Revisiting Parallel Context Windows: A Frustratingly Simple Alternative and Chain-of-Thought Deterioration

Kejuan Yang[*], Xiao Liu[*], Kaiwen Men, Aohan Zeng, Yuxiao Dong, Jie Tang
Tsinghua University
ykj22@mails.tsinghua.edu.cn, shawliu9@gmail.com

Abstract

We identify two crucial limitations in the evaluation of recent parallel-integrated method Parallel Context Windows (PCW) (Ratner et al., 2023), which extends the maximum context lengths of language models, e.g., 2048 for LLaMA, by harnessing window-wise attention and positional embedding techniques. We first show that a simple yet strong baseline, weighted sum ensemble, is missing for the in-context few-shot classification. Moreover, on more challenging Chain-of-Thought (CoT) reasoning (e.g., HotpotQA), PCW would present unexpected deterioration regarding question miscomprehension and false inference. Based on our findings, we suggest that the existing PCW design may not guarantee sufficient improvement and practicality in handling lengthy documents in real-world applications. More community efforts on enabling language models' long context understanding ability should be paid.

1 Introduction

Over the past few months, the field of Large Language Models (LLMs) (Brown et al., 2020; Chowdhery et al., 2022; Zhang et al., 2022; Scao et al., 2022; Zeng et al., 2022) has undergone a remarkable resurgence, primarily GPT-4, which has proved reasoning abilities akin to human, spanning a variety of professional fields from law to mathematics and physics (OpenAI, 2023). LLMs experience a paradigm shift, from individual tasks such as machine translation (Lopez, 2008), text summarization (Allahyari et al., 2017), and information extraction (Sarawagi et al., 2008), and gravitate toward a unified solution where users engage and interact in dialogues with chatbots to query anything.

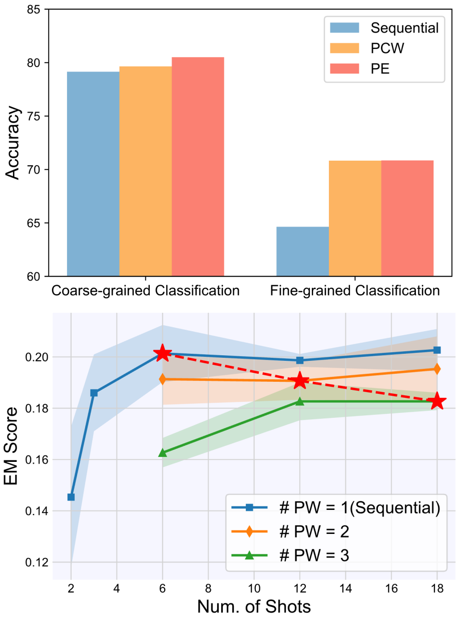


Figure 1: (a) PCW is comparable with Parallel Ensemble (PE) on both coarse-grained and fine-grained classification benchmarks; (b) PCW deteriorates closed-book HotpotQA. The red dashed line illustrates degradation in this challenging multi-hop reasoning task, despite doubling or tripling the number of demonstrations. An increased number of parallel windows (higher #PW) leads to sparser attention but worse accuracy, while a single window indicates the sequential baseline.

Still, a major challenge remains in LLMs — their abilities are constrained by their maximum context lengths. For example, GPT-3 (Brown et al., 2020) mentions its few demonstration samples in in-context learning (ICL) due to length limit. Recent Auto-GPT (Significant-Gravitas, 2023) is also observed to suffer from lengthy histories induced by CoT (Wei et al., 2022), which shepherds the LMs to mirror human cognition through a step-by-step progression of thinking and reflection to solve challenging reasoning missions. Hence it is vital to develop techniques to extend the context length of existing LLMs for reasoning.

[*]Corresponding to: Jie Tang(jietang@tsinghua.edu.cn) and Yuxiao Dong (yuxiaod@tsinghua.edu.cn)

[†]Kejuan and Xiao contributed equally.

[‡]Codes are available at https://github.com/kejuanyang1/Revisit_PCW.

Findings of the Association for Computational Linguistics: ACL 2024, pages 8841–8852
August 11-16, 2024 ©2024 Association for Computational Linguistics

Error Type	Sequential	Parallel
Reasoning Error	16.28%	34.09%
- False Reasoning	2.33%	10.23%
- Question Misinterpretation	10.47%	19.32%
- No CoT Reasoning	3.49%	4.55%
Non-reasoning Error	81.40%	59.09%
Other	2.33%	6.82%

Table 1: Analysis on closed-book HotpotQA errors. We classify them into five sub-categories and record their frequencies. PCW diminishes reasoning by more false reasoning, misinterpretation of the question, and even a complete lack of CoT reasoning.

A recent related attempt is PCW (Ratner et al., 2023), which brings the idea of parallel contexts to mitigate the length limitation problem in GPTs. PCW segments the text sequence into windows, constraining the attention to be visible within each window while all windows share the same positional embeddings. It reports improvements in few-shot ICL classification and generation tasks over the conventional sequential baseline, especially on fine-grained classification tasks with large label space such as BANKING77 (Casanueva et al., 2020) and CLINIC150 (Larson et al., 2019). By introducing over-length number of demonstration samples in one sequence, LMs can access more labels from context and thus outperform the sequential ICL where fewer samples could be seen.

However, in this work we identify limitations in PCW's evaluation, especially from two aspects:

- **Unequal Comparison**: As PCW sees more demonstrations, it is better to compare sequential methods receiving equal number of samples (e.g., ensembling multiple sequences) instead of a single sequence with fewer samples.
- **Unchallenging Tasks**: PCW evaluates on traditional classification and generation tasks only, but leaves untouched more challenging and practical problems in current LLMs concerning lengthy context of CoT reasoning.

Contributions. In light of the current limitations, we re-examine PCW's effectiveness in few-shot text classification against a fairer baseline and in more challenging CoT problems.

