Naive Bayes-based Context Extension for Large Language Models

Anonymous ACL submission

Abstract

Large Language Models (LLMs) have shown promising in-context learning abilities. However, conventional in-context learning approaches are often impeded by length limitations of transformer architecture, which pose challenges when attempting to effectively integrate supervision from a substantial number of demonstration examples. In this paper, we introduce a novel framework, called Naive Bayes-based Context Extension (NBCE), to enable existing LLMs to significantly expand their context size without the necessity for fine-tuning or reliance on specific model architectures, while maintaining linear efficiency. NBCE initially splits the context into equal-sized windows fitting the target LLM's maximum length. Then, it introduces a voting mechanism to select the most relevant window, regarded as the posterior context. Finally, it employs Bayes' theorem to generate the test task. Our experimental results demonstrate that NBCE substantially enhances performance, particularly as the number of demonstration examples increases, consistently outperforming alternative methods. The NBCE code will be made publicly accessible.

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities in in-context learning (ICL), a paradigm that enables them to excel in various unseen tasks based on task examples or instructions within their context (Han et al., 2021; Qiu et al., 2020). Unlike traditional fine-tuning methods, ICL leverages LLMs for downstream tasks solely through inference, eliminating the need for parameter updates and making it computationally efficient, bringing us closer to the goal of general AI. This approach has gained prominence as LLMs continue to grow in scale (Brown et al., 2020; Zhang et al., 2022a; Chowdhery et al., 2022).

The 2048-token context limit in popular LLMs like GPT-3 poses challenges for scaling up ICL with more demonstration examples in ICL, due to architectural constraints and computational complexity. Recent studies (Garg et al., 2022; Min et al., 2022b; Chen et al., 2022) improve ICL through meta-learning and fine-tuning on downstream tasks, but the limited diversity of annotated tasks and biases hinder generalization. Another line of research has explored various approaches to retraining long-range language models with extrapolation, extending them to 128 times the limit of existing LLMs (Li et al., 2023; Gu et al., 2023). However, these approaches require additional training over several steps, which can be time-consuming. Recently, Ratner et al. (2023) have introduced the concept of structured prompting that encodes demonstration examples individually with a designed position embeddings. Building upon this concept, Hao et al. (2022) proposed a mechanism in which these examples are collectively attended to by the test example through a scaled attention mechanism. Addressing this issue is crucial for leveraging ICL effectively, especially in scenarios with ample examples.

In this paper, we introduce a novel framework called Naive Bayes-based Context Extension (NBCE) for large language models to significantly expand the number of demonstrations by orders of magnitude while greatly enhancing stability. Instead of simply merging all demonstrations, we partition the vast number of demonstrations into multiple groups, each independently processed by the language model. This approach ensures that the encoding complexity scales linearly with the number of groups, avoiding the quadratic complexity associated with considering all examples simultaneously. Following Ratner et al. (2023); Hao et al. (2022), we align the position embeddings of grouped prompts to the right, placing them next to the test input. Subsequently, we leverage the Naive Bayes to encode the input by conditioning it on these grouped prompts. We conducted experiments across various tasks, including text classification, multi-choice, and open-ended tasks. NBCE effectively scales up the number of demonstrations, outperforming conventional in-context learning across different model sizes and tasks, while also significantly enhancing stability.

In brief, the contributions can be summarized as follows:

1. We introduce an innovative framework known as Naive Bayes-based Context Extension (NBCE), designed to substantially increase the volume of demonstrations for large language models, thus enhancing stability on a significant scale.

2. We provide detailed technical insights to enable context expending of in-context learning tasks. The idea is to encode the test sample by conditioning it on a vast array of demonstrations sourced from the training dataset.

3. We conducted extensive experiments on benchmark NLP datasets, and our findings clearly highlight NBCE's remarkable capability to efficiently scale up the number of demonstrations, while significantly enhancing overall stability.

2 Related Work

2.1 In-Context Learning

In recent years, in-context learning has received significant attention in the research community. Brown et al. (2020) introduced this concept, sparking a wave of investigations. Zhao et al. (2021); Han et al. (2023) addressed the issue of LLM miscalibrations and explored various calibration methods. However, few-shot performance can vary based on the order of demonstrations and template choices (Lu et al., 2022). In this context, Zhao et al. (2021) identified three biases and suggested content-free output calibration. Min et al. (2022a) demonstrated how these biases shift decision boundaries and proposed calibrating through prototypical cluster distribution estimation. Others focused on prompt engineering, such as selecting optimal demonstration permutations (Lu et al., 2022) and using retrieval modules for semantically similar in-context examples (Liu et al., 2022; Rubin et al., 2022). One promising direction is to improve in-context learning by increasing the number of demonstrations.

2.2 Context Extension

Expanding the contextual capabilities of LLM continues to pose a formidable challenge and has attracted considerable research attention. Various studies have introduced to tackle the memory limitations associated with self-attention mechanisms. These approaches can be broadly classified into two categories: fine-tuned approaches and few-shot approaches. Zaheer et al. (2020); Guo et al. (2022), have suggested using sparse attention as a solution to this issue. Press et al. (2022) took a novel approach by incorporating positional information using relative factors in attention weights instead of relying on absolute positional encoding. Despite the impressive capabilities of Press et al. (2022)'s model for extrapolation, it remains computationally intensive due to its quadratic self-attention cost, making it slow and resource-demanding for longer prompts. Ivgi et al. (2022) introduced an alternative approach called SLED, which is an encoder-decoder model specifically designed for handling lengthy texts. This model encodes short overlapping segments of input text and integrates this information within the decoder, similar to the Fusion-in-Decoder concept by Izacard and Grave (2021). However, these researches require additional training.

More recently, Ratner et al. (2023) have introduced the concept of Parallel Context Windows (PCW), which enables the concurrent utilization of multiple context windows without requiring additional training. PCW has been purposefully tailored for self-attention models, involving modifications to both position encoding and attention mask mechanisms to enhance the performance. NBCE and PCW share noteworthy similarities, as they both treat contexts as unordered and apply equal weighting. Notably, when NBCE is employed within the context of a single-layer, single-head attention model, the resulting outcomes closely approximate those achieved through the utilization of PCW. To substantiate this claim, we can formulate the language model tailored to a single-layer, single-head attention configuration.

$$p(x_t|x_{<t}) = \text{softmax}\left(\sum_{i=1}^{t} a_{t,i}v_iW\right) \quad (1)$$

hence, approximately: $\log p(x_t|x_{<t}) \sim \sum_{i=1}^{t} a_{t,i}v_iW$. Substituting this into Equation 11 and setting $\beta = 0$, we obtain:

$$\log p(T|S_1, S_2, \ldots, S_n) \sim \frac{1}{n}\sum_{k=1}^{n}\left(\sum_{i\in S_k} a_{T,i}v_i\right)W$$
$$= \left(\sum_{i\in S_1\oplus\ldots\oplus S_n} \frac{a_{T,i}}{n}v_i\right)W \quad (2)$$

here, we assume T represents a single sequence (i.e., the query), However, this assumption does not lack generality. The symbol $\oplus$ denotes concatenation and $S_k \oplus T$ is used for reasoning as a continuous segment (as per NBCE's setup), so their positional encodings are adjacent. Additionally, $a_{T,i}/n$ forms a collective attention for T with all S_i (with a sum equal to 1). These characteristics are consistent with PCW, which is essentially integrated into each layer more elegantly through an attention mask. Therefore, PCW can be thought of as a version of NBCE that utilizes average pooling.

3 Approach

An example of our proposed NBCE is depicted in 1. Assume that we have a sequence, denoted as T, which we intend to generate. Furthermore, we have multiple relatively independent context sets, denoted as $S_1, S_2, \ldots, S_n$ (e.g., n different paragraphs), each of which is sufficiently long and does not split a sentence into fragments. Suppose that the total length of these context sets exceeds the training length, but when combined with an individual S_k and T, they still fall within the training length. Our objective is to generate T based on the information contained in $S_1, S_2, \ldots, S_n$. In essence, we seek to estimate the conditional probability of T given $S_1, S_2, \ldots, S_n$, which can be represented as $p(T|S_1, S_2, \ldots, S_n)$.

In straightforward terms, Naive Bayes can be understood as a combination of two key elements: Bayes' formula and an independence assumption:

$$p(T|S_1, S_2, \ldots, S_n) \propto p(S_1, S_2, \ldots, S_n|T)p(T), \quad (3)$$

where, the symbol $\propto$ denotes proportionality, signifying that we are focusing solely on the relevant factors in a proportion while disregarding constant factors unrelated to the token sequence T. This approach aligns with the underlying assumption of conditional independence:

$$p(S_1, S_2, \ldots, S_n|T) = \prod_{k=1}^{n} p(S_k|T). \quad (4)$$

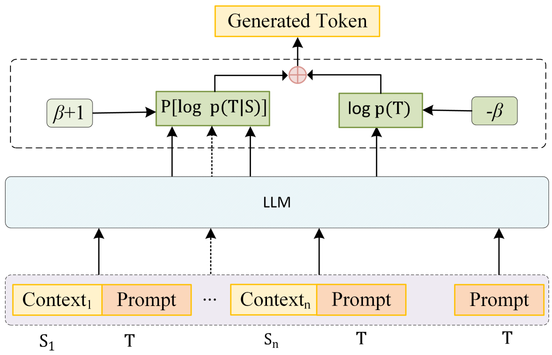

Figure 1: An example for our NBCE. Initially, NBCE divides the context into equal-sized windows, each with the maximum length compatible with LLM in-target. Subsequently, a voting mechanism is introduced to select the most relevant context window, regarded as the posterior context. Finally, it employs Bayes' theorem to generate the test task.

