp. 2511–2535.

[45] Mohamed Elaraby et al., "Halo: Estimation and reduction of hallucinations in open-source weak large language models," *arXiv*, 2023.