Fill-in Language Models

Anonymous ACL submission

Abstract

Language models have become the backbone of today's AI systems. However, their predominant left-to-right (L2R) generation limits the use of bidirectional context, which is essential for tasks that involve filling text in the middle. We propose the **F**ill-**i**n **L**anguage **M**odel (FiLM) that allows for flexible generation at any position without adhering to a specific generation order. Its training extends the masked language modeling objective by adopting varying mask probabilities sampled from the Beta distribution to enhance the generative capabilities. During inference, FiLM can seamlessly insert missing phrases, sentences, or paragraphs, ensuring that the outputs are fluent and are coherent with the surrounding context. In both automatic and human evaluations, FiLM outperforms existing infilling methods that rely on L2R language models trained on rearranged text segments. FiLM is easy to implement and can be either trained from scratch or fine-tuned from a L2R language model. Notably, as the model size grows, FiLM's perplexity approaches that of strong L2R language models of similar sizes, indicating FiLM's scalability and potential as a large language model.[1]

1 Introduction

Large language models (LLMs) have demonstrated remarkable success in open-ended text generation and a variety of natural language understanding and reasoning tasks (Brown et al., 2020; Ouyang et al., 2022). The next word prediction objective, inherent in the training of these models, has positioned them as predominantly Causal Language Models (CLMs). However, this confines their generation order to left-to-right, constraining their versatility and applicability, particularly in tasks that require filling in the middle (Zhu et al., 2019). We present the **F**ill-**i**n **L**anguage **M**odel (FiLM), designed for flexible sequence generation at any given positions, taking into account both the preceding and subsequent context. This distinctive attribute opens up avenues for a myriad of applications, including assisting in text editing and revision, automating template filling, and facilitating code completion.

Training FiLM takes insights from both Masked Language Models (MLMs) (Devlin et al., 2018) and text diffusion models (Li et al., 2022; Austin et al., 2021; Dieleman et al., 2022). Unlike MLMs that are trained with a fixed mask ratio, FiLM adopts a strategy inspired by diffusion models that utilize varying noise levels (§3.1). For each training sequence, a mask probability is drawn from the Beta distribution, and each token is masked with this probability. The model then learns to predict the masked tokens based on the surrounding context. This adaptive masking strategy significantly enhances FiLM's generative capacity (§4.1). At decoding time, FiLM has the flexibility to start with either a sequence entirely of masks or a partially complete text interspersed with masks. It progressively replaces one mask with a predicted token at each step. We explore various decoding orders for FiLM (§3.2), and our analysis shows that, besides proceeding from the leftmost mask to the right, selecting the mask position with the minimum entropy is also an effective strategy (§4.1).

FiLM can be either trained from scratch or fine-tuned from off-the-shelf MLMs or CLMs. We experiment with the latter in this paper, which avoids the expensive pretraining stage and is appealing in practice. We develop a method to evaluate the perplexity of any-order language models, enabling a direct comparison between FiLM and CLM (§3.3). When fine-tuned from GPT2-xl, FiLM yields perplexity of 14.03 and 20.32 on the WikiText-103 and One Billion Word datasets respectively, outperforming state-of-the-art text diffusion models of the same size (Gulrajani and Hashimoto, 2023). While these numbers lag behind those by CLM fine-tuned from GPT2-xl on the same data (11.29

[1] Our code will be made public after the review process.

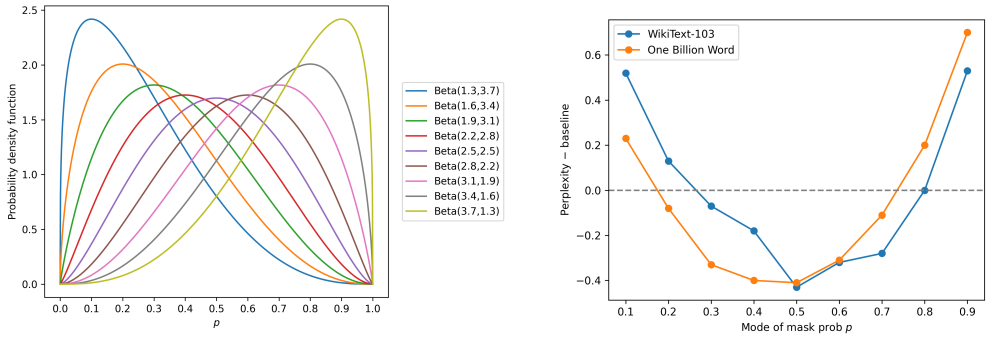

Figure 1: Left: Beta distribution with varying modes. Right: Perplexity of FiLM trained with different Beta distributions, each value depicted as the deviation from the baseline perplexity achieve by the uniform distribution.

and 16.46), we observe a diminishing disparity as model size increases, indicating FiLM's promising potential with scaling (§4.2). FiLM excels in filling text in the middle and significantly surpasses previous infilling methods that employ a CLM trained on rearranged data (Donahue et al., 2020; Fried et al., 2022; Bavarian et al., 2022). In particular, on story completion, FiLM is favored by humans in 48% of the cases over a four times larger baseline model, while the latter is preferred in only 21%, with the remaining cases resulting in ties.

2 Related work

MLMs utilizing bidirectional context have shown exceptional performance in language understanding tasks (Devlin et al., 2018). However, their application in generation has been limited. A line of work represented by T5 (Raffel et al., 2020) and BART (Lewis et al., 2019) employs an MLM-style encoder but remains dependent on a CLM decoder. Another approach attempts to transform MLMs into generators by interpreting them as Markov random fields or energy-based models and using Markov-chain Monte-Carlo (MCMC) sampling algorithms for decoding (Wang and Cho, 2019; Goyal et al., 2021). Despite the intricacy of these techniques, the generative performance of such models lags significantly behind that of CLM.