For text classification, we introduce a simple yet strong alternative—Parallel Ensemble (PE), which directly ensembles predictions from each context window as individual sequences, to achieve the same improvement as PCW, without modifying transformers and adding computation complexity (Cf. Figure 1). Results show that PE achieves comparable and even better average performance to PCW in evaluation. For more challenging missions, we follow ReAct (Yao et al., 2023) setting to evaluate pure CoT reasoning on closed-book HotpotQA. Unfortunately, PCW makes no improvement, and even deteriorates LMs CoT reasoning (Cf. Figure 1). Careful investigation unveils that PCW might weaken LMs' language reasoning, yielding issues including false inference, question misunderstanding, and absence of CoT (Cf. Figure 2).

In conclusion, our contributions are two-fold. Firstly, we propose that Parallel Ensemble, a direct weighted-sum ensemble on the logits of generated labels, is comparable to PCW on most classification benchmarks without any architecture modification. Secondly, we examine that PCW unintentionally results in a decline in LM's reasoning ability, raising questions about its practical benefit to current chat-based LLMs. We appeal to the community for more comprehensive study on the problem of LLMs' length extension challenge.

2 Preliminary

2.1 In-Context Learning

A language model ϕ is pre-trained to predict the conditional probability $p_\phi(\psi|C)$ where C represents the text input and ψ represents the word distribution over the given vocabulary.

In addition to the direct zero-shot inference, LMs also exhibit in-context learning capabilities where they tailor to corresponding tasks by seeing demonstrations(examples). In few-shot inference, C is extended into two parts: N-shot demonstrations $D = \{d_1, d_2, ..., d_N\}$ formatted as $d_i = \{input : x_i; output : y_i\}$, and the test input x_{test}. Conceptually, in-context learning equates to the text generation of $p_\phi(y_{test}|D, x_{test})$.

2.2 Sequential ICL

The language model reads context input $I = \{T, A, P\}$, which includes text tokens T, attention matrix A, and positional embedding P.

- Text tokens T: tokenized input text.
- Attention matrix A: a two-dimensional matrix that determines the visibility between input and output tokens—$A_{i,j} = 1$ suggests the j-th output token relates to the i-th input token, and $A_{i,j} = 0$ suggests no attention between them.
- Positional Embedding P: a sequence of IDs indicating the position for every text token.

Dataset	#Labels	LLaMA 7B			LLaMA 13B			LLaMA 33B		
		Seq	PCW	PE	Seq	PCW	PE	Seq	PCW	PE
RTE	2	72.2 (3.5)	**74.8** (2.1)	73.7 (2.7)	75.1 (2.2)	74.2 (1.5)	**75.8** (0.8)	79.9 (2.2)	79.0 (2.1)	**80.3** (1.6)
CB	3	82.6 (6.2)	84.6 (6.0)	**85.4** (3.4)	72.9 (8.9)	75.3 (9.4)	**76.4** (8.8)	87.8 (2.0)	87.3 (2.0)	**88.9** (1.3)
AGNews	4	87.0 (1.7)	87.4 (1.7)	**88.5** (1.0)	85.6 (2.6)	**87.8** (1.6)	87.7 (1.5)	88.1 (4.6)	89.9 (0.8)	**90.2** (0.6)
SST5	5	48.0 (3.8)	47.8 (4.9)	**48.6** (2.1)	49.9 (1.2)	**51.4** (0.9)	50.5 (1.2)	50.2 (1.6)	50.3 (1.3)	**51.2** (1.2)
TREC	6	83.3 (5.9)	83.6 (2.1)	**87.4** (1.2)	**83.5** (2.3)	82.2 (3.7)	82.3 (2.8)	86.6 (2.2)	86.1 (1.8)	**86.9** (1.2)
DBPedia	14	87.1 (6.4)	**95.2** (2.8)	93.6 (3.6)	88.9 (4.8)	**92.8** (4.2)	92.3 (4.6)	87.9 (7.1)	**94.9** (2.7)	94.7 (2.6)
NLU Scenario	18	79.7 (2.8)	82.0 (1.5)	**85.1** (1.3)	83.8 (2.1)	87.2 (1.1)	**87.6** (1.2)	83.9 (2.5)	86.8 (1.3)	**87.8** (0.9)
TREC Fine	50	53.9 (7.8)	53.8 (4.6)	**65.6** (4.0)	56.0 (6.7)	**63.6** (5.9)	**63.6** (4.8)	61.8 (6.4)	**68.9** (4.5)	68.6 (4.8)
NLU Intent	68	60.3 (3.5)	61.9 (2.9)	**69.2** (2.5)	66.9 (3.4)	73.7 (1.8)	**74.3** (2.1)	69.7 (4.0)	75.8 (2.3)	**77.4** (2.0)
BANKING77	77	41.4 (3.4)	48.0 (2.1)	**48.9** (1.4)	43.8 (2.9)	**55.2** (2.2)	**55.2** (2.2)	47.5 (3.1)	**61.3** (2.1)	57.3 (2.1)
CLINIC150	150	62.9 (2.3)	64.6 (2.2)	**66.0** (1.4)	66.9 (3.5)	**73.1** (1.3)	71.5 (1.9)	67.6 (3.1)	**75.0** (2.0)	72.4 (2.1)
AVG Gain		-	+2.30	+4.89	-	+3.90	+3.96	-	+4.01	+4.06

Table 2: Results on coarse-grained (#Labels $\leq$ 15) and fine-grained (#Labels >15) classification tasks utilizing three ICL methods: Sequential baseline, Parallel Context Window (PCW) (Ratner et al., 2023), and Parallel Ensemble (PE). We set the number of parallel windows to 3 as it is the best selection according to (Ratner et al., 2023).