Thus, we have:

$$p(T|S_1, S_2, \ldots, S_n) \propto p(T)\prod_{k=1}^{n} p(S_k|T). \quad (5)$$

Furthermore, based on Bayes' formula $p(S_k|T) \propto \frac{p(T|S_k)}{p(T)}$, we get:

$$p(T|S_1, S_2, \ldots, S_n) \propto \frac{1}{p^{n-1}(T)}\prod_{k=1}^{n} p(T|S_k). \quad (6)$$

Or:

$$\log p(T|S_1, S_2, \ldots, S_n) = \sum_{k=1}^{n} \log p(T|S_k) - (n-1)\log p(T) + \text{constant}, \quad (7)$$

where both $p(T|S_k)$ and $p(T)$ can be computed directly utilizing existing LLMs, independent of their architecture, and without the need for fine-tuning on extensive textual data. Specifically, $p(T|S_k)$ represents the probability predicted by an individual contextual set, while $p(T)$ signifies the probability in the absence of any context or with an empty context. It is noteworthy that multiple contextual sets can be concurrently processed within the same batch, with computational complexity scaling linearly with the number of contexts. Certainly, Naive Bayes leans heavily on the independence assumption, which can restrict its practical utility. To aspire to enhance its performance beyond the initial state, we further refine Equation 7.

To commence this refinement, we shall introduce the following notations:

$$\log p(T|S) = [\log p(T|S_1), \ldots, \log p(T|S_n)], \tag{8}$$

and

$$\overline{\log p(T|S)} = \frac{1}{n}\sum_{k=1}^{n} \log p(T|S_k), \tag{9}$$

where $\overline{\log p(T|S)}$ denotes the Average Pooling of $\log p(T|S)$. Let $\beta = n - 1$, then Equation 7 can be rewritten as

$$\begin{aligned} \log p(T|S_1, S_2, \ldots, S_n) = (\beta + 1)\overline{\log p(T|S)} \\ - \beta \log p(T) \\ + \text{constant}. \end{aligned} \tag{10}$$

However, the reformulation may prompt the emergence of two inherent inquiries:

- If we consider β as a hyperparameter subject to tuning, could this potentially yield superior results?
- Is it conceivable that employing alternative pooling techniques, denoted as P, might potentially yield enhancements in performance? That is:

$$\begin{aligned} \log p(T|S_1, S_2, \ldots, S_n) = (\beta + 1)P[\log p(T|S)] \\ - \beta \log p(T) \\ + \text{constant} \end{aligned} \tag{11}$$

To delve deeper into these inquiries, we conducted a series of experiments employing the 7B model and garnered preliminary insights. In the realm of reading comprehension, a consistent trend of robust performance emerges when employing Max Pooling with a β value of 0.25 in conjunction with Greedy Search. Conversely, outcomes generated via Random Sampling frequently yield results that are challenging to interpret.

The observed disparities in outcomes can be attributed to the inherent characteristics of these two methods. Random Sampling, characterized by its selection of tokens based on their probability distribution, tends to exhibit lackluster performance, signaling that the output of Max Pooling may not align with a plausible probability distribution. In contrast, Greedy Search operates distinctively by prioritizing the token with the highest probability, disregarding the holistic distribution. Its commendable performance suggests that the token with the highest probability is more likely to be the accurate choice. Larger probabilities are indicative of lower uncertainty. To enhance the performance of Random Sampling, we modify the pooling method to directly output the probability distribution with the lowest uncertainty:

$$\begin{aligned} P[\log p(T|S)] = \log p(T|S_k), \\ k = \text{argmin}\{H_1, H_2, \ldots, H_n\}, \\ H_i = -\sum_T p(T|S_i) \log p(T|S_i), \end{aligned} \tag{12}$$

By substituting this expression into Equation 11, we arrive at the conclusive formulation of the NBCE. It is noteworthy that while the initial inspiration for this approach stemmed from Naive Bayes, the generalized Equation 11 transcends the conventional boundaries of traditional Naive Bayes, yet maintains its inherent interpretability. Equation 11 assumes an intuitive form: Predictions originating from various contextual sources are collectively amalgamated (or weighted) through the utilization of the method denoted as P (with a weight factor of $\beta + 1$). Subsequently, this amalgamation is counterbalanced by subtracting the prediction in the absence of context, weighted by β. The rationale behind subtracting the context-less prediction lies in enhancing the model's reliance on contextual information, reducing its dependency on inherent knowledge (Shi et al., 2023).

The choice of values for β can be tailored to different scenarios. For tasks necessitating comprehensive reading comprehension and robust context integration, a larger β value may be deemed appropriate. Conversely, tasks leaning towards creative writing may benefit from a smaller β value. In our experiments, we set $\beta = 0.25$.

4 Experimental Setup

4.1 Datasets

In our experiments, we employed a diverse range of benchmark datasets to evaluate our approach. These datasets encompassed various tasks, including text classification and multiple-choice questions. Fifteen Text Classification Datasets: SST-2 (Socher et al., 2013), CR (Ding et al., 2008), RTE (Bar-Haim et al., 2014), Subj (Pang and Lee, 2004), CB (De Marneffe et al., 2019), AGNews (Zhang et al., 2015), SST-5 (Socher et al., 2013), YELP (Zhang et al., 2015), TREC (Li and Roth, 2002), DBPedia (Zhang et al., 2015), NLU (Liu et al., 2019), BANKING77 (Casanueva et al., 2020), CLINIC150 (Larson et al., 2019), TREC (fine-grained labels) and NLU (fine and coarse-grained labels). Five datasets from Multiple-choice Domain. Specifically, we consider sentence completion: HellaSwag (Zellers et al., 2019); commensense reasoning: PIQA (Bisk et al., 2020), OpenBookQA (Mihaylov et al., 2018), StoryCloze (Mostafazadeh et al., 2017), MMLU (Hendrycks et al., 2021), ARC-Easy (Bhakthavatsalam et al., 2021); and COPA from SuperGLUE benchmark (Wang et al., 2019). It is worth noting that we conducted evaluations using the standard test sets or validation sets when a public test set was not available. It is important to mention that all the datasets used in our experiments are in the English language.

4.2 Training Sampling and Models

The effectiveness of ICL has been observed to be highly dependent on the selection of training examples (Zhao et al., 2021). To ensure a fair and consistent comparison, we maintain the approach employed in the PCW (Ratner et al., 2023), a common practice in prior research (Zhao et al., 2021; Lu et al., 2022). Specifically, we randomly selected 30 sets from the training datasets and report the mean and standard deviation calculated across these sampled sets.

Given our limited computational resources, our experiments were conducted using eight large models: GPT2-Large (0.75B), GPT2-XL(1.5B)(Radford et al., 2019), there LLAMA models, including 7B, 13B and 30B (Touvron et al., 2023), and three OPT models with 1.3B, 6.7B and 13B parameters (Zhang et al., 2022b).

4.3 Comparative Baseline

Note that our proposed solution does not require any additional training. As far as our knowledge extends, Ratner et al. (2023) initiated the work in this line of research. Therefore, we compare our approach with methods that also do not require further training, as follows.

- **ICL**. A traditional ICL approach employs a conventional single context window, which essentially utilizes the full capacity of the positional embedding in the LLM.
- **PCW**(Ratner et al., 2023). PCW introduces strategic adjustments to both position encoding and attention mask mechanisms to enable multiple context windows without requiring additional training.

4.4 Prompt Formats

We have employed the same prompt formats as those adapted by the comparative baseline, PCW. For the sake of brevity, we have omitted specific details about the prompt format; for a more comprehensive understanding, we kindly refer you to Ratner et al. (2023).

5 Evaluation

We evaluate our proposed solution based on two primary criteria:

- **Ability to Extend the Length of Large Models:** Does our solution effectively enable the expansion of the size or capacity of large models?
- **Impact of Additional Demonstrations on ICL Task Performance:** Does the inclusion of more demonstrations have a positive effect on the performance of the ICL task?