Apart from FiLM, models like XLNet (Yang et al., 2019), GLM (Du et al., 2021), and BLM (Shen et al., 2020) can generate text in a non left-to-right order. XLNet and GLM have primarily been developed for language understanding tasks and generating text from scratch, with a limited focus on text infilling performance. BLM is explicitly designed for filling in blanks, but it has a markedly higher perplexity compared to CLM.

3 Fill-in language model (FiLM)

3.1 Training

Given a training sequence x consisting of tokens $(x_1, \ldots, x_n)$, we first sample the mask probability p according to a noise schedule, then independently mask each token x_i with probability p. Let $\tilde{x} = (\tilde{x}_1, \ldots, \tilde{x}_n)$ denote the resulting masked sequence. FiLM takes $\tilde{x}$ as input and is trained to predict the original tokens in x at the masked positions of $\tilde{x}$, leveraging bidirectional contextual information.

Unlike BERT, which uses a fixed mask probability $p = 0.15$ (Devlin et al., 2018), FiLM samples p from $[0, 1]$ to ensure that the model learns to generate from different numbers of masks. One choice is sampling p from the uniform distribution $U[0, 1]$ (Liao et al., 2020). However, assigning equal weights across different mask probabilities can be suboptimal, as a high mask ratio leaves the model with insufficient information for predictions, while a low mask ratio oversimplifies the task. We therefore turn to the beta distribution characterized by two shape parameters α, β. By tweaking α, β, we can adjust the distribution of p to skew towards different mask probabilities. Specifically, we keep $\alpha + \beta$ at a constant sum of 5 and adjust them to produce modes between 0.1 and 0.9 at intervals of 0.1 (see Fig. 1 Left). We empirically compare the performance of FiLM trained using different noise schedules to determine the most effective strategy.

3.2 Decoding

To complete a partial text $\tilde{x}$, FiLM fills one mask at a step conditioned on the context and previous predictions, until all masks are filled. Technically, FiLM can fill multiple masks per iteration, similar to previous application of conditional MLMs for non-autoregressive machine transla-

tion (Ghazvininejad et al., 2019). However, without source sentence anchors, tokens have strong interdependence and generation quality degrades with independent predictions. Recognizing the challenge, we use a sequential decoding approach here.

We explore four strategies to determine which mask to fill in first. (1) random selection, akin to FiLM's training process; (2) unidirectional generation, either from the leftmost mask to the right or vice versa. This is similar a CLM, but now with access to the subsequent context. (3) selecting the mask position with the minimum entropy, acting as an "easy-first" heuristic; (4) selecting the position with the maximum entropy, a "hard-first" strategy. In the last two approaches, the generation order is not predetermined but dynamically established during the model's decoding process. Once the order is determined, conventional decoding algorithms of CLM (e.g., sampling, greedy decoding, beam search) are also applicable to FiLM. We use sampling in our experiments.

3.3 Perplexity

To evaluate FiLM's perplexity, we compute the probability it assigns to a sequence x under a specific decoding order, as marginalizing over all possible orders is intractable. We start by calculating $p_{\text{len}}(n)$, the frequency of sequences of length n in the training set (adjusted with add-one smoothing to avoid assigning probability 0 on unseen lengths). Then we can determine $\log p_\theta(x;\sigma)$, the log-probability of generating x with order σ, where θ denotes the model parameters. Note that σ can be deterministic, such as left-to-right and right-to-left; random; or adaptive, like the min-entropy and max-entropy strategies discussed in §3.2.

Let $p_\theta(x_{\sigma_t}|x_{\sigma_1},\ldots,x_{\sigma_{t-1}},n)$ be the probability that FiLM predicts x_{σ_t} given an n-length sequence with $x_{\sigma_1},\ldots,x_{\sigma_{t-1}}$ filled. For example, for $n = 4, \sigma = (3,1,4,2), t = 3$, $p_\theta(x_4|x_3,x_1,4)$ is the probability of predicting x_4 at the last position in $(x_1, \texttt{[MASK]}, x_3, \texttt{[MASK]})$. We have $\log p_\theta(x;\sigma) = \log p_{\text{len}}(n) + \sum_{t=1}^{n} \log p_\theta(x_{\sigma_t}|x_{\sigma_1},\ldots,x_{\sigma_{t-1}},n)$. The perplexity is then $\exp\left(-\frac{1}{n+1}\log p_\theta(x;\sigma)\right)$. We divide by $n+1$ to ensure comparability with CLM, which appends an [EOS] token to $(x_1,\ldots,x_n)$ to signify the end of generation, resulting in a total sequence length of $n+1$. In contrast, FiLM determines the sequence length through $p_{\text{len}}(n)$ and then fills in an n-length sequence.

4 Experiments

Datasets We experiment on two language modeling benchmarks WikiText-103 (WT-103; Merity et al., 2016), One Billion Word (1BW; Chelba et al., 2013), and a commonsense story dataset ROCStories (ROC; Mostafazadeh et al., 2016).

Experimental setup We evaluate FiLM's performance when fine-tuned from RoBERTa (Liu et al., 2019) to represent MLMs, and GPT2 (Radford et al., 2019) to represent CLMs. Note that the causal masking should be deactivated when fine-tuning FiLM from a CLM. We use the Adam optimizer with a learning rate of $2\text{e}-5$ and a batch size of 20K tokens, and train for 500K steps on WT-103, 1BW and 50K steps on ROCStories.