Denote input token length $l = len(C)$. The standard sequential ICL input I_{seq} is formed as:

$$\begin{aligned} T_{\text{seq}} &= \{T(x_{\text{test}}), T(d_1), \cdots, T(d_N)\}, \\ A_{\text{seq}} &= [a_{ij}]_{l\times l} = \begin{cases} 0 & \text{for } 0 \leq j < i < l \\ 1 & \text{otherwise} \end{cases}, \\ P_{\text{seq}} &= \{0, 1, \cdots, l-1\}. \end{aligned} \tag{1}$$

2.3 Parallel ICL

Parallel ICL reconfigures two fundamental inputs of LMs: the attention matrix A and positional embedding P. All demonstrations D are segmented into separate windows $\{W_1, W_2, ..., W_\phi\}$ (Ratner et al., 2023), denoting the number of windows as ϕ, where $\phi = N$ is the most fine-grained division. The straightforward parallel approach is to block attention between demonstration windows, but allow the test input x_{test} to attend to every window. For positional embedding, we modify the test input to begin after the longest window's position p_{max}.

The input of Parallel ICL I_{prl} is formulated as:

$$\begin{aligned} T_{\text{prl}} &= T_{\text{seq}} = \{T(x_{\text{test}}), T(d_1), \cdots, T(d_N)\}, \\ A_{\text{prl}} &= [a_{ij}]_{l\times l} \\ &= \begin{cases} 0 & \text{for } 0 \leq j < i < l, \\ 0 & \text{between } W_m \text{ and } W_k, m \neq k \in [1, \phi] \\ 1 & \text{otherwise} \end{cases}, \\ P_{\text{prl}} &= \underbrace{\{0, 1, \cdots, p_{\text{max}}\}, \cdots, \{0, 1, \cdots, p_{\text{max}}\}}_{\phi \text{ times}}, \\ &\quad \{p_{\text{max}} + 1, \cdots, l-1\}. \end{aligned} \tag{2}$$

3 Experiments

3.1 Experiment Setup

Classification. We perform ICL evaluation on 11 classification datasets spread among diverse domains — SST5 (Socher et al., 2013), CB (Wang et al., 2019), RTE (Bentivogli et al., 2009), BANKING77 (Casanueva et al., 2020), NLU & NLU Scenario (Liu et al., 2019), CLINIC150 (Larson et al., 2019), AGNews (Zhang et al., 2015), DBPedia (Zhang et al., 2015), TREC & TREC Fine(Li and Roth, 2002). The selection of datasets follows PCW (Ratner et al., 2023). For prompt engineering, we follow PCW (Ratner et al., 2023) setting. See more details in Appendix A.3.

Reasoning. HotpotQA (Yang et al., 2018) is a challenging knowledge-intensive multi-hop reasoning task designed for complex reasoning scenarios. Unlike traditional QA tasks, HotpotQA requires LMs to not only locate relevant information from multiple Wikipedia documents but also to understand and connect these pieces of information in a logical and meaningful way. For instance, to answer the question "What movie starring Nicole Kidman won her an Academy Award", we will execute Hop 1: Identify the movies in which Nicole Kidman has acted, and then Hop 2: Determine which of these films led to Nicole Kidman winning an Academy Award. By synthesizing these two pieces of information from separate sources, we obtain the final answer "The Hour".

We aim for a more advanced setting to evaluate both the knowledge level and reasoning abil-

#Shots	LLaMA 7B			Vicuna 13B			LLaMA 33B		
	#PW = 1 (Sequential)	#PW = 2	#PW = 3	#PW = 1 (Sequential)	#PW = 2	#PW = 3	#PW = 1 (Sequential)	#PW = 2	#PW = 3
2	**14.5** (2.7)	0.1 (0.1)	-	**16.9** (3.8)	0.2 (0.2)	-	**28.6** (0.9)	0.3 (0.2)	-
3	**18.6** (1.5)	-	0.7 (0.8)	**23.0** (1.4)	-	3.3 (3.2)	**32.1** (1.2)	-	0.7 (0.3)
6	**20.1** (1.1)	19.1 (1.0)	16.3 (0.6)	**23.6** (0.5)	23.4 (1.6)	22.5 (0.8)	**33.2** (0.3)	32.1 (0.7)	30.5 (1.2)
12	**19.9** (0.3)	19.1 (0.7)	18.3 (0.7)	**24.1** (0.8)	23.1 (0.3)	22.8 (0.0)	**33.7** (0.4)	**33.7** (0.4)	32.9 (0.8)
18	**20.3** (0.8)	19.5 (1.3)	18.3 (0.3)	**24.6** (1.3)	24.1 (0.8)	22.8 (1.1)	**35.8** (0.4)	35.0 (0.3)	32.5 (0.3)

Table 3: CoT results on HotpotQA evaluated in Exact Match score. #PW denotes the number of parallel windows, higher PW means finer-grained windows, and #PW = 1 demonstrates the sequential baseline.

ity leveraging CoT as in ReAct (Yao et al., 2023), given that current LLaMAs (cf., Table 3) have already achieved performance comparable to PLMs ((Ratner et al., 2023), ranging from 20% to 30%), even when LLaMAs have no access to golden supporting paragraphs.

Adhering to the popular CoT evaluation (Yao et al., 2023; Wei et al., 2022), we manually crafted 18 multi-step thinking trajectories, as creating hundreds of high-quality demonstrations to reach the maximum token length of the language model(2048) is too expensive. See more details in Appendix A.3.

3.2 Result Analysis

PCW is Weighted Sum Ensemble for classification. As indicated in Table 2 (with complete results provided in Table 4), the strength of parallel-integrated methods is significant mostly in classification tasks featuring many labels, e.g., BANKING77, CLINIC150. To identify the underlying cause, we introduce another parallel method, Parallel Ensemble (PE), which directly applies a weighted sum after the test instance's label is predicted using each context window. The weights for each label candidate are determined by the logits of the newly generated tokens, averaged among the sequence. See detailed formulation in Appendix A.2.

We find PCW and PE have similar performances across most tasks, and sometimes PE even outperforms PCW, with a higher overall average gain among all LMs. This might suggest that PCW is simply doing a weighted sum ensemble among all the windows. But in larger models such as LLaMA 33B, we notice that PE slightly underperforms PCW in BANKING77 and CLINIC150, which hints at the potential strength of PCW in larger LMs with massive labels.