5.1 Classification Task Evaluation

5.1.1 Main Results

We conducted an analysis in which we calculated the average accuracy from 30 different runs, each with a unique seed. We compiled the accuracy and standard deviation for various text classification datasets, which are presented in Tables 1, 2, and 3. **Due to space constraints, the results of more scaled models are presented in the Appendix Section: GPT2-XL Table 6, LLAMA-13B Table 7, LLAMA-30B Table 8, and OPT-6.7B Table 12**. To highlight significant findings, we marked statistical significance with an asterisk (*), based on a t-test with a p-value of less than 0.05. Our key observations are as follows. (1) Vanilla ICL consistently showed the lowest performance across all models and datasets, underscoring the critical need for expanded context in ICL tasks. (2) For models with fewer parameters (like GPT-2-Large and OPT-1.3B) and when dealing with a limited number of output classes (five or fewer), we noted minor

Dataset	# Labels	ICL	B=3		B=6		B=9	
			PCW	NBCE	PCW	NBCE	PCW	NBCE
SST-2	2	80.2 ± 11.7	84.1 ± 8.2	**85.2** ± 6.7	81.2 ± 7.0	**83.6** ± 7.0	78.9 ± 5.3	**84.3** ± 5.9*
CR	2	81.3 ± 6.3	81.2 ± 6.4	**82.7** ± 6.3	82.3 ± 5.2	**84.7** ± 4.6	81.2 ± 3.4	**84.1** ± 4.4*
SUBJ	2	65.1 ± 11.9	**67.0** ± 12.2	66.1 ± 13.2	62.9 ± 10.9	**66.2** ± 10.7	60.1 ± 2.8	**64.4** ± 9.9*
CB	2	43.9 ± 3.7	43.9 ± 3.2	**45.2** ± 3.7	42.8 ± 2.1	**44.8** ± 3.3*	42.1 ± 2.2	**45.1** ± 5.0*
RTE	2	52.5 ± 2.2	**53.5** ± 1.7	52.9 ± 2.9	**54.4** ± 1.0*	53.0 ± 2.4	53.9 ± 2.6	**54.2** ± 2.5
AGNews	4	61.7 ± 14.2	70.9 ± 9.4	**71.0** ± 8.9*	**67.7** ± 7.0	67.1 ± 10.6	64.8 ± 3.1	**72.9** ± 7.6*
SST5	5	40.8 ± 2.5	41.5 ± 3.1	**41.8** ± 2.4	37.4 ± 4.1	**42.5** ± 1.9*	35.9 ± 2.8	**41.9** ± 2.4*
TREC	6	56.6 ± 7.9	59.0 ± 4.7	**63.1** ± 7.0*	53.9 ± 3.1	**65.3** ± 3.0*	50.9 ± 3.4	**66.5** ± 2.9*
DBPedia	14	58.7 ± 20.2	**78.9** ± 6.6*	71.1 ± 13.7	**79.3** ± 4.2	75.9 ± 8.2	68.1 ± 1.9	**76.7** ± 5.7*
NLU Scenario	18	34.8 ± 7.6	28.5 ± 4.3	**45.7** ± 6.7*	26.9 ± 3.2	**41.7** ± 8.5*	24.4 ± 1.6	**44.1** ± 6.1*
TREC Fine	50	31.2 ± 7.9	33.9 ± 4.4	**36.9** ± 6.3*	31.3 ± 3.5	**40.3** ± 5.1*	26.5 ± 4.2	**39.3** ± 3.9*
NLU Intent	68	24.5 ± 6.1	22.3 ± 5.6	**27.5** ± 4.6*	19.8 ± 4.7	**28.6** ± 6.1*	15.5 ± 3.4	**31.1** ± 4.7*
BANKING77	77	28.9 ± 5.1	28.0 ± 3.7	**36.0** ± 3.2*	23.0 ± 3.3	**37.1** ± 3.4*	18.5 ± 2.7	**38.5** ± 3.6*
CLINIC150	150	43.9 ± 3.2	44.1 ± 1.9	**48.5** ± 2.3*	40.4 ± 1.7	**49.4** ± 1.5*	35.0 ± 1.9	**49.7** ± 1.8*

Table 1: Comparative Analysis of Classification Accuracy (in %) for GPT2-Large Using Various Context Windows (B=3, B=6, B=9). **Note: A single window (B) includes K examples**, falling within the model's capacity (e.g., 1024 tokens in GPT-2). For detailed information on the maximum number of examples (K) for each dataset and model, refer to Appendix Section A.2. Best scores are highlighted in bold. An asterisk (*) denotes statistical significance, as determined by a t-test with a p-value < 0.05. The results of GPT-2-Xl are presented in Appendix Table 6.

Dataset	# Labels	ICL	B=3		B=6		B=9	
			PCW	NBCE	PCW	NBCE	PCW	NBCE
SST-2	2	93.4 ± 1.3	**94.9** ± 0.6*	93.8 ± 0.9	91.7 ± 1.0	**94.0** ± 0.9*	84.5 ± 0.9	**94.1** ± 0.7*
CR	2	93.9 ± 0.7	93.5 ± 0.6	**94.1** ± 0.6*	90.0 ± 1.0	**94.0** ± 0.5*	79.3 ± 3.3	**94.2** ± 0.5*
SUBJ	2	70.1 ± 9.9	60.5 ± 7.6	**74.2** ± 7.5*	49.8 ± 1.8	**69.8** ± 7.3*	48.4 ± 0.0	**71.4** ± 6.9*
CB	2	81.3 ± 5.7	**81.9** ± 7.4	77.8 ± 8.3	76.4 ± 5.2	**78.4** ± 7.5	62.2 ± 3.0	**83.9** ± 3.7*
RTE	2	72.9 ± 3.1	**73.8** ± 1.9	73.1 ± 3.1	67.2 ± 2.5	**74.4** ± 1.8*	57.5 ± 1.4	**74.2** ± 2.4*
AGNews	4	87.9 ± 2.8	87.3 ± 1.7	**88.6** ± 1.6	87.4 ± 1.1	**88.8** ± 1.6*	83.1 ± 1.8	**89.3** ± 1.0*
SST5	5	40.8 ± 5.6	**44.6** ± 3.8*	43.1 ± 3.5	40.4 ± 4.4	**42.5** ± 3.2	22.9 ± 3.0	**42.9** ± 2.6*
TREC	6	83.4 ± 5.4	81.1 ± 3.9	**83.5** ± 4.7	55.1 ± 3.8	**86.4** ± 3.7*	41.2 ± 4.0	**88.8** ± 3.0*
DBPedia	14	86.7 ± 6.8	**94.9** ± 3.0*	93.2 ± 3.3	**95.7** ± 1.6	95.6 ± 2.4	92.7 ± 1.3	**96.8** ± 1.3*
NLU Scenario	18	79.6 ± 3.0	79.7 ± 2.5	**83.8** ± 2.2*	58.4 ± 2.9	**85.0** ± 1.6*	40.4 ± 4.9	**86.3** ± 1.4*
TREC Fine	50	55.6 ± 6.1	49.5 ± 5.4	**57.8** ± 6.8*	33.5 ± 3.6	**59.8** ± 5.0*	16.9 ± 2.9	**60.9** ± 4.5*
NLU Intent	68	59.9 ± 5.2	**62.9** ± 3.9*	54.3 ± 2.9	37.3 ± 5.6	**56.6** ± 3.1*	14.8 ± 3.4	**57.9** ± 2.5*
BANKING77	77	46.3 ± 4.0	**51.2** ± 3.3*	50.5 ± 3.1	26.6 ± 4.5	**54.6** ± 3.3*	11.2 ± 3.2	**58.9** ± 2.5*
CLINIC150	150	**61.3** ± 2.5*	57.0 ± 3.2	55.4 ± 2.6	32.8 ± 4.8	**57.2** ± 1.8*	17.1 ± 4.0	**60.8** ± 1.9*

Table 2: Comparative Analysis of Classification Accuracy (in %) for LLAMA-7B Across Various Context Windows. The results of LLAMA-13B and LLAMA-30B are presented in Appendix Section Tables 7 and 8.

or negligible differences between both PCW and NBCE, compared to vanilla ICL. Conversely, in models with a larger number of parameters, NBCE generally demonstrated superior performance in most cases. However, it is important to note that several of these differences did not reach statistical significance. (3) NBCE enhances ICL by accommodating a greater number of examples. This improvement becomes particularly evident when B=9, where both accuracy and stability generally show marked improvements. We observed that larger models benefit more substantially from our approach. This favorable scaling trend of NBCE is particularly notable when contrasted with previous efforts to enhance ICL (refer to (Zhao et al., 2021; Lu et al., 2022)), where improvements in 178B-scale models were less marked compared to those in smaller models

5.1.2 PCW enables ICL with a Large Number of Classes

To investigate the relationship between the number of classes and our NBCE's performance, we conducted a detailed analysis, which was adapted by Ratner et al. (2023). In each experiment, we calculated the difference between NBCE and PCW and then averaged the results across all datasets on GPT2 models sharing the same number of classes. As illustrated in Figure 2, a robust positive correlation emerged between the quantity of classes