4.1 Analysis of FiLM

We assess FiLM's perplexity on WT-103 and 1BW using various training noise schedules and decoding orders to identify the best strategy, employing RoBERTa-base and GPT2-small models for efficiency. We first set the decoding order to left-to-right and examine the impact of training noise schedules. Table 1 in the appendix shows that a fixed schedule $\delta(0.15)$ leads to substantially higher perplexity, underscoring the importance of varying mask probability p for improved generative performance. To elucidate the effects of training with a Beta distribution of different modes, we plot the difference in perplexity compared to $U[0,1]$ in Fig. 1 (Right). The findings corroborate our hypothesis that overly small or large values for the mode of p are suboptimal. The lowest perplexity is achieved at $\text{Beta}(2.5, 2.5)$ with mode 0.5, which we adopt in subsequent experiments.

Table 2 presents the results of different decoding orders. While FiLM is trained to predict a random subset of words, decoding from left to right substantially outperforms decoding in a random order. When fine-tuned from an order-agnostic MLM, FiLM demonstrates near-optimal performance with right-to-left decoding. However, when fine-tuned from a left-to-right CLM, the efficacy of right-to-left decoding significantly degrades, illustrating the resistance in altering a CLM's behavior from left-to-right to right-to-left. The min-entropy order consistently emerges as the second-best strategy, while the max-entropy order proves to be the least effective. Given the simplicity and superior performance of left-to-right decoding, we select this order for FiLM in subsequent experiments.

Dataset	Model	ROUGE-1/2/L
WT-103	CM-1.6B	21.73/7.14/16.66
	FiLM-1.6B	**35.34/13.29/27.26**
1BW	CM-1.6B	33.69/8.88/33.54
	FiLM-1.6B	**42.55/13.35/42.25**
ROC	CM-1.6B	33.15/9.03/31.94
	FiLM-355M	**38.19/13.24/36.49**

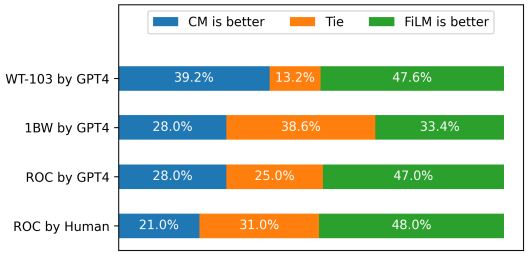

Figure 2: Text infilling results. Left: ROUGE scores. Right: comparative evaluation by GPT4 and human judges.

4.2 Language modeling

We extend the perplexity evaluation across various model sizes and plot the results in Fig. 4. FiLM fine-tuned from bidirectional RoBERTa demonstrates superior performance compared to fine-tuning from unidirectional GPT2, and we hypothesize that directly pretraining with the FiLM objective may further enhance model performance. While FiLM exhibits higher perplexity than CLM, this gap diminishes as the model size increases. Specifically, as the pretrained model grows from 124M GPT2-small to 1558M GPT2-xl, the perplexity gap drops from 5.85 to 2.74 on WT-103 and from 7.96 to 3.86 on 1BW. This trend suggests that FiLM could benefit from further scaling and holds considerable potential as an alternative LLM.

4.3 Text infilling

We remove random spans from WT-103 and 1BW or a random sentence from ROC and task models to fill them in. Previous state-of-the-art methods for infilling, referred to as causal masking (CM), rely on training CLMs on rearranged data (Donahue et al., 2020; Aghajanyan et al., 2022; Fried et al., 2022; Bavarian et al., 2022). For example, the sentence 'They have really good ice cream" would be transformed to "They `[MASK:0]` good `[MASK:1]` cream `[FILL:0]` have really `[FILL:1]` ice" for training. At test time, given prompt "They `[MASK:0]` good `[MASK:1]` cream `[FILL:0]`", the generated tokens are re-integrated into the corresponding mask positions.

We fine-tune FiLM from GPT2-xl on WT-103 and 1BW. Due to ROC's limited size, FiLM fine-tuned from 355M RoBERTa-large achieves a lower loss (2.63) compared to fine-tuning from 1.6B GPT2-xl (2.90), and thus we use RoBERTa-large here. For CM, we fine-tune from GPT2-xl on all three datasets.

To evaluate the model outputs, we compute the ROUGE scores (Lin, 2004) against the original text to measure their overlap. In addition, we conduct GPT4 (Liu et al., 2023) and human evaluations on 500 examples to compare which model output is more grammatically fluent and coherent with the surrounding context, or neither is more fitting than the other. As shown in Fig. 2, FiLM significantly outperforms CM, with average ROUGE score improvements of 10.12 on WT-103, 7.35 on 1BW, and 4.6 on ROC. GPT4 also shows a preference for FiLM over CM by 8.4%, 5.4%, and 19% across these datasets, respectively. Human evaluators further endorse FiLM, with a 27% preference margin on ROC. Table 3 provides details of the agreement analysis between GPT4 and human evaluations.

Fig. 5, 6, 7 showcase example infillings and illustrate qualitative difference between CM and FiLM. Despite CM's attempt to consider subsequent context by artificially altering the text order, it still introduces redundancy by inserting "stress" between "Depression," and "and stress" in the first example in Fig. 5, and commits a logical error by adding "farming, and a world leader in" between "Switzerland is the source of Europe's biggest rivers, supporting agriculture and" and "nuclear power stations" in the second example. In contrast, FiLM generates apt fillings such as "loneliness" and "the construction of new", resulting in coherent sentences. On story completion displayed in Fig. 7, while CM's generated sentence aligns with the prior context, it struggles to link appropriately with the sentences that follow. In contrast, FiLM's outputs exhibit seamless integration, like mentioning "started to itch" considering the subsequent allergy, and adding "But he didn't like to eat it until it was ready" before "So he decided this time he would sneak a piece before dinner", thereby preserving the story's logical flow.