PCW deteriorates CoT Reasoning. We conducted experiments to explore how parallel windows influence the reasoning chain. HotpotQA, a knowledge-intensive multi-hop reasoning task known for its difficulty, even for models like GPT3.5 and PaLM 540B, merely achieves around 30% EM accuracy (Yao et al., 2023; Shinn et al., 2023). This makes it an ideal task to detect if language models' performance degrades throughout the reasoning chain. Here we encourage LMs to progressively solve problems utilizing their inherent knowledge through CoT, following (Yao et al., 2023) to minimize the noises induced by the accuracy and authenticity of provided or retrieved supporting paragraphs.

As illustrated in Table 3, we notice a significant gap between the Sequential baseline(# PW = 1) and PCW. When exposed to the same number of demonstrations, the raised number of windows implies sparser attention, resulting in worse performance because the repetitive positional embeddings might confuse the LM. Even when comparing 6-shots with 12- or 18-shots that offer double or triple the examples, the parallel method still falls short.

Further error analysis depicted in Figure 2 reveals that PCW easily misinterprets the basic logical relation between contexts, sometimes even disregards the question, and provides unrelated answers. None-reasoning error is mainly caused by hallucination, which is less relevant to the rationality of CoT reasoning. Other includes the generation of repetitive sentences or meaningless symbols.

4 Conclusion

We raise concerns about the use of parallel-integrated methods to address context length restriction: (1) PCW is functionally equal with a simple weighted sum ensemble on label distribution among context windows; (2) PCW degrades the multi-step reasoning capabilities of LLMs in complex tasks requiring knowledge understanding. Despite the fact that parallel-integrated methods show better classification performance when the label space is large, they merely brute-force ensemble each window's context, consequently weakening logical reasoning and knowledge comprehension.

Limitations

The limitations of our experimental considerations are as follows:

Firstly, we currently only evaluate language models under 50B parameters due to our computational constraints. A more comprehensive analysis should extend to larger models, such as LLaMA 65B, known for powerful understanding and CoT reasoning capabilities, and potentially some bidirectional language models (Du et al., 2022; Raffel et al., 2020).

Secondly, since LLaMA models employ rotary positional embedding, differing from the absolute positional embedding used by GPT2 in (Ratner et al., 2023), the enhancement brought by PCW may vary.

Thirdly, our experimental scope was restricted to knowledge-intensive tasks like HotpotQA and did not extend to mathematical tasks such as GSM8K (Cobbe et al., 2021), which necessitates multi-step reasoning to solve grade-school math word problems. We will include more CoT tasks in the next version of the evaluation.

Lastly, in our CoT experimental configuration, a limited number of examples are employed, due to the tedious efforts required to construct numerous demonstrations. This does not quite accord with the original PCW (Ratner et al., 2023) setting, where every window is populated with examples. Consequently, our observation might differ from theirs.

Therefore, the degradation phenomenon on reasoning tasks caused by parallel windows still requires further exploration and validation.

ACKNOWLEDGEMENTS

We would like to thank the suggestions and feedback from anonymous reviewers. This work was supported by Technology and Innovation Major Project of the Ministry of Science and Technology of China under Grant 2020AAA0108400, Natural Science Foundation of China (NSFC) 62276148 and 62425601, the New Cornerstone Science Foundation through the XPLORER PRIZE.

References

Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D Trippe, Juan B Gutierrez, and Krys Kochut. 2017. Text summarization techniques: a brief survey. *arXiv preprint arXiv:1707.02268*.

Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. 2009. The fifth pascal recognizing textual entailment challenge. In *TAC*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

Iñigo Casanueva, Tadas Temcinas, Daniela Gerz, Matthew Henderson, and Ivan Vulic. 2020. Efficient intent detection with dual sentence encoders. In *Proceedings of the 2nd Workshop on NLP for ConvAI - ACL 2020*. Data available at https://github.com/PolyAI-LDN/task-specific-datasets.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Stefan Larson, Anish Mahendran, Joseph J. Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael A. Laurenzano, Lingjia Tang, and Jason Mars. 2019. An evaluation dataset for intent classification and out-of-scope prediction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1311–1316, Hong Kong, China. Association for Computational Linguistics.

Xin Li and Dan Roth. 2002. Learning question classifiers. In *COLING 2002: The 19th International Conference on Computational Linguistics*.

Xingkun Liu, Arash Eshghi, Pawel Swietojanski, and Verena Rieser. 2019. Benchmarking natural language understanding services for building conversational agents. *ArXiv*, abs/1903.05566.

Adam Lopez. 2008. Statistical machine translation. *ACM Computing Surveys (CSUR)*, 40(3):1–49.

OpenAI. 2023. Gpt-4 technical report.

Ofir Press, Noah Smith, and Mike Lewis. 2022. Train short, test long: Attention with linear biases enables input length extrapolation. In *International Conference on Learning Representations*.

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.

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.

Sunita Sarawagi et al. 2008. Information extraction. *Foundations and Trends® in Databases*, 1(3):261–377.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Noah Shinn, Beck Labash, and Ashwin Gopinath. 2023. Reflexion: an autonomous agent with dynamic memory and self-reflection. *arXiv preprint arXiv:2303.11366*.

Significant-Gravitas. 2023. Auto-gpt. `https://github.com/charlespwd/project-title`.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

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

A Appendix

A.1 Prompts

A.1.1 Reasoning

We manually write 18 Chain-of-Thoughts demonstrations for the HotpotQA task including two subcategories — comparison and bridge. In bridge reasoning, the answer to the question requires making a connection between two or more pieces of information that are not directly related. The model needs to "bridge" the gap between these pieces of information in order to arrive at the correct answer. Comparison reasoning involves comparing two or more entities based on their attributes or related facts. This requires the model to understand and compare information from different facts. They are selected from the distractor test set while ensuring no overlap with the evaluation data pool. See Table 9 for details.