Dataset	# Labels	ICL	B=3		B=6		B=9	
			PCW	NBCE	PCW	NBCE	PCW	NBCE
SST-2	2	85.0 ± 8.5	81.7 ± 10.6	**86.0 ± 7.2**	81.1 ± 7.7	**88.1 ± 5.7***	79.9 ± 9.8	**88.8 ± 5.2***
CR	2	89.1 ± 2.4	88.8 ± 2.3	**89.7 ± 1.7**	88.5 ± 3.3	**88.8 ± 1.6**	85.6 ± 3.6	**89.1 ± 1.5***
SUBJ	2	**78.8 ± 9.0***	68.3 ± 7.5	69.0 ± 7.9	68.5 ± 6.6	**70.5 ± 7.4**	65.2 ± 8.3	**70.9 ± 6.3***
CB	2	**53.0 ± 6.0**	50.5 ± 3.3	50.8 ± 3.3	**51.6 ± 5.2**	51.5 ± 4.3	49.1 ± 1.0	**51.6 ± 3.6***
RTE	2	51.1 ± 3.7	51.8 ± 3.8	**52.7 ± 3.2**	50.6 ± 3.1	**51.4 ± 2.9**	50.9 ± 2.1	**51.3 ± 2.5**
AGNews	4	61.3 ± 10.3	**67.4 ± 6.7***	59.6 ± 7.2	**65.1 ± 5.9***	60.3 ± 9.0	**69.4 ± 5.0***	62.9 ± 6.7
SST-5	5	44.0 ± 3.9	42.7 ± 4.6	**44.8 ± 2.8**	42.4 ± 4.0	**44.8 ± 2.2***	41.6 ± 4.3	**45.1 ± 2.0***
TREC	6	**59.4 ± 6.3***	55.0 ± 4.3	56.8 ± 4.7	55.2 ± 3.2	**55.7 ± 4.3**	52.5 ± 2.8	**57.1 ± 3.9***
DBPedia	14	86.3 ± 3.8	87.7 ± 2.1	**87.9 ± 2.2**	**88.1 ± 2.6**	87.5 ± 2.6	87.0 ± 3.1	**87.9 ± 2.6**
NLU Scenario	18	67.8 ± 4.0	69.9 ± 3.5	**70.2 ± 4.0**	**69.9 ± 2.6**	69.3 ± 4.3	67.7 ± 4.0	**72.8 ± 3.8***
TREC Fine	50	39.7 ± 4.5	38.8 ± 4.7	**41.5 ± 6.0**	40.5 ± 5.8	**43.1 ± 6.4**	35.3 ± 3.5	**42.0 ± 4.7***
NLU Intent	68	45.3 ± 4.9	50.0 ± 4.2	**50.9 ± 4.0**	48.8 ± 4.2	**51.0 ± 4.7**	45.4 ± 3.2	**54.5 ± 3.3***
BANKING77	77	25.9 ± 4.9	24.8 ± 4.0	**28.8 ± 4.5**	26.0 ± 3.5	**30.1 ± 3.5***	28.9 ± 3.1	**32.5 ± 3.5***
CLINIC150	150	50.8 ± 3.0	52.4 ± 2.3	**57.7 ± 2.0**	52.6 ± 2.0	**57.2 ± 2.5***	49.3 ± 2.5	**58.4 ± 2.0***

Table 3: Comparative Analysis of Classification Accuracy (in %) for OPT-1.3B models. The results of OPT-6.7B are presented in Appendix Tables 12.

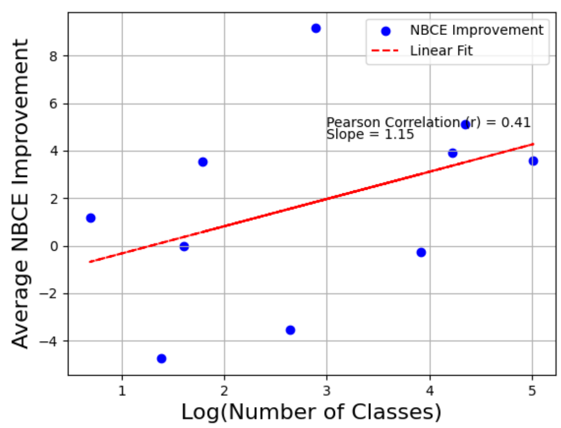

Figure 2: Average Performance Enhancements with NBCE over PCW as a Function of Label Count: Each data point in our analysis signifies the average improvement observed across all datasets on GPT2 models. It is worth noting a clear and positive correlation between the quantity of unique labels and the benefits derived from our NBCE.

and the improvements achieved by NBCE. Specifically, the Pearson correlation coefficient (r) was 0.41 when considering the logarithm of class numbers in relation to the average improvement, with a slope of 1.15. Remarkably, for datasets featuring numerous labels, such as NLU Intent (Liu et al., 2019), Banking77 (Casanueva et al., 2020), and CLINIC150 (Larson et al., 2019), we observed substantial improvements ranging from 3.6 to 5.1 points in most cases.

When comparing results across datasets with varying numbers of classes, it is crucial to account for potential confounding factors, such as variations in domain, style, or genre. To mitigate these effects, we conducted a comparison using two datasets, each featuring both fine-grained and coarse-grained labels. The TREC dataset (Li and Roth, 2002), which includes 6 coarse-grained classes. The NLU dataset (Liu et al., 2019), comprising 18 scenarios coarse-grained classes and 68 intents coarse-grained classes. Our analysis on GPT2 models, as presented in Table 10, reveals that NBCE outperforms PCW by 4.1 and 3.0 improvements on GPT2-Large and GPT2-XLarge, respectively. Similarly, in the context of NLU, we observe average improvements of 17.2 and 5.2 points on GPT2-XLarge, respectively. These findings underscore the effectiveness of our approach, particularly when confronted with a large number of output classes.

5.2 Multi-Choice Tasks

Table 4 shows the evaluation of multi-choice tasks. It is important to note that the improvements made by both PCW and our NBCE in these tasks, compared to text classification, are relatively modest, with a slight edge for NBCE. Furthermore, employing a greater number of demonstrations does not consistently translate to better performance in multi-choice tasks. Instead, we observe that scaling up the model size (Appendix Section Table 9), rather than increasing the number of demonstrations, tends to yield more substantial improvements in these tasks.

5.3 Impact of more Demonstrations on ICL

We conducted experiments to validate the impact of additional demonstrations on ICL in NLP mod-

Dataset	ICL	B=2		B=3		B=4		B=6	
		PCW	NBCE	PCW	NBCE	PCW	NBCE	PCW	NBCE
PIQA	81.6 ± 0.6	80.6 ± 0.7	**82.1** ± 0.4*	79.6 ± 0.7	**82.9** ± 0.6*	79.1 ± 0.6	**82.9** ± 0.6*	77.5 ± 0.8	**83.0** ± 0.5*
OpenBookAQ	41.9 ± 0.8	41.3 ± 1.0	**46.3** ± 0.9*	40.9 ± 0.9	**49.2** ± 0.8*	39.4 ± 0.6	**49.3** ± 0.9*	35.1 ± 0.8	**50.3** ± 1.1*
COPA	77.8 ± 1.2	**78.3** ± 1.1	78.2 ± 1.5	**78.9** ± 1.7*	77.5 ± 1.2	**77.8** ± 1.3	77.6 ± 1.6	65.9 ± 3.6	**76.6** ± 0.8*
HellaSwag	79.4 ± 1.1	**80.4** ± 1.1*	78.9 ± 0.9	**80.2** ± 0.8*	79.6 ± 0.7	**80.1** ± 0.9	79.9 ± 0.8	78.5 ± 0.8	**79.9** ± 0.7*
ARCE	**74.4** ± 1.1*	73.8 ± 1.2	72.8 ± 0.7	**73.7** ± 1.4	73.5 ± 0.6	**74.1** ± 0.8	73.7 ± 0.8	70.8 ± 1.5	**73.5** ± 0.8*
StoryCloze	46.0 ± 0.0	46.1 ± 0.1	**78.7** ± 0.9*	46.1 ± 0.2	**78.9** ± 0.8*	46.1 ± 0.2	**78.8** ± 1.0*	46.3 ± 0.2	**79.6** ± 0.7*
MMLU	33.8 ± 1.9	34.1 ± 2.2	**34.3** ± 1.5	33.6 ± 2.3	**33.7** ± 1.7	34.1 ± 1.9	**34.7** ± 1.9	32.5 ± 3.0	**33.9** ± 1.9

Table 4: Comparative Results of Task Completion (e.g., Multiple Choices Task) for LLAMA-7B Using Various Context Windows. Best scores are highlighted in bold. An asterisk (*) denotes statistical significance, as determined by a t-test with a p-value < 0.05. The results of LLAMA-13B are presented in Appendix Tables 9.