Limitations

FiLM requires a specified number of tokens for infilling, which suits applications where the infilling length is predetermined. For instance, in text editing, users may want to replace a phrase, yet CM might generate excessively lengthy outputs. When the infill length is not known, we will need to either test various lengths or employ an additional model for length prediction. Addressing this challenge and improving FiLM's flexibility in handling variable infill lengths are left to future work.

Ethics Statement

This research complies with the ACL Code of Ethics. FiLM's advancements offer potential benefits in enhancing flexibility of language models to aid content creation. However, we acknowledge the risks associated with the misuse of language models, including the generation of disinformation and automated spam. We advocate for responsible use, transparency, and the development of detection methods for AI-generated content.

References

Armen Aghajanyan, Bernie Huang, Candace Ross, Vladimir Karpukhin, Hu Xu, Naman Goyal, Dmytro Okhonko, Mandar Joshi, Gargi Ghosh, Mike Lewis, et al. 2022. Cm3: A causal masked multimodal model of the internet. *arXiv preprint arXiv:2201.07520*.

Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. 2021. Structured denoising diffusion models in discrete state-spaces. *Advances in Neural Information Processing Systems*, 34:17981–17993.

Mohammad Bavarian, Heewoo Jun, Nikolas Tezak, John Schulman, Christine McLeavey, Jerry Tworek, and Mark Chen. 2022. Efficient training of language models to fill in the middle. *arXiv preprint arXiv:2207.14255*.

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.

Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. 2013. One billion word benchmark for measuring progress in statistical language modeling. *arXiv preprint arXiv:1312.3005*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Sander Dieleman, Laurent Sartran, Arman Roshannai, Nikolay Savinov, Yaroslav Ganin, Pierre H Richemond, Arnaud Doucet, Robin Strudel, Chris Dyer, Conor Durkan, et al. 2022. Continuous diffusion for categorical data. *arXiv preprint arXiv:2211.15089*.

Chris Donahue, Mina Lee, and Percy Liang. 2020. Enabling language models to fill in the blanks. *arXiv preprint arXiv:2005.05339*.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2021. Glm: General language model pretraining with autoregressive blank infilling. *arXiv preprint arXiv:2103.10360*.

Daniel Fried, Armen Aghajanyan, Jessy Lin, Sida Wang, Eric Wallace, Freda Shi, Ruiqi Zhong, Wen-tau Yih, Luke Zettlemoyer, and Mike Lewis. 2022. Incoder: A generative model for code infilling and synthesis. *arXiv preprint arXiv:2204.05999*.

Marjan Ghazvininejad, Omer Levy, Yinhan Liu, and Luke Zettlemoyer. 2019. Mask-predict: Parallel decoding of conditional masked language models. *arXiv preprint arXiv:1904.09324*.

Kartik Goyal, Chris Dyer, and Taylor Berg-Kirkpatrick. 2021. Exposing the implicit energy networks behind masked language models via metropolis–hastings. *arXiv preprint arXiv:2106.02736*.

Ishaan Gulrajani and Tatsunori B Hashimoto. 2023. Likelihood-based diffusion language models. *arXiv preprint arXiv:2305.18619*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori B Hashimoto. 2022. Diffusion-lm improves controllable text generation. *arXiv preprint arXiv:2205.14217*.

Yi Liao, Xin Jiang, and Qun Liu. 2020. Probabilistically masked language model capable of autoregressive generation in arbitrary word order. *arXiv preprint arXiv:2004.11579*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. Gpteval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and evaluation framework for deeper understanding of commonsense stories. *arXiv preprint arXiv:1604.01696*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Tianxiao Shen, Victor Quach, Regina Barzilay, and Tommi Jaakkola. 2020. Blank language models. *arXiv preprint arXiv:2002.03079*.

Alex Wang and Kyunghyun Cho. 2019. Bert has a mouth, and it must speak: Bert as a markov random field language model. *arXiv preprint arXiv:1902.04094*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Wanrong Zhu, Zhiting Hu, and Eric Xing. 2019. Text infilling. *arXiv preprint arXiv:1901.00158*.

APPENDIX

Table 1: Perplexity of FiLM under different training noise schedules.

Dataset	**Pretrained model**	$\delta(0.15)$	$U[0,1]$	$\text{Beta}(2.5, 2.5)$
WT-103	RoBERTa-base	26.76	19.02	**18.59**
1BW	RoBERTa-base	34.68	30.56	**30.15**

Table 2: Perplexity of FiLM using different decoding orders. The **bold** numbers highlight the best perplexity, while the underlined ones denote the second best.

Dataset	**Pretrained model**	**Random**	**L2R**	**R2L**	**Min-Ent**	**Max-Ent**
WT-103	RoBERTa-base	19.95	**18.59**	19.04	18.70	21.38
	GPT2-small	28.42	**21.68**	29.34	23.47	30.68
1BW	RoBERTa-base	31.85	30.15	**29.92**	30.07	33.16
	GPT2-small	41.52	**32.55**	42.73	34.59	44.04

Min-entropy	Max-entropy
[M] [M] [M] [M] [M] [M] [M] [M] [M] [M]	[M] [M] [M] [M] [M] [M] [M] [M] [M] [M]
[M] [M] [M] [M] [M] [M] [M] [M] [M] .	[M] [M] [M] [M] [M] [M] [M] [M] service [M]
Mr [M] [M] [M] [M] [M] [M] [M] [M] .	[M] . [M] [M] [M] [M] [M] [M] service [M]
Mr . [M] [M] [M] [M] [M] [M] [M] .	[M] . [M] [M] thank [M] [M] [M] service [M]
Mr . [M] [M] thank [M] [M] [M] [M] .	[M] . Chairman [M] thank [M] [M] [M] service [M]
Mr . [M] [M] thank you [M] [M] [M] .	Mr . Chairman [M] thank [M] [M] [M] service [M]
Mr . [M] , thank you [M] [M] [M] .	Mr . Chairman [M] thank [M] [M] your service [M]
Mr . [M] , thank you for [M] [M] .	Mr . Chairman , thank [M] [M] your service [M]
Mr . [M] , thank you for your [M] .	Mr . Chairman , thank [M] [M] your service .
Mr . [M] , thank you for your service .	Mr . Chairman , thank you [M] your service .
Mr . Chairman , thank you for your service .	Mr . Chairman , thank you for your service .