A.1.2 Classification

We strictly follow the prompting from (Ratner et al., 2023) in order to make a fair comparison. Therefore, we encourage a read of the original paper for details.

A.2 Parallel Ensemble

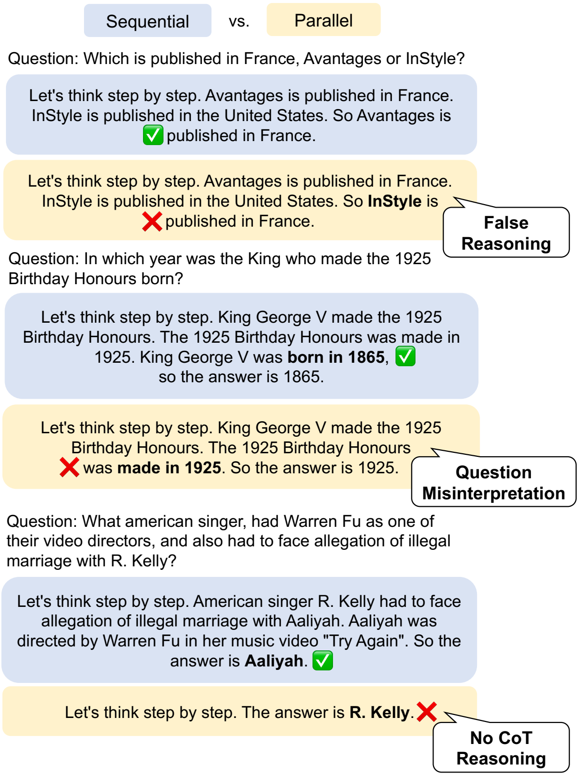
We introduce a simple but effective baseline, Parallel Ensemble(PE). The weighted sum for a specific classification label c is given by

$$P(c|x) = \sum_{i=1}^{\phi} w_i \cdot p_i(c|x) \tag{3}$$

where ϕ denotes the number of windows, and w_i denotes the logits of the newly generated tokens, averaged among the sequence.

A.3 Experiment Details

Language Models. We choose the LLaMA models including 7B, 13B, and 33B (Touvron et al., 2023) for evaluation due to their alignment with human preferences and strong ability to reason. Furthermore, we also test Vicuna 13B for reasoning. It is fine-tuned upon LLaMA 13B on user-shared conversations, which achieves nearly 90% quality of ChatGPT. While LLaMAs employ rotational positional embedding, they still accommodate parallel modifications and can potentially benefit from them, as handling longer texts results in degradation in models with relative positional embeddings (Press et al., 2022). We use LLaMAs and Vicuna 13B v1.1 checkpoint from HuggingFace for evaluation. Figure 1 shows Vicuna 13B results.



Question: Which is published in France, Avantages or InStyle?

Let's think step by step. Avantages is published in France. InStyle is published in the United States. So Avantages is ✅ published in France.

Let's think step by step. Avantages is published in France. InStyle is published in the United States. So **InStyle** is ❌ published in France.

Question: In which year was the King who made the 1925 Birthday Honours born?

Let's think step by step. King George V made the 1925 Birthday Honours. The 1925 Birthday Honours was made in 1925. King George V was **born in 1865**, ✅ so the answer is 1865.

Let's think step by step. King George V made the 1925 Birthday Honours. The 1925 Birthday Honours ❌ was **made in 1925**. So the answer is 1925.

Question: What american singer, had Warren Fu as one of their video directors, and also had to face allegation of illegal marriage with R. Kelly?

Let's think step by step. American singer R. Kelly had to face allegation of illegal marriage with Aaliyah. Aaliyah was directed by Warren Fu in her music video "Try Again". So the answer is **Aaliyah**. ✅

Let's think step by step. The answer is **R. Kelly**. ❌

Figure 2: Case study on closed-book HotpotQA CoT reasoning, where the sequential method succeeds but PCW fails in the reasoning due to reasons above.

Classification. We sample 10 times from the training set for classification tasks, limiting the maximum test samples to 1000. In the absence of a validation set, the test set is used. Our evaluation metric is multi-choice accuracy. We record the mean and variance for each seed run across all experimental results.

Reasoning. For the reasoning task, we sample from the manually designed demonstration pool with 3 seeds, restricting the size of the test samples to 500. The predictions are generated using greedy decoding at 0 temperature for reproducibility. We randomly select 100 samples to derive Table 1.

A.4 Supplementary Results

A.4.1 Ablation Study

We have observed that the Parallel ICL is significantly impacted by specific evaluation configurations. Consequently, we conducted an ablation study to gain a comprehensive understanding of these influences.