Dataset	# Labels	GPT2-Large		GPT2-XL		LLAMA-7B		LLAMA-13B	
		NBCE (RAN)	NBCE	NBCE (RAN)	NBCE	NBCE (RAN)	NBCE	NBCE (RAN)	NBCE
SST-2	2	80.5 ± 4.5	**84.3** ± 5.9*	91.6 ± 1.5	**92.5** ± 1.5	92.3 ± 1.5	**94.1** ± 0.7*	92.2 ± 1.0	**94.9** ± 0.5*
CR	2	78.0 ± 3.9	**84.1** ± 4.4*	81.0 ± 2.2	**81.9** ± 2.0	91.9 ± 1.2	**94.2** ± 0.5*	91.1 ± 1.3	**93.1** ± 0.6*
SUBJ	2	57.0 ± 3.8	**64.4** ± 9.9*	72.0 ± 5.0	**76.0** ± 7.0	69.0 ± 3.4	**71.4** ± 6.9	89.9 ± 3.0	**93.0** ± 1.7*
CB	2	**46.1** ± 4.4	45.1 ± 5.0	**55.3** ± 6.2	54.8 ± 8.5	81.6 ± 5.1	**83.9** ± 3.7*	81.7 ± 4.0	**84.1** ± 3.5*
RTE	2	52.5 ± 2.8	**54.2** ± 2.5	53.9 ± 2.9	**55.3** ± 2.2	68.2 ± 1.9	**74.2** ± 2.4*	72.9 ± 2.3	**75.1** ± 1.5*
AGNews	4	66.4 ± 7.5	**72.9** ± 7.6*	69.5 ± 5.9	**76.3** ± 4.7*	83.4 ± 2.1	**89.3** ± 1.0*	85.3 ± 2.3	**87.9** ± 1.1*
SST5	5	41.3 ± 1.8	**41.9** ± 2.4	39.1 ± 3.6	**41.7** ± 5.3	40.4 ± 2.7	**42.9** ± 2.6*	44.5 ± 2.1	**47.7** ± 2.0*
TREC	6	61.0 ± 2.8	**66.5** ± 2.9*	50.7 ± 2.8	**51.6** ± 3.0	84.1 ± 3.5	**88.8** ± 3.0*	81.7 ± 4.4	**85.0** ± 2.4*
DBPedia	14	68.9 ± 8.2	**76.7** ± 5.7*	84.1 ± 2.5	**89.0** ± 2.8*	82.8 ± 2.7	**96.8** ± 1.3*	89.2 ± 3.4	**96.9** ± 1.3*
NLU Scenario	18	40.8 ± 4.8	**44.1** ± 6.1	45.3 ± 3.9	**55.1** ± 5.4*	82.0 ± 2.1	**86.3** ± 1.4*	81.7 ± 1.8	**88.7** ± 1.0*
TREC Fine	50	33.2 ± 4.2	**39.3** ± 3.9*	35.2 ± 4.4	**41.9** ± 3.7*	56.7 ± 3.1	**60.9** ± 4.5*	57.1 ± 3.5	**63.3** ± 4.1*
NLU Intent	68	28.3 ± 0.8	**31.1** ± 4.7*	35.1 ± 1.2	**40.3** ± 3.6*	57.2 ± 2.1	**57.9** ± 2.5*	**62.6** ± 2.4*	61.8 ± 2.1
BANKING77	77	29.3 ± 1.6	**38.5** ± 3.6*	33.6 ± 1.3	**38.9** ± 2.4*	47.0 ± 1.5	**58.9** ± 2.5*	48.7 ± 3.2	**63.5** ± 2.3*
CLINIC150	150	43.8 ± 1.7	**49.7** ± 1.8*	47.7 ± 1.1	**51.6** ± 1.7*	58.7 ± 2.1	**60.8** ± 1.9*	62.5 ± 2.2	**66.2** ± 2.2*

Table 5: Ablation Study with Context Window B=9. Best scores are highlighted in bold. An asterisk (*) denotes statistical significance, as determined by a t-test with a p-value < 0.05.

els. Our focus was to show how extra demonstrations (B=6 and B=9, where B is the window size) enhance model performance by improving context understanding and robustness. **Note that each window contains K samples within the model's token limit (e.g., 2024 tokens for LLAMA).** For detailed information on the maximum value of K for each model and dataset, please see Appendix Table **??** . This approach aligns with the importance of training example quantity in model adaptability and generalization (Murtadha et al., 2023). Our observations indicate that NBCE mostly outperforms its counterpart, PCW, and these improvements can be considered significant. Additionally, scaling up the model size (Appendix Section Tables 6,7, 8, and 12) leads to improved performance, especially on larger and more complex datasets.

5.4 Ablation Study

To better evaluate the proposed voting mechanism, i.e., selecting the best k contexts as the posterior in Equation 12, we conducted an ablation study introducing a new variant, referred to as NBCE (RAND). In this variant, rather than deliberately choosing k, we randomly select one context from the context windows. The results are presented in Table 5. The experimental outcomes across a variety of models and datasets demonstrate that a careful selection of k significantly contributes to the quality of the generated tokens. It is noteworthy that, in this setting, NBCE can be considered as a standard ICL, where only one context window is considered. However, the performance may slightly differ due to the likelihood of the generated text $p(T)$, as outlined in Equation 11, affecting the final performance.

6 Conclusion

This paper introduces a novel framework called Naive Bayes-based Context Extension (NBCE) for large language models. NBCE innovatively incorporates a voting mechanism to select the most appropriate window context, and then utilizes Bayes' theorem to generate the task text. Our results show that NBCE outperforms its alternative PCW across a diverse set of multi-class classification tasks. For future work, while PCW shows effective without additional training, ICL could potentially benefit from more demonstrations in fine-tuning settings; however, further investigation is required to fully comprehend the extent of its advantages.

Limitations

NBCE facilitates ICL tasks by allowing for more demonstrations without the need for fine-tuning. However, there are still some limitations to this approach:

- Since NBCE essentially functions as a voting mechanism, its effectiveness is constrained in tasks that require ordered or interrelated contexts, such as code generation. This is due to its inherent nature, which may not adequately handle sequential or dependent information in certain contexts.
- Increasing the number of shots does not necessarily lead to improved performance. Experimental results have indicated that expanding the context window size does not significantly enhance performance in completion tasks. This suggests a diminishing return on performance gains with an increased number of contexts.

References

Roy Bar-Haim, Ido Dagan, and Idan Szpektor. 2014. Benchmarking applied semantic inference: The PASCAL recognising textual entailment challenges. In *Language, Culture, Computation. Computing - Theory and Technology - Essays Dedicated to Yaacov Choueka on the Occasion of His 75th Birthday, Part I*, volume 8001 of *Lecture Notes in Computer Science*, pages 409–424. Springer.

Sumithra Bhakthavatsalam, Daniel Khashabi, Tushar Khot, Bhavana Dalvi Mishra, Kyle Richardson, Ashish Sabharwal, Carissa Schoenick, Oyvind Tafjord, and Peter Clark. 2021. Think you have solved direct-answer question answering? try arc-da, the direct-answer AI2 reasoning challenge. *CoRR*, abs/2102.03315.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. PIQA: Reasoning about Physical Commonsense in Natural Language. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7432–7439. AAAI Press.

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 *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Iñigo Casanueva, Tadas Temcinas, Daniela Gerz, Matthew Henderson, and Ivan Vulic. 2020. Efficient intent detection with dual sentence encoders. *CoRR*, abs/2003.04807.

Yanda Chen, Ruiqi Zhong, Sheng Zha, George Karypis, and He He. 2022. Meta-learning via language model in-context tuning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 719–730. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. PaLM: Scaling Language Modeling with Pathways. *CoRR*, abs/2204.02311.

Marie-Catherine De Marneffe, Mandy Simons, and Judith Tonhauser. 2019. The commitmentbank: Investigating projection in naturally occurring discourse. In *proceedings of Sinn und Bedeutung*, volume 23, pages 107–124.

Xiaowen Ding, Bing Liu, and Philip S. Yu. 2008. A holistic lexicon-based approach to opinion mining. In *Proceedings of the International Conference on Web Search and Web Data Mining, WSDM 2008, Palo Alto, California, USA, February 11-12, 2008*, pages 231–240. ACM.

Shivam Garg, Dimitris Tsipras, Percy Liang, and Gregory Valiant. 2022. What Can Transformers Learn In-Context? A Case Study of Simple Function Classes. In *NeurIPS*.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. 2023. Pre-Training to Learn in Context. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 4849–4870. Association for Computational Linguistics.

Mandy Guo, Joshua Ainslie, David C. Uthus, Santiago Ontañón, Jianmo Ni, Yun-Hsuan Sung, and Yinfei Yang. 2022. Longt5: Efficient text-to-text transformer for long sequences. In *Findings of the Association for Computational Linguistics: NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 724–736. Association for Computational Linguistics.

Xu Han, Zhengyan Zhang, Ning Ding, Yuxian Gu, Xiao Liu, Yuqi Huo, Jiezhong Qiu, Yuan Yao, Ao Zhang, Liang Zhang, Wentao Han, Minlie Huang, Qin Jin, Yanyan Lan, Yang Liu, Zhiyuan Liu, Zhiwu Lu, Xipeng Qiu, Ruihua Song, Jie Tang, Ji-Rong Wen, Jinhui Yuan, Wayne Xin Zhao, and Jun Zhu. 2021. Pre-trained models: Past, present and future. *AI Open*, 2:225–250.