Figure 3: An illustration of FiLM decoded using min-entropy and max-entropy orders. The mask position selected to be filled in at each step is highlighted in the green color. The min-entropy strategy generates text in a segmented manner, sequentially predicting cohesive phrases such as "thank you" and "for your service", deferring the more uncertain name after "Mr." to the final stage. In contrast, the max-entropy strategy opts for a "challenging" order, selecting distant positions at each step.

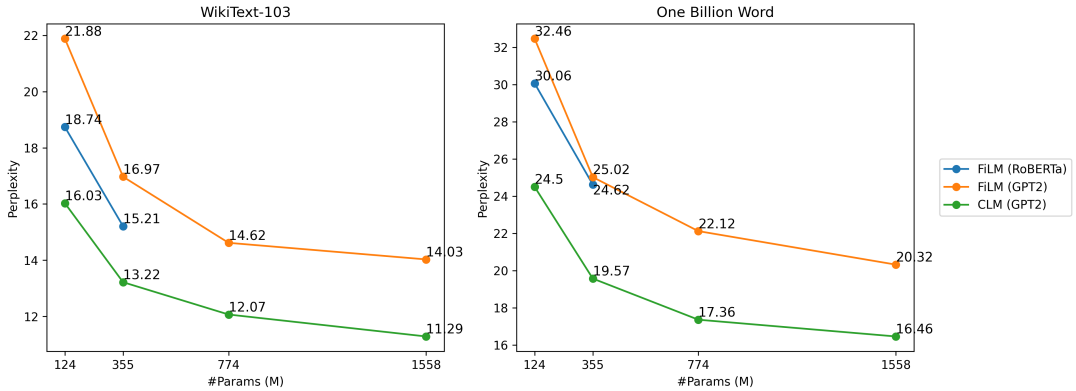

Figure 4: Perplexity of FiLM and CLM on WikiText-103 (Left) and One Billion Word (Right).

Table 3: Agreement percentages between GPT4-human and human-human evaluations on ROCStories. Each cell shows the percentage where the row head matches the first evaluator's preference and the column head the second's. **Bold** entries highlight agreements, and underlined entries mark opposite preferences for options 1 and 2. GPT4 and human evaluators reach consensus 49.5% of the time, while the agreement between humans is 62%.