Dataset	**LLaMA 7B**			**LLaMA 13B**			**Vicuna 13B**			**LLaMA 33B**		
	Seq	**PCW**	**PE**	**Seq**	**PCW**	**PE**	**Seq**	**PCW**	**PE**	**Seq**	**PCW**	**PE**
RTE	72.2 (3.5)	**74.8** (2.1)	73.7 (2.7)	75.1 (2.2)	74.2 (1.5)	**75.8** (0.8)	80.7 (1.4)	78.1 (1.2)	**79.3** (1.5)	79.9 (2.2)	79.0 (2.1)	**80.3** (1.6)
CB	82.6 (6.2)	84.6 (6.0)	**85.4** (3.4)	72.9 (8.9)	75.3 (9.4)	**76.4** (8.8)	82.6 (2.2)	83.3 (3.0)	**85.3** (3.3)	87.8 (2.0)	87.3 (2.0)	**88.9** (1.3)
AGNews	87.0 (1.7)	87.4 (1.7)	**88.5** (1.0)	85.6 (2.6)	**87.8** (1.6)	87.7 (1.5)	84.8 (1.7)	86.2 (2.4)	**86.5** (1.5)	88.1 (4.6)	89.9 (0.8)	**90.2** (0.6)
SST5	48.0 (3.8)	47.8 (4.9)	**48.6** (2.1)	49.9 (1.2)	**51.4** (0.9)	50.5 (1.2)	48.1 (1.6)	50.7 (1.5)	50.7 (1.5)	50.2 (1.6)	50.3 (1.3)	**51.2** (1.2)
TREC	83.3 (5.9)	83.6 (2.1)	**87.4** (1.2)	**83.5** (2.3)	82.2 (3.7)	82.3 (2.8)	83.6 (3.0)	82.5 (4.0)	**84.6** (3.3)	86.6 (2.2)	86.1 (1.8)	**86.9** (1.2)
DBPedia	87.1 (6.4)	**95.2** (2.8)	93.6 (3.6)	88.9 (4.8)	**92.8** (4.2)	92.3 (4.6)	95.2 (2.0)	**97.2** (2.4)	96.6 (1.5)	87.9 (7.1)	**94.9** (2.7)	94.7 (2.6)
NLU Scenario	79.7 (2.8)	82.0 (1.5)	**85.1** (1.3)	83.8 (2.1)	87.2 (1.1)	**87.6** (1.2)	80.5 (2.5)	**85.6** (1.1)	83.9 (1.7)	83.9 (2.5)	86.8 (1.3)	**87.8** (0.9)
TREC Fine	53.9 (7.8)	53.8 (4.6)	**65.6** (4.0)	56.0 (6.7)	**63.6** (5.9)	**63.6** (4.8)	60.0 (5.3)	65.0 (3.9)	**67.6** (3.5)	61.8 (6.4)	**68.9** (4.5)	68.6 (4.8)
NLU Intent	60.3 (3.5)	61.9 (2.9)	**69.2** (2.5)	66.9 (3.4)	73.7 (1.8)	**74.3** (2.1)	69.2 (2.8)	74.6 (1.5)	**75.1** (1.7)	69.7 (4.0)	75.8 (2.3)	**77.4** (2.0)
BANKING77	41.4 (3.4)	48.0 (2.1)	**48.9** (1.4)	43.8 (2.9)	**55.2** (2.2)	**55.2** (2.2)	47.4 (2.5)	56.4 (1.6)	**56.9** (1.6)	47.5 (3.2)	**61.3** (2.1)	57.3 (2.1)
CLINIC150	62.9 (2.3)	64.6 (2.2)	**66.0** (1.4)	66.9 (3.5)	**73.1** (1.3)	71.5 (1.9)	66.1 (2.3)	**72.5** (1.9)	70.8 (2.0)	67.6 (3.1)	**75.0** (2.0)	72.4 (2.1)
AVG Gain	-	+2.30	+4.89	-	+3.90	+3.96	-	+3.09	+3.56	-	+4.01	+4.06

Table 4: Complete results on classification tasks utilizing three ICL methods: Sequential baseline, Parallel Context Window (PCW) (Ratner et al., 2023), and Parallel Ensemble (PE).

	Seq		**PCW**		**PE**	
	w/o blank	w/ blank	w/o blank	w/ blank	w/o blank	w/ blank
RTE	72.2 (±3.5)	72.5 (±3.3)	74.8 (±2.1)	70.0 (±4.5)	73.7 (±2.7)	69.8 (±4.6)
CB	82.6 (±6.2)	75.6 (±10.8)	84.6 (±6.0)	70.7 (±14.1)	85.4 (±3.4)	70.9 (±13.8)
AGNews	87.0 (±1.7)	86.9 (±1.8)	87.4 (±1.7)	88.0 (±0.7)	88.5 (±1.0)	88.0 (±0.7)
SST5	48.0 (±3.8)	47.8 (±1.0)	47.8 (±4.9)	47.6 (±2.0)	48.6 (±2.1)	47.6 (±1.9)
TREC	83.3 (±5.9)	82.5 (±2.6)	83.6 (±2.1)	73.7 (±6.1)	87.4 (±1.2)	73.4 (±5.9)
DBPedia	87.1 (±6.4)	90.4 (±4.7)	95.2 (±2.8)	93.8 (±1.9)	93.6 (±3.6)	93.8 (±1.8)
NLU Scenario	79.7 (±2.8)	79.9 (±2.8)	82.0 (±1.5)	82.5 (±1.8)	85.1 (±1.3)	82.5 (±1.8)
TREC Fine	53.9 (±7.8)	52.6 (±10.4)	53.8 (±4.6)	44.5 (±8.6)	65.6 (±4.0)	43.4 (±8.4)
NLU Intent	60.3 (±3.5)	60.3 (±2.9)	61.9 (±2.9)	59.0 (±3.6)	69.2 (±2.5)	59.0 (±3.6)
BANKING77	41.4 (±3.4)	41.8 (±2.4)	48.0 (±2.1)	46.7 (±2.2)	48.9 (±1.4)	46.6 (±2.1)
CLINIC150	62.9 (±2.3)	62.0 (±1.6)	64.6 (±2.2)	58.0 (±1.9)	66.0 (±1.4)	58.0 (±1.9)
AVG Δ		-0.54		-4.46		-7.18

Table 5: Ablation study on the existence of blank space in front of the label text. The experiments are performed using LLaMA 7B.