Zhixiong Han, Yaru Hao, Li Dong, Yutao Sun, and Furu Wei. 2023. Prototypical calibration for few-shot learning of language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Yaru Hao, Yutao Sun, Li Dong, Zhixiong Han, Yuxian Gu, and Furu Wei. 2022. Structured Prompting: Scaling In-Context Learning to 1, 000 Examples. *CoRR*, abs/2212.06713.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Maor Ivgi, Uri Shaham, and Jonathan Berant. 2022. Efficient long-text understanding with short-text models. *CoRR*, abs/2208.00748.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 874–880. Association for Computational Linguistics.

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 2019, Hong Kong, China, November 3-7, 2019*, pages 1311–1316. Association for Computational Linguistics.

Mukai Li, Shansan Gong, Jiangtao Feng, Yiheng Xu, Jun Zhang, Zhiyong Wu, and Lingpeng Kong. 2023. In-Context Learning with Many Demonstration Examples. *CoRR*, abs/2302.04931.

Xin Li and Dan Roth. 2002. Learning question classifiers. In *19th International Conference on Computational Linguistics, COLING 2002, Howard International House and Academia Sinica, Taipei, Taiwan, August 24 - September 1, 2002*.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for gpt-3? In *Proceedings of Deep Learning Inside Out: The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures, DeeLIO@ACL 2022, Dublin, Ireland and Online, May 27, 2022*, pages 100–114. Association for Computational Linguistics.

Xingkun Liu, Arash Eshghi, Pawel Swietojanski, and Verena Rieser. 2019. Benchmarking natural language understanding services for building conversational agents. In *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction - 10th International Workshop on Spoken Dialogue Systems, IWSDS 2019, Syracuse, Sicily, Italy, 24-26 April 2019*, volume 714 of *Lecture Notes in Electrical Engineering*, pages 165–183. Springer.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 8086–8098. Association for Computational Linguistics.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? A new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2381–2391. Association for Computational Linguistics.

Sewon Min, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022a. Noisy channel language model prompting for few-shot text classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 5316–5330. Association for Computational Linguistics.

Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2022b. MetaICL: Learning to

Learn In Context. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 2791–2809. Association for Computational Linguistics.

Nasrin Mostafazadeh, Michael Roth, Annie Louis, Nathanael Chambers, and James F. Allen. 2017. LSDSem 2017 Shared Task: The Story Cloze Test. In *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics, LSDSem@EACL 2017, Valencia, Spain, April 3, 2017*, pages 46–51. Association for Computational Linguistics.

Ahmed Murtadha, Shengfeng Pan, Wen Bo, Jianlin Su, Xinxin Cao, Wenze Zhang, and Yunfeng Liu. 2023. Rank-Aware Negative Training for Semi-Supervised Text Classification. *Transactions of the Association for Computational Linguistics*, 11:771–786.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, 21-26 July, 2004, Barcelona, Spain*, pages 271–278. ACL.

Ofir Press, Noah A. Smith, and Mike Lewis. 2022. Train short, test long: Attention with linear biases enables input length extrapolation. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. 2020. Pre-trained Models for Natural Language Processing: A Survey. *CoRR*, abs/2003.08271.

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.

Nir Ratner, Yoav Levine, Yonatan Belinkov, Ori Ram, Inbal Magar, Omri Abend, Ehud Karpas, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. Parallel context windows for large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 6383–6402. Association for Computational Linguistics.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2022. Learning to retrieve prompts for in-context learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 2655–2671. Association for Computational Linguistics.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Scott Wen-tau Yih. 2023. Trusting your evidence: Hallucinate less with context-aware decoding. *CoRR*, abs/2305.14739.

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, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1631–1642. ACL.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. *CoRR*, abs/2302.13971.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 3261–3275.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontañón, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020. Big bird: Transformers for longer sequences. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4791–4800. Association for Computational Linguistics.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona T. Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022a. OPT: Open Pre-trained Transformer Language Models. *CoRR*, abs/2205.01068.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona T. Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022b. OPT: open pre-trained transformer language models. *CoRR*, abs/2205.01068.

Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 649–657.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate Before Use: Improving Few-shot Performance of Language Models. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR.

A Appendix

A.1 Scaling Model Parameters

A.2 Prompt Format

Dataset	# Labels	ICL	B=3		B=6		B=9	
			PCW	NBCE	PCW	NBCE	PCW	NBCE
SST-2	2	90.6 ± 3.5	92.4 ± 2.5	**92.7** ± 2.3*	89.4 ± 3.5	**92.5** ± 2.2*	83.7 ± 1.7	**92.5** ± 1.5*
CR	2	79.2 ± 5.9	81.3 ± 4.6	**82.5** ± 2.9*	81.6 ± 2.4	**81.9** ± 2.1	**82.7** ± 1.7	81.9 ± 2.0
SUBJ	2	68.8 ± 11.6	64.9 ± 7.3	**74.5** ± 8.3*	57.0 ± 4.1	**78.7** ± 4.8*	65.6 ± 3.0	**76.0** ± 7.0*
CB	2	51.9 ± 7.4	**57.2** ± 8.5*	56.1 ± 7.9	49.6 ± 3.6	**55.8** ± 7.8*	42.2 ± 2.1	**54.8** ± 8.5*
RTE	2	55.4 ± 2.4	**55.6** ± 1.6	54.9 ± 2.5	54.2 ± 1.3	**55.2** ± 2.3*	50.4 ± 2.0	**55.3** ± 2.2*
AGNews	4	67.2 ± 13.2	**79.6** ± 3.4*	70.0 ± 9.6	**80.4** ± 2.3*	74.1 ± 5.8	71.6 ± 2.5	**76.3** ± 4.7*
SST5	5	38.0 ± 6.1	**41.4** ± 4.3*	41.1 ± 4.7	38.1 ± 3.6	**41.5** ± 5.4*	35.3 ± 2.2	**41.7** ± 5.3*
TREC	6	47.9 ± 5.1	48.7 ± 2.8	**51.7** ± 5.0*	45.5 ± 2.3	**51.8** ± 4.6*	43.1 ± 1.9	**51.6** ± 3.0*
DBPedia	14	77.5 ± 9.8	87.0 ± 4.0	**87.7** ± 3.8*	**88.9** ± 3.3	88.6 ± 3.3	81.4 ± 2.1	**89.0** ± 2.8*
NLU Scenario	18	45.1 ± 9.3	50.0 ± 6.1	**51.1** ± 8.1*	46.7 ± 5.9	**50.3** ± 6.8*	38.7 ± 6.3	**55.1** ± 5.4*
TREC Fine	50	36.4 ± 6.2	40.0 ± 3.0	**40.1** ± 5.1*	35.5 ± 2.6	**41.7** ± 3.6*	31.0 ± 2.8	**41.9** ± 3.7*
NLU Intent	68	30.2 ± 5.4	33.8 ± 4.6	**36.4** ± 4.9*	33.4 ± 4.3	**38.5** ± 5.4*	24.3 ± 3.7	**40.3** ± 3.6*
BANKING77	77	30.7 ± 4.1	33.3 ± 3.5	**35.5** ± 2.8*	26.8 ± 3.1	**37.6** ± 2.4*	16.7 ± 2.6	**38.9** ± 2.4*
CLINIC150	150	46.6 ± 2.5	47.1 ± 2.3	**49.9** ± 1.9*	40.8 ± 2.3	**50.9** ± 2.1*	34.5 ± 2.5	**51.6** ± 1.7*

Table 6: Comparative Analysis of Classification Accuracy (in %) for GPT-2-XL Across Various Context Windows (B=3, B=6, B=9). Best scores are highlighted in bold. An asterisk (*) denotes statistical significance, as determined by a t-test with a p-value < 0.05.

Dataset	# Labels	ICL	B=3		B=6		B=9	
			PCW	NBCE	PCW	NBCE	PCW	NBCE
SST-2	2	94.5 ± 0.7	94.1 ± 0.7	**94.8** ± 0.5*	94.0 ± 0.9	**95.0** ± 0.4*	90.1 ± 1.2	**94.9** ± 0.5*
CR	2	92.0 ± 1.4	92.2 ± 0.9	**92.9** ± 1.0*	92.5 ± 0.5	**93.0** ± 1.0*	91.1 ± 0.9	**93.1** ± 0.6*
SUBJ	2	90.2 ± 3.8	87.5 ± 3.3	**90.8** ± 2.9*	79.0 ± 7.2	**92.5** ± 1.7*	67.1 ± 5.4	**93.0** ± 1.7*
CB	2	80.3 ± 8.0	**84.6** ± 4.1*	79.8 ± 4.9	**83.1** ± 4.0*	80.3 ± 6.4	74.1 ± 6.3	**84.1** ± 3.5*
RTE	2	**74.6** ± 2.7	73.5 ± 2.0	74.0 ± 2.5	71.9 ± 1.6	**74.6** ± 1.6*	66.4 ± 2.0	**75.1** ± 1.5*
AGNews	4	86.9 ± 2.9	**87.9** ± 1.7	86.6 ± 1.8	**88.0** ± 0.9	87.3 ± 1.8	87.7 ± 1.1	**87.9** ± 1.1
SST5	5	48.0 ± 3.3	**49.2** ± 2.6	48.0 ± 3.3	**48.4** ± 2.1	47.3 ± 3.4	44.0 ± 2.9	**47.7** ± 2.0*
TREC	6	83.1 ± 3.1	**83.7** ± 2.9*	81.5 ± 3.4	75.5 ± 3.6	**83.0** ± 3.8*	49.5 ± 5.4	**85.0** ± 2.4*
DBPedia	14	88.6 ± 6.1	**93.6** ± 3.9*	93.2 ± 3.9	94.4 ± 2.7	**94.7** ± 2.6	94.5 ± 2.7	**96.9** ± 1.3*
NLU Scenario	18	82.1 ± 2.7	85.9 ± 1.8	**86.7** ± 1.8*	81.2 ± 2.4	**87.4** ± 1.4*	74.1 ± 2.9	**88.7** ± 1.0*
TREC Fine	50	55.4 ± 5.3	**60.1** ± 5.1*	57.7 ± 4.7	56.8 ± 5.4	**60.4** ± 4.7*	47.6 ± 9.0	**63.3** ± 4.1*
NLU Intent	68	68.3 ± 4.1	**73.0** ± 2.6*	58.1 ± 2.3	**65.2** ± 2.6*	60.7 ± 2.7	52.6 ± 3.6	**61.8** ± 2.1*
BANKING77	77	46.6 ± 4.2	**56.4** ± 2.8*	52.8 ± 3.5	50.8 ± 3.1	**59.2** ± 2.8*	40.2 ± 2.5	**63.5** ± 2.3*
CLINIC150	150	63.7 ± 2.5	**66.0** ± 2.7*	59.2 ± 2.3	57.5 ± 2.9	**62.4** ± 1.7*	48.7 ± 2.3	**66.2** ± 2.2*