GPT4\Human	1	2	Tie
1	**21.0**	8.5	11.5
2	6.0	**19.0**	10.0
Tie	7.0	7.5	**9.5**

Human\Human	1	2	Tie
1	**24.0**	3.0	8.0
2	5.0	**22.0**	14.0
Tie	4.0	4.0	**16.0**

1. CM	Depression, stress and stress increase the risk of alcoholism, gambling, excessive drinking and drug abuse.
2. FiLM	Depression, loneliness and stress increase the risk of, say, drug abuse.
GPT4	Option 1 is grammatically correct, but it contains redundancy by stating that stress increases the risk of excessive drinking and alcoholism, which essentially means the same thing. Option 2 is also grammatically correct, but includes 'say' which adds unnecessary hesitation to the sentence. Nonetheless, option 2 resolves the redundancy issue present in option 1, making it the better completion for the given sentence. For those reasons, the final decision is: <2>.
2. CM	In the same way as the Himalayas are " Asia's water-tower, " Switzerland is the source of Europe's biggest rivers, supporting agriculture and farming, and a world leader in nuclear power stations.
1. FiLM	In the same way as the Himalayas are " Asia's water-tower, " Switzerland is the source of Europe's biggest rivers, supporting agriculture and the construction of new nuclear power stations.
GPT4	Option 1 is superior in terms of grammatical fluency and coherence. It expresses that the water from Switzerland's rivers supports both agriculture and the construction of new nuclear power stations, which makes clear and logical sense. Option 2, by contrast, seems a bit disjointed. It indicates that the water supports agriculture, farming (a redundancy since farming is a type of agriculture), and then asserts that Switzerland is a world leader in nuclear power stations. The conjunction "and" makes it seem as if being a world leader in nuclear power stations is another thing that the water is supporting, which isn't logically coherent. The final decision is <1>.
1. CM	The situation has been further complicated by growing indications that the US Treasury is preparing to inject capital directly into banks.
2. FiLM	The situation has been further complicated by growing indications that the US Treasury is preparing to sell stakes in some of America's banks.
GPT4	Both options are grammatically correct. However, they suggest different outcomes. In option 1, the US treasury is preparing to support the banks by injecting capital directly into them. This implies a strategy of financially bolstering banks. In option 2, the US treasury is preparing to divest its stake in some banks, which implies a different strategy of reducing government involvement or exposure in the banking sector. Both fit the context and maintain the coherence of the sentence. Hence, the choice between the two hinges on the intended meaning rather than on grammar or coherence. Therefore, the answer is <->.
2. CM	WASHINGTON (Reuters) - The United States has been secretly detaining an Iraqi held by the U.S. military, a British legal charity charged on Monday, but the Pentagon described the report as inaccurate.
1. FiLM	WASHINGTON (Reuters) - The United States has been secretly detaining terror suspects for longer than previously reported, a leading British legal charity charged on Monday, but the Pentagon described the report as inaccurate.
GPT4	Option 1 seems to refer to the prolonged detention of multiple terror suspects, while option 2 specifies the detainment of a particular Iraqi individual by the U.S military. Both options could potentially fit depending on the broader context, but without further information, option 1 seems slightly more coherent as it aligns better with the criticism made by the British legal charity and the Pentagon's response to that criticism. Hence, I conclude with <1>.
1. CM	The Yankees needed three games to beat Cleveland in last year 's AL opener this year and then won Game 1 against the Indians.
2. FiLM	The Red Sox came from behind to beat the Tigers in the American League 's Series opener this year and then won Game 1 against the Indians.
GPT4	Both options are grammatically correct. However, they both present different scenarios. Choice 1 states that "The Yankees" needed three games to defeat Cleveland in the AL opener that happened last year, and then it leaps to discuss a victory in Game 1 against the Indians this year. This could potentially cause confusion due to the time jump in the sports events it discusses. Choice 2 says, "The Red Sox came from behind", implying that they played catch-up in a game to defeat the Tigers in the Series opener this year. They then, presumably in the same year, won a game against the Indians. This is more logically coherent and consistent in regard to its timeline. So, option 2 is the better choice. Final decision: <2>.
2. CM	The satellite weighed some 660 pounds, Israeli Haaretz newspaper reported, citing company officials.
1. FiLM	The satellite weighs some 660 pounds, the Israeli Haaretz newspaper reported, citing defense company officials.
GPT4	Both options are grammatically correct and coherent, however, the difference lies in the tense used. Option 1 uses the present tense 'weighs' suggesting the satellite still exists in its described state, while option 2 uses the past tense 'weighed' implying the satellite may no longer exist or have changed in some way. The best choice largely depends on the overall context and temporal setting of the sentence. Given that there's insufficient context to determine the most appropriate tense, both options are equally appropriate. For this reason, my final decision is <->.
2. CM	As North Korea's close neighbor, traditional ally and major provider of economic aid, China is widely believed to hold the key to solving the North Korea conundrum.
1. FiLM	As North Korea's close neighbor, traditional military ally and provider of economic aid, China is widely believed to hold the key to solving the North Korea conundrum.
GPT4	Both sentences are grammatically correct, and both retain the intended meaning of the sentence. However, option 1 specifies that China is a "military" ally while option 2 does not specify the type of ally but states it's a "major" provider of economic aid. Both are fitting and coherent with the context. However, depending on the specific context, it may be more accurate to emphasize the military alliance (option 1) or the significant economic aid (option 2). Therefore, without further context, I conclude my decision with <->.
2. CM	Puerto Rico's unemployment rate hit 14.8 % in January, according to the U.S. Bureau of Labor Statistics.
1. FiLM	Puerto Rico had a record low rate of employment in the last year, according to the Bureau of Labor Statistics.
GPT4	Both options are grammatically correct and fit sensibly into the context. However, option 2 provides a more detailed, specific statistic (14.8% unemployment in January) while option 1 gives a general statement that could be improved with more precise figures. Therefore, based on coherency and information value of the statement, option 2 is a better fit. Final decision: <2>.

Figure 5: Text infilling examples from One Billion Word, evaluated by GPT4. The provided context is in black, and the model generated text is in color. The number before each model denotes the option number presented to GPT4 for evaluation.