	Seq		**PCW**		**PE**	
	INT8	FP16	INT8	FP16	INT8	FP16
RTE	72.2 (±3.5)	73.8 (±1.7)	74.8 (±2.1)	74.6 (±1.9)	73.7 (±2.7)	74.6 (±1.9)
CB	82.6 (±6.2)	82.4 (±5.6)	84.6 (±6.0)	84.0 (±6.3)	85.4 (±3.4)	84.0 (±2.9)
AGNews	87.0 (±1.7)	87.0 (±1.6)	87.4 (±1.7)	87.6 (±1.5)	88.5 (±1.0)	88.3 (±0.9)
SST5	48.0 (±3.8)	47.6 (±4.0)	47.8 (±4.9)	47.2 (±5.6)	48.6 (±2.1)	48.8 (±2.0)
TREC	83.3 (±5.9)	83.8 (±6.8)	83.6 (±2.1)	83.1 (±2.9)	87.4 (±1.2)	86.8 (±1.9)
DBPedia	87.1 (±6.4)	87.0 (±6.1)	95.2 (±2.8)	95.5 (±2.6)	93.6 (±3.6)	93.9 (±3.4)
NLU Scenario	79.7 (±2.8)	80.1 (±2.6)	82.0 (±1.5)	82.6 (±1.5)	85.1 (±1.3)	85.7 (±1.4)
TREC Fine	53.9 (±7.8)	55.5 (±8.1)	53.8 (±4.6)	57.2 (±5.2)	65.6 (±4.0)	66.7 (±4.1)
NLU Intent	60.3 (±3.5)	61.4 (±3.2)	61.9 (±2.9)	63.4 (±3.4)	69.2 (±2.5)	69.8 (±2.6)
BANKING77	41.4 (±3.4)	42.0 (±3.2)	48.0 (±2.1)	49.6 (±2.0)	48.9 (±1.4)	50.0 (±1.4)
CLINIC150	62.9 (±2.3)	62.8 (±2.1)	64.6 (±2.2)	64.9 (±2.8)	66.0 (±1.4)	66.5 (±1.3)
AVG Δ		0.45		0.56		0.29

Table 6: Ablation study on the quantization of LM, i.e., INT8 or FP16 precision. The experiments are performed using LLaMA 7B.

Firstly, we discovered that whether or not a blank space is added before the label text in LLaMA classification tasks significantly influences the performance of the parallel methods due to the use of SentencePiece tokenizer. As shown in Table 5, while sequential ICL does not degrade much(-0.54%), notable declines are observed in the PCW and PE methods, with decreases of 4.46% and 7.18%, respectively. Therefore, we choose to delete the blank space for LLaMA classification.

Additionally, we also explored whether the quantization of the LM would impact the performance of the parallel methods. Comparisons were made between results derived from FP16 and INT8 quantization, as shown in Table 6. The results suggest that the discrepancy is relatively insignificant, confined within a range of 0.6%. Hence, to reduce the budget, we conduct our experiments using INT8 quantization for LLaMA 7B. Yet, for a precise evaluation of larger models, i.e., LLaMA 13B and 33B, we opt for FP16 quantization.

A.4.2 Case Study Analysis

Taking advantage of overlapping positional IDs, the parallel-integrated method is designed to concentrate on the cross-relation between windows. Based on our experiments in 3.2, PCW is functionally equal to a simple weighted-sum ensemble in classification tasks. But the drawback is that PCW tends to overlook semantic logic encapsulated within each example, especially how to think step-by-step and arrive at a correct answer.

Our case study (Fig. 2) found that this drawback of poor thinking includes three issues - false reasoning, question misinterpretation, and no CoT reasoning. False reasoning fails to make causal inferences between generated sentences, and question misinterpretation arises when there is a disconnection between the question and newly generated thoughts. We attribute these errors to the repetitive use of positional IDs, which confuses LLM's inference ability, leading to misinterpretations and inconsistent reasoning patterns.

A.4.3 PCW Single

We evaluate the most fine-grained parallel window method, i.e., PCW Single, where the window span is 1. We find that under such conditions, the parallel method drastically declines due to excessive repetition of positional embeddings in context windows, as shown in Table 7. We choose n_{max} for each dataset to be the shot number that fills in the maximum token length of LMs, i.e., 2048 for Vicuna. We set the window size as 3 to align with the main results in Section 3.

Method	Seq	PCW	PCW Single
# shots	n_{max}	$3 * n_{max}$	n_{max}
RTE	72.2 (±3.5)	74.8 (±2.1)	56.2 (±2.1)
CB	82.6 (±6.2)	84.6 (±6.0)	61.8 (±8.3)
AGNews	87.0 (±1.7)	87.4 (±1.7)	66.1 (±18.0)
SST5	48.0 (±3.8)	47.8 (±4.9)	26.0 (±3.2)
TREC	83.3 (±5.9)	83.6 (±2.1)	12.6 (±0.6)
DBPedia	87.1 (±6.4)	95.2 (±2.8)	82.2 (±15.1)
NLU Scenario	79.7 (±2.8)	82.0 (±1.5)	4.8 (±0.0)
TREC Fine	53.9 (±7.8)	53.8 (±4.6)	10.8 (±0.4)
NLU Intent	60.3 (±3.5)	61.9 (±2.9)	0.4 (±0.2)
BANKING77	41.4 (±3.4)	48.0 (±2.1)	2.2 (±0.7)
CLINIC150	62.9 (±2.3)	64.6 (±2.2)	0.6 (±0.0)

Table 7: Supplementary PCW results on ICL classification tasks for LLaMA 7B.

It is evident that as the number of parallel windows increases, there is a dramatic drop in In-Context Learning performance. This decline is especially notable in datasets such as BANKING77 and CLINIC150, which contain more than 50 labels. This is because of a prediction bias favoring one certain label. Above results demonstrate the negative effects of repeated positional embeddings for language models.

Method	Seq	PCW
RTE	64.2 (±6.2)	54.0 (±3.0)
CB	75.4 (±7.9)	64.7 (±12.0)
AGNews	62.2 (±9.2)	56.2 (±8.2)
SST5	43.1 (±1.8)	45.9 (±1.1)
TREC	44.3 (±3.2)	48.5 (±5.0)
DBPedia	82.0 (±3.5)	81.9 (±2.9)
NLU Scenario	55.4 (±4.4)	66.4 (±2.8)
TREC Fine	35.5 (±6.8)	37.2 (±4.5)
NLU Intent	49.8 (±3.9)	57.1 (±2.7)
BANKING77	14.4 (±2.3)	13.6 (±2.4)
CLINIC150	27.3 (±3.1)	28.8 (±5.2)

Table 8: PCW results on ICL classification tasks for ChatGLM 6B.