Table 7: Comparative Analysis of Classification Accuracy (in %) for LLAMA-13B Across Various Context Windows (B=3, B=6, B=9). Best scores are highlighted in bold. An asterisk (*) denotes statistical significance, as determined by a t-test with a p-value < 0.05.

Dataset	# Labels	ICL	B=3		B=6	
			PCW	NBCE	PCW	NBCE
SST-2	2	94.7 ± 0.5	94.9 ± 0.7	**95.0** ± 0.3	92.9 ± 0.7	**95.0** ± 0.3*
CR	2	**93.8** ± 0.5	93.6 ± 0.5	93.8 ± 0.5	93.3 ± 1.1	**93.7** ± 0.4
SUBJ	2	90.3 ± 4.5	91.0 ± 2.7	**93.8** ± 1.7*	83.7 ± 5.1	**94.5** ± 1.6*
CB	2	**88.8** ± 2.5	88.7 ± 1.9	88.0 ± 3.3	83.9 ± 2.4	**89.1** ± 2.2*
RTE	2	**79.9** ± 1.9	79.0 ± 1.8	79.4 ± 2.1	73.8 ± 3.4	**80.6** ± 1.8*
AGNews	4	88.0 ± 4.7	**89.4** ± 0.7	88.9 ± 1.3	88.0 ± 0.8	**88.8** ± 1.4
SST5	5	47.0 ± 2.6	**47.5** ± 2.3	45.0 ± 2.8	**48.4** ± 1.0*	44.5 ± 2.4
TREC	6	87.2 ± 3.3	**90.1** ± 1.7*	88.8 ± 2.8	67.2 ± 4.8	**88.6** ± 1.7*
DBPedia	14	88.4 ± 8.6	94.5 ± 3.0	**95.4** ± 2.6*	96.2 ± 3.0	**96.7** ± 1.4
NLU Scenario	18	82.6 ± 2.0	**85.3** ± 1.5*	84.6 ± 1.7	80.2 ± 2.1	**85.8** ± 1.2*
TREC Fine	50	60.7 ± 4.8	**67.7** ± 4.3*	64.7 ± 3.7	50.1 ± 4.2	**68.6** ± 4.2*
NLU Intent	68	68.6 ± 4.4	**74.4** ± 2.7*	60.1 ± 2.7	**61.6** ± 3.2	61.0 ± 2.2
BANKING77	77	50.3 ± 3.1	**63.2** ± 2.5*	55.3 ± 3.5	58.1 ± 2.7	**63.7** ± 3.6*
CLINIC150	150	67.0 ± 3.6	**71.0** ± 4.2*	65.6 ± 3.0	57.2 ± 2.9	**67.3** ± 2.3*

Table 8: Comparative Analysis of Classification Accuracy (in %) for LLAMA-30B Across Various Context Windows (B=3, B=6, B=9). Best scores are highlighted in bold. An asterisk (*) denotes statistical significance, as determined by a t-test with a p-value < 0.05.

Dataset	ICL	B=2		B=3		B=4		B=6	
		PCW	NBCE	PCW	NBCE	PCW	NBCE	PCW	NBCE
PIQA	83.0 ± 0.6	**83.6** ± 0.6*	83.2 ± 0.6	**83.5** ± 0.6	83.2 ± 0.7	**83.3** ± 0.5	83.2 ± 0.6	81.9 ± 1.0	**83.2** ± 0.5*
OpenBookAQ	51.0 ± 1.7	**51.1** ± 1.2*	47.0 ± 1.1	50.2 ± 1.3	**50.2** ± 1.3	48.8 ± 1.1	**49.8** ± 1.0*	46.7 ± 1.3	**51.1** ± 1.0*
COPA	79.9 ± 2.5	**81.8** ± 2.4*	79.0 ± 0.9	**86.0** ± 1.9*	79.8 ± 2.2	**86.5** ± 1.5*	79.8 ± 2.1	74.9 ± 3.1	**78.4** ± 1.5*
HellaSwag	82.3 ± 0.7	**82.5** ± 1.0	82.5 ± 0.7	**82.3** ± 0.7	82.2 ± 0.5	82.2 ± 0.6	**82.4** ± 0.5	81.7 ± 0.8	**82.2** ± 0.5*
ARCE	80.3 ± 0.6	**80.5** ± 0.7*	77.4 ± 0.7	**79.8** ± 0.5	79.7 ± 0.5	78.9 ± 0.6	**79.8** ± 0.5*	76.8 ± 0.9	**80.5** ± 0.4*
StoryCloze	80.5 ± 0.8	**82.1** ± 0.9*	80.1 ± 0.9	**82.0** ± 0.6*	80.0 ± 0.9	**81.9** ± 0.8*	80.1 ± 1.0	**81.2** ± 0.8*	80.1 ± 0.9
MMLU	45.3 ± 1.8	**46.4** ± 1.9*	43.6 ± 1.3	**45.5** ± 1.9*	44.4 ± 1.3	**44.7** ± 2.1	44.4 ± 2.0	43.6 ± 2.8	**44.6** ± 1.4

Table 9: Comparative Results of Task Completion (e.g., Multiple Choices Task) for LLAMA-13B Using Various Context Windows. Best scores are highlighted in bold. An asterisk (*) denotes statistical significance, as determined by a t-test with a p-value < 0.05.

Dataset	# Labels	GPT2-Large			GPT2-XLarge		
		ICL	PCW	NBCE	ICL	PCW	NBCE
SST-2	2	80.2 ± 11.7	84.1 ± 8.2	**85.2 ± 6.7**	90.6 ± 3.5	92.4 ± 2.5	**92.7 ± 2.3***
CR	2	81.3 ± 6.3	81.2 ± 6.4	**82.7 ± 6.3**	79.2 ± 5.9	81.3 ± 4.6	**82.5 ± 2.9***
SUBJ	2	65.1 ± 11.9	**67.0 ± 12.2**	66.1 ± 13.2	68.8 ± 11.6	64.9 ± 7.3	**74.5 ± 8.3***
CB	2	43.9 ± 3.7	43.9 ± 3.2	**45.2 ± 3.7**	51.9 ± 7.4	**57.2 ± 8.5***	56.1 ± 7.9
RTE	2	52.5 ± 2.2	**53.5 ± 1.7**	52.9 ± 2.9	55.4 ± 2.4	**55.6 ± 1.6**	54.9 ± 2.5
AGNews	4	61.7 ± 14.2	70.9 ± 9.4	**71.0 ± 8.9 ***	67.2 ± 13.2	**79.6 ± 3.4***	70.0 ± 9.6
SST-5	5	40.8 ± 2.5	41.5 ± 3.1	**41.8 ± 2.4**	38.0 ± 6.1	**41.4 ± 4.3***	41.1 ± 4.7
TREC	6	56.6 ± 7.9	59.0 ± 4.7	**63.1 ± 7.0***	47.9 ± 5.1	48.7 ± 2.8	**51.7 ± 5.0***
DBPedia	14	58.7 ± 20.2	**78.9 ± 6.6**	71.1 ± 13.7	77.5 ± 9.8	87.0 ± 4.0	**87.7 ± 3.8***
NLU Scenario	18	34.8 ± 7.6	28.5 ± 4.3	**45.7 ± 6.7***	45.1 ± 9.3	50.0 ± 6.1	**51.1 ± 8.1***
TREC Fine	50	36.9 ± 6.3	**37.4 ± 4.8***	36.9 ± 6.3	36.4 ± 6.2	**40.1 ± 3.0***	40.1 ± 5.1
NLU Intent	68	24.5 ± 6.1	22.3 ± 5.6	**27.5 ± 4.6***	30.2 ± 5.4	33.8 ± 4.6	**36.4 ± 4.9***
BANKING77	77	28.9 ± 5.1	28.0 ± 3.7	**36.0 ± 3.2***	30.7 ± 4.1	33.3 ± 3.5	**35.5 ± 2.8***
CLINIC150	150	43.9 ± 3.2	44.1 ± 1.9	**48.5 ± 2.3***	46.6 ± 2.5	47.1 ± 2.3	**49.9 ± 1.9***

Table 10: Comparative analysis of classification results in terms of accuracy (in %) for both the GPT2-Large and GPT2-XLarge models using a context window of B = 3. Notably, a single window comprises a set of examples with a total number of tokens equal to the maximum capacity of conventional in-context learning (e.g., 1024 tokens in GPT-2). The best-performing scores for each model and dataset are highlighted in bold, while '*' indicates statistical significance, determined by a t-test with a p-value < 0.05.