1. CM	= Kiss You (One Direction song) = " Kiss You " is a song recorded by English @-@ Irish boy band One Direction for their second studio album, Take Me Home (2012). It was released as the record's second single in Germany and the third overall single on 7 January 2013. The song was composed by Kristoffer Fogelmark, Kristian Lundin, Albin Nedler, Savan Kotecha, Shellback and its producers, Carl Falk and Rami Yacoub. " Kiss You " is an upbeat power pop song with electronic effects ; the lyrics detail a protagonist's infatuation with a significant other. Critics praised the song for its production, calling it a stand @-@ out track on Take Me Home. The track became the group's sixth top @-@ ten hit in Ireland and the United Kingdom, while attaining top @-@ forty positions in both Belgian territories (Flanders and Wallonia), as well as in Australia, Canada, Denmark, France, New Zealand, and the Netherlands. The single peaked at number 46 on the US Billboard Hot 100 and has been certified gold by the Recording Industry Association of America (RIAA) for shipments of 500 @,@ 000 copies. One Direction performed " Kiss You " on both the Vevo channel and during their 2013 Take Me Home Tour. In 2013, the song was performed on the first leg of their 2013 @-@ 2014 Take Me Home Tour. = = Background and release = = After releasing their debut album, Up All Night (2011), One Direction embarked on a promotional campaign for the album, including several music video releases and a series of live appearances. They released the lead single, " What Makes You Beautiful ", on 3 December 2011. A music video for " Kiss You ", directed by Anthony Mandler, was filmed for the single and its accompanying music video. The music video for " Kiss You " was released on YouTube on 5 December, and featured cameo appearances by Liam Payne, Leona Lewis, Alesha Dixon, and other music videos. The clip depicts the band shooting various scenes via a green screen, which include sequences reminiscent of iconic music videos, such as the ones for the Beatles'" You Never Give Me Your Money ", the dancing game Just Dance 2014, and is also one of the select songs available on the demo version. Additionally, it is the final main track on the US edition of Now That's What I Call Music! 46.
2. FiLM	= Kiss You (One Direction song) = " Kiss You " is a song recorded by English @-@ Irish boy band One Direction for their second studio album, Take Me Home (2012). It was released as the record's second single in Germany and the third overall single on 7 January 2013. The song was composed by Kristoffer Fogelmark, Kristian Lundin, Albin Nedler, Savan Kotecha, Shellback and its producers, Carl Falk and Rami Yacoub. " Kiss You " is an upbeat power pop song with electronic effects ; the lyrics detail a protagonist's infatuation with a significant other. Critics praised the song for its production, calling it a stand @-@ out track on Take Me Home. The track became the group's sixth top @-@ ten hit in Ireland and the United Kingdom, while attaining top @-@ forty positions in both Belgian territories (Flanders and Wallonia), as well as in Australia, Canada, Denmark, France, New Zealand, and the Netherlands. The single peaked at number 46 on the US Billboard Hot 100 and has been certified gold by the Recording Industry Association of America (RIAA) for shipments of 500 @,@ 000 copies. One Direction performed " Kiss You " on both the 2012 and 2013 versions of the 2012 MTV Video Music Awards and on the red carpet at the 2012 Royal Variety Performance. The group performed the track live on The X Factor in the United Kingdom. The accompanying music video was directed by Ed Bales and features a series of concept and extravagant filmmaking techniques, including film grain, motion blur and other filters reminiscent of other music videos. The clip depicts the band shooting various scenes via a green screen, which include sequences reminiscent of iconic music videos by Michael Jackson, Michael Jackson, Diana Ross and The Supremes. Upon its release, " Kiss You " charted in both the Irish and UK Singles Chart, reaching number one on the UK Singles Chart. The song has charted in the top twenty of Australia, Canada, the United States, Sweden and the United Kingdom. The song is featured on the soundtrack of the dancing game Just Dance 2014, and is also one of the select songs available on the demo version. Additionally, it is the final main track on the US edition of Now That's What I Call Music! 46.
GPT4	Both options provide a smooth transition that follows the flow of the given text. However, option 2 shows a better understanding of the song and its relevance at events such as the 2012 MTV Video Music Awards and the 2012 Royal Variety Performance. It also correctly mentioned that the music video was directed by Ed Bales. In contrast, option 1 references confusing elements, like appearances by Leona Lewis and Alesha Dixon without further explanation of their roles. The suggestion "the Beatles'" You Never Give Me Your Money " in option 1 also feels awkward and is not clear in its relation to the narrative. Therefore, my final decision is <2>.
2. CM	= = = Aircraft = = = Ise was briefly fitted with an aircraft flying deck during her reconstruction and was initially equipped with four floatplanes, but they were removed during the modernization and only one floatplane, a Nakajima E8N2 biplane, was installed in late 1936. The conversion was not completed until late 1937. The first Ise @-@ class battleships were fitted with the Type 3 floatplane MMF3 as their aircraft carrier escorts until the initial Nakajima E4N2 biplanes were replaced by Nakajima E8N2 biplanes in 1938. = = Conversion to aircraft carrier = = = The sinking of the British capital ships Prince of Wales and Repulse by Japanese land @-@ based aircraft on the night of 8 / 9 June 1942 and the loss of an estimated 39 aircraft to aircraft carrier aircraft during the Battle of the Eastern Solomons in May 1943 increased concerns about the ability of Japanese battleships to defeat enemy aircraft. The loss of several aircraft on 10 December 1941 led the IJN to realize that battleships could not operate in the face of enemy aircraft and required friendly air support to protect them.
1. FiLM	= = = Aircraft = = = Ise was briefly fitted with an aircraft flying @-@ off deck equipped with Nakajima EMF3 biplanes. At the start of the war, two of the ships embarked a total of eleven Mitsubishi F1M2 fighters. In 1933, the ship embarked a total of twelve Mitsubishi A2N2 biplanes. Two years later, the ship embarked a total of 24 Mitsubishi A1M3 fighters and nine Mitsubishi F1M2 biplanes. In 1935, the ship was fitted with three more fighters and three more in 1938 - 39. The initial Nakajima E4N2 biplanes were replaced by Nakajima E8N2 biplanes in 1938. = = Conversion to aircraft carrier = = The sinking of the British capital ships Prince of Wales and Repulse by Japanese land @-@ based aircraft on 10 December 1941 led the IJN to realize that battleships could not operate in the face of enemy aircraft and required friendly air support to protect them.
GPT4	The first option provides a more thorough and chronological history of the aircraft on the Ise, complete with the models of the biplanes and fighters and the timeline of when they were embarked. The mention of the Nakajima E4N2 biplanes being replaced by Nakajima E8N2 in 1938 also fits neatly with the earlier detail. In addition, the use of '@-@' instead of '/' in the word 'off-deck' and 'land-based' is more grammatically correct. However, in the second paragraph of option 1, the word 'aircraft carrier' is not enclosed by brackets correctly. The second option doesn't seem to provide a complete account of the history of aircraft on Ise, mentioning only a few floatplanes. It also refers to the change of aircraft in 1938, but less distinctly than option 1. The '@-@' symbol is improperly used as a substitute for a slash in phrases such as '8 / 9 June 1942'. The use of 'aircraft carrier' in brackets is semantically correct, but grammatically incorrect as it is not enclosed by brackets properly, similar to option 1. Given these considerations, option 1 is a clear candidate as it provides a more extensive history of aircraft on the Ise. Therefore, despite the bracket error which is a shared mistake in both options, I choose <1>.

Figure 6: Text infilling examples from WikiText-103, evaluated by GPT4. The provided context is in black, and the model generated text is in color. The number before each model denotes the option number presented to GPT4 for evaluation.