A.4.4 Bi-directional Models

We test the efficacy of parallel ICL on ChatGLM 6B, a bi-directional language model developed within GLM (Zeng et al., 2022) framework. ChatGLM 6B has been trained in Chinese and English contexts, demonstrating proficiency in both languages.

As illustrated in Table 8, we have noted that the few-shot performance of ChatGLM is less proficient when compared to LLaMA. This trend indicates a decline when a substantially larger number of examples are included in the sequential ICL. Also, it is reasonable to find out that the PCW method maintains its improvement in most of the fine-grained classification tasks.

Table 9: CoT demonstrations for HotpotQA

Question	What is the elevation range for the area that the eastern sector of the Colorado orogeny extends into?
Thought	The eastern sector of Colorado orogeny extends into the High Plains. High Plains rise in elevation from around 1,800 to 7,000 ft, so the answer is 1,800 to 7,000 ft.
Answer	1,800 to 7,000 ft
Question	Musician and satirist Allie Goertz wrote a song about the "The Simpsons" character Milhouse, who Matt Groening named after who?
Thought	The character Milhouse was named after U.S. president Richard Nixon, so the answer is Richard Nixon.
Answer	Richard Nixon
Question	Which documentary is about Finnish rock groups, Adam Clayton Powell or The Saimaa Gesture?
Thought	Clayton Powell (film) is a documentary about an African-American politician, not Finnish rock groups. So the documentary about Finnish rock groups must instead be The Saimaa Gesture.
Answer	The Saimaa Gesture
Question	What profession does Nicholas Ray and Elia Kazan have in common?
Thought	Professions of Nicholas Ray are director, screenwriter, and actor. Professions of Elia Kazan are director, producer, screenwriter, and actor. So profession Nicholas Ray and Elia Kazan have in common is director, screenwriter, and actor.
Answer	director, screenwriter, actor
Question	Which magazine was started first Arthur’s Magazine or First for Women?
Thought	Arthur’s Magazine was started in 1844. First for Women was started in 1989. 1844 (Arthur’s Magazine) < 1989 (First for Women), so Arthur’s Magazine was started first.
Answer	Arthur’s Magazine
Question	Were Pavel Urysohn and Leonid Levin known for the same type of work?
Thought	Pavel Urysohn is a mathematician. Leonid Levin is a mathematician and computer scientist. So Pavel Urysohn and Leonid Levin have the same type of work.
Answer	Yes
Question	Which of Jonny Craig and Pete Doherty has been a member of more bands?
Thought	Jonny Craig has been a member of four bands, Dance Gavin Dance, Emarosa, Rain City Drive, Isles, and Glacierswhile. Pete Doherty has been a member of three bands, Babyshambles, Peter Doherty and the Puta Madres. So Jonny Craig has been a member of more bands.
Answer	Jonny Craig
Question	If You Ever Get Lonely was covered by what Lyric Street Records-affiliated band?
Thought	If You Ever Get Lonely was covered by American country music duo Love and Theft, which is a Lyric Street Records-affiliated band, so the answer is Love and Theft.
Answer	Love and Theft
Question	Jaclyn Stapp is married to the former frontman of a band that disbanded in what year?
Thought	Jaclyn Stapp is married to Scott Stapp, the voice of the band Creed. Creed was an American rock band from Tallahassee, Florida, active from 1994 to 2004. So Creed disbanded in 2004.
Answer	2004

Question	The W. H. Shipman House is in what Hawaii county?
Thought	W. H. Shipman House is located at 141 Kaiulani Street, Hilo. Hawaii County, Hawaii is the county that Hilo is in, so the answer is Hawaii County.
Answer	Hawaii County
Question	The trophy given to the winner of the college football game between University of Idaho Vandals and University of Montana Grizzlies is derived from which European culture?
Thought	The trophy given to the winner of the college football game between University of Idaho Vandals and University of Montana Grizzlies is the Little Brown Stein. The Little Brown Stein is a large stein mug, which is derived from German culture.
Answer	German
Question	In what year was the singer who popularized "I'll Keep the Lovelight Burning" born?
Thought	"I'll Keep the Lovelight Burning (In My Heart)" is a popular song written by Bennie Benjamin and George David Weiss, popularized by Patti Page in 1949. Patti Page was born in 1927, so the answer is 1927.
Answer	1927
Question	Stronger Together was used for the campaign comprised of the Vice Presidential candidate who was a Senator from what state?
Thought	Tim Kaine was a Senator from Virginia who used Stronger Together for the campaign, so the answer is Virginia.
Answer	Virginia
Question	Which facility was founded in Missouri, Discovery Zone or Valentino's?
Thought	Valentino's was founded in Lincoln, Nebraska in 1957. 1989 (Discovery Zone) < 1957 (Valentino's), so Discovery Zone was founded in Missouri first.
Answer	Discovery Zone
Question	Kam Heskin plays Paige Morgan in a 2004 film directed by who?
Thought	Kam Heskin played Paige Morgan in The Prince and Me film franchise. The Prince and Me film franchise was directed by Martha Coolidge in 2004. So the answer is Martha Coolidge.
Answer	Martha Coolidge
Question	Otterington railway station was on a 393 mi major railway that linked Edinburgh to what city?
Thought	Otterington railway station was on the East Coast Main Line. The East Coast Main Line linked Edinburgh to London King's Cross station. So the answer is London.
Answer	London
Question	Who wrote the book that inspired the name of the CD/DVD "Here at the End of All Things"?
Thought	The title comes from a quote from the book The Lord of the Rings: The Return of the King. The Lord of the Rings was written by J. R. R. Tolkien, so the answer is J. R. R. Tolkien.
Answer	J. R. R. Tolkien
Question	Are both Lygodium or Maxillaria a genus of orchids?
Thought	Maxillaria is a genus of orchids. Lygodium is not a genus of orchids. So the answer is No.
Answer	No