2. CM	I tried going to the park the other day. The weather seemed nice enough for a walk. I walked for a few hours. My eyes were watery and it was hard to breathe. My allergies were too bad and I had to go back home.
1. FiLM	I tried going to the park the other day. The weather seemed nice enough for a walk. However, when I got there I started to itch. My eyes were watery and it was hard to breathe. My allergies were too bad and I had to go back home.
GPT4	Option 1 introduces a sudden onset of the narrator's allergies, implicating that the park's environment may be causing it. This makes sense as it leads directly to the narrator's symptoms and decision to return home. Option 2 introduces a time lapse - the narrator's long walk - but this clashes with the sudden onset of the symptoms. Hence, option 1 maintains the logical flow of the story and makes the most sense in this context. Therefore, my final decision is <1>.
2. CM	Josh loved when his mom baked apple pie. She always made it the same way. So he decided this time he would sneak a piece before dinner. The eggs his mom used must have been bad though. Josh got sick.
1. FiLM	Josh loved when his mom baked apple pie. But he didn't like to eat it until it was ready. So he decided this time he would sneak a piece before dinner. The eggs his mom used must have been bad though. Josh got sick.
GPT4	Option 1 suggests that Josh doesn't usually eat the apple pie before dinner because he prefers it when it's properly done, hence he deviated from his norm which resulted in him getting sick. This setup lends more of a cause-and-effect relationship between him deciding to sneak a piece early and getting sick, keeping the connection between story events clearer. Option 2 points out Josh's mom's consistency in making the apple pie but doesn't adequately explain why he chose to sneak a piece before dinner this time. It also doesn't tie as effectively to the ending where Josh gets sick. In terms of maintaining a logical flow and coherence, Option 1 seems to do a better job. Final decision: <1>.
1. CM	Avery was married with children. She was tired of her boring life. One day, she decided to meet up with an old boyfriend from college. They had a great time together. Avery regretted what she did the next day.
2. FiLM	Avery was married with children. She was tired of her boring life. One day, she decided to meet up with an old boyfriend from college. They had a lot of fun and ended up spending all night hanging out. Avery regretted what she did the next day.
GPT4	Both options have similar themes and continue the story logically. Option 2 is slightly more detailed and specific about what occurred during the meet-up. It explicitly highlights the carefree time Avery and her old boyfriend had, which strongly reinforces her regret in the following sentence. Therefore, option 2 maintains the story's coherence slightly more than option 1. So, my final decision is <2>.
2. CM	Feliciano went olive picking with his grandmother. While they picked, she told him stories of his ancestors. Before he realized it, the sun was going down. He had a nice dinner and went to bed. Feliciano was happy about his nice day.
1. FiLM	Feliciano went olive picking with his grandmother. While they picked, she told him stories of his ancestors. Before he realized it, the sun was going down. He had a great time with his grandmother that day. Feliciano was happy about his nice day.
GPT4	Option 1 keeps the focus of the story on the olive picking day and the relationship between Feliciano and his grandmother, thus adding to the coherence. Option 2, on the other hand, introduces a new event (dinner and going to bed) which does not contribute to the main plot and seems somewhat unrelated. Therefore, option 1 is more fitting. Final decision: <1>.
1. CM	Kelly was playing her new Mario game. She was having a lot of fun. She was playing for so long without beating the level. Finally she beat the last level. Kelly was so happy to finally beat it.
2. FiLM	Kelly was playing her new Mario game. She was trying to beat a level. She was playing for so long without beating the level. Finally she beat the last level. Kelly was so happy to finally beat it.
GPT4	Both options seem to flow well with Kelly's actions throughout the story, filling in the missing part in a coherence manner. However, option 2 presents a more logical flow as it introduces the fact that Kelly was trying to beat a level. This ties in more directly with the remaining parts of the story where she is playing for so long without beating the level and finally manages to do it. Option 1 simply comments on her emotional state, which doesn't seem to link as directly to the following sentences. Therefore, option 2 seems to be the more fitting choice. Final decision: <2>.
1. CM	Oliver was getting married. He was worried that he would stutter during the vows. When the time came, he took a deep breath and began to speak. He stuttered, but his wife smiled and hugged him and he was okay. Oliver was so grateful for his wife's love.
2. FiLM	Oliver was getting married this weekend. He was worried that he would stutter during the vows. When the time came, he took a deep breath and began to speak. He stuttered, but his wife smiled and hugged him and he was okay. Oliver was so grateful for his wife's love.
GPT4	Both options are very similar, and both provide continuum and coherence for the rest of the story. The difference between them is slight: option 1 states that Oliver was getting married, while option 2 specifies that the wedding was taking place this weekend. Both are logical and fit into the story seamlessly. Thus, neither is more fitting than the other. Answer: <->.
1. CM	Brandon was given a 10 page research paper to write for class. He begun working on the paper on his 2003 Dell laptop. Brandon stayed up until 1 in the morning to try and finish his paper. Suddenly, his computer crashed and died. Brandon was so upset that he had to turn in his paper late.
2. FiLM	Brandon was given a 10 page research paper to write for class. He begun working on the paper on his 2003 Dell laptop. Brandon stayed up until 1 in the morning to try and finish his paper. Suddenly, his computer crashed and died. Brandon was devastated that he would never finish his paper.
GPT4	Option 1 suggests that Brandon had his paper completed, or nearly completed, but was unable to turn it in on time due to his computer crashing. This is a logical flow of events. Option 2 suggests that because of the computer crashing, Brandon was unable to finish his paper, but this does not necessarily mean he would never be able to finish it. Therefore, it seems less logical and coherent. Hence, option 1 is the most suitable choice. Therefore, the answer is <1>.

Figure 7: Story completion examples from ROCStories, evaluated by GPT4. The provided context is in black, and the model generated text is in color. The number before each model denotes the option number presented to GPT4 for evaluation.