Dataset	# Labels	OPT-1.3B			OPT-6.7B			OPT-13B		
		ICL	PCW	NBCE	ICL	PCW	NBCE	ICL	PCW	NBCE
SST-2	2	85.0 ± 8.5	81.7 ± 10.6	**86.0 ± 7.2**	93.8 ± 2.6	93.7 ± 3.3	**95.8 ± 1.7***	93.1 ± 4.4	93.8 ± 3.1	**94.9 ± 2.3**
CR	2	89.1 ± 2.4	88.8 ± 2.3	**89.7 ± 1.7**	90.3 ± 2.5	90.7 ± 2.4	**91.7 ± 1.5***	92.7 ± 1.5	92.3 ± 2.5	**93.1 ± 1.4**
SUBJ	2	**78.8 ± 9.0***	68.3 ± 7.5	69.0 ± 7.9	**72.3 ± 10.6***	70.9 ± 13.9	64.0 ± 10.7	86.4 ± 9.2	88.0 ± 8.3	**90.1 ± 5.9**
CB	2	**53.0 ± 6.0**	50.5 ± 3.3	50.8 ± 3.3	52.4 ± 10.1	**59.9 ± 12.1**	59.3 ± 10.8	50.5 ± 8.5	49.3 ± 5.8	**62.5 ± 10.2**
RTE	2	51.1 ± 3.7	51.8 ± 3.8	**52.7 ± 3.2**	56.1 ± 2.2	56.2 ± 1.6	**56.8 ± 2.0**	53.0 ± 6.0	56.3 ± 4.9	**56.8 ± 6.2**
AGNews	4	61.3 ± 10.3	**67.4 ± 6.7***	59.6 ± 7.2	74.8 ± 6.7	**76.7 ± 4.8***	72.7 ± 5.7	78.6 ± 5.6	**82.4 ± 2.3**	78.8 ± 3.9
SST-5	5	44.0 ± 3.9	42.7 ± 4.6	**44.8 ± 2.8**	42.7 ± 5.1	**45.2 ± 4.2**	42.5 ± 4.6	45.6 ± 3.4	**45.7 ± 2.6**	42.9 ± 4.2
TREC	6	**59.4 ± 6.3***	55.0 ± 4.3	56.8 ± 4.7	70.3 ± 3.3	**73.1 ± 2.2***	71.8 ± 3.5	56.7 ± 7.2	**62.4 ± 6.2**	57.1 ± 6.8
DBPedia	14	86.3 ± 3.8	87.7 ± 2.1	**87.9 ± 2.2**	89.8 ± 3.5	**94.3 ± 2.0***	93.5 ± 2.6	87.3 ± 4.0	**94.1 ± 2.1**	94.0 ± 2.2
NLU Scenario	18	67.8 ± 4.0	69.9 ± 3.5	**70.2 ± 4.0**	74.9 ± 3.0	**79.0 ± 2.0**	77.9 ± 3.0	78.5 ± 3.2	81.8 ± 2.0	**83.7 ± 1.8**
TREC Fine	50	39.7 ± 4.5	38.8 ± 4.7	**41.5 ± 6.0**	45.7 ± 6.7	49.6 ± 6.6	**50.1 ± 6.7**	49.7 ± 6.0	**55.5 ± 6.6**	51.7 ± 6.6
NLU Intent	68	45.3 ± 4.9	50.0 ± 4.2	**50.9 ± 4.0**	55.8 ± 3.9	62.5 ± 3.1	**63.3 ± 3.1**	61.5 ± 2.8	**71.8 ± 2.5**	**71.8 ± 2.7**
BANKING77	77	25.9 ± 4.9	24.8 ± 4.0	**28.8 ± 4.5**	43.6 ± 3.1	51.9 ± 2.8	**53.7 ± 3.3**	43.3 ± 3.4	53.0 ± 3.8	**56.0 ± 3.4**
CLINIC150	150	50.8 ± 3.0	52.4 ± 2.3	**57.7 ± 2.0**	60.4 ± 2.4	63.0 ± 1.9	**65.5 ± 1.9**	59.7 ± 2.3	65.1 ± 2.7	**66.1 ± 2.1**

Table 11: Comparative analysis of classification results measured by accuracy (in %) for OPT models with B = 3. The best scores are highlighted in bold, while '*' indicates p-value < 0.05.

Dataset	# Labels	ICL	B=3		B=4		B=5	
			PCW	NBCE	PCW	NBCE	PCW	NBCE
SST-2	2	93.8 ± 2.6	93.7 ± 3.3	**95.8 ± 1.7***	93.9 ± 2.7	**96.1 ± 0.9***	92.3 ± 4.2	**96.3 ± 0.9***
CR	2	90.3 ± 2.5	90.7 ± 2.4	**91.7 ± 1.5***	90.8 ± 2.3	**91.9 ± 1.6***	90.0 ± 2.7	**91.5 ± 1.4***
SUBJ	2	**72.3 ± 10.6***	70.9 ± 13.9	64.0 ± 10.7	**66.6 ± 13.2**	65.7 ± 9.7	67.3 ± 14.2	**68.4 ± 9.8**
CB	2	52.4 ± 10.1	**59.9 ± 12.1**	59.3 ± 10.8	55.6 ± 10.4	**59.8 ± 12.0**	**60.7 ± 8.7**	56.1 ± 9.9
RTE	2	56.1 ± 2.2	56.2 ± 1.6	**56.8 ± 2.0**	55.7 ± 1.6	**56.6 ± 2.0**	55.0 ± 1.4	**56.9 ± 1.9***
AGNews	4	74.8 ± 6.7	**76.7 ± 4.8***	72.7 ± 5.7	75.7 ± 5.3	**73.0 ± 5.6**	**77.7 ± 3.9**	77.1 ± 5.1
SST-5	5	42.7 ± 5.1	**45.2 ± 4.2**	42.5 ± 4.6	**44.3 ± 4.5***	41.3 ± 3.5	**46.3 ± 3.6***	42.8 ± 3.4
TREC	6	70.3 ± 3.3	**73.1 ± 2.2***	71.8 ± 3.5	**72.1 ± 2.9**	72.0 ± 3.4	**73.6 ± 2.7**	72.9 ± 2.9
DBPedia	14	89.8 ± 3.5	**94.3 ± 2.0***	93.5 ± 2.6	**94.4 ± 2.1**	93.4 ± 2.3	**94.7 ± 1.5***	93.7 ± 2.0
NLU Scenario	18	74.9 ± 3.0	**79.0 ± 2.0**	77.9 ± 3.0	**76.8 ± 4.3***	**76.8 ± 3.1***	77.7 ± 3.8	**79.3 ± 2.1***
TREC Fine	50	45.7 ± 6.7	49.6 ± 6.6	**50.1 ± 6.7**	48.2 ± 6.7	**49.4 ± 6.9**	**51.5 ± 6.9**	50.7 ± 5.2
NLU Intent	68	55.8 ± 3.9	62.5 ± 3.1	**63.3 ± 3.1**	61.8 ± 3.6	**62.4 ± 3.9**	61.1 ± 3.7	**66.4 ± 2.3***
BANKING77	77	43.6 ± 3.1	51.9 ± 2.8	**53.7 ± 3.3**	51.5 ± 3.2	53.8 ± 3.2	52.2 ± 2.0	**56.4 ± 2.6**
CLINIC150	150	60.4 ± 2.4	63.0 ± 1.9	**65.5 ± 1.9**	62.7 ± 2.2	**65.5 ± 2.5***	61.9 ± 1.8	**67.1 ± 2.2***

Table 12: The comparative results of context extension, measured by accuracy (in %), for OPT-6.7B models with windows (B = 4 and B = 5).

Dataset	# Labels	GPT2-Large				GPT2-XLarge			
		B = 4		B = 5		B = 4		B = 5	
		PCW	NBCE	PCW	NBCE	PCW	NBCE	PCW	NBCE
SST-2	2	83.3 ± 7.8	**83.9 ± 7.9**	**85.0 ± 6.9**	83.7 ± 8.6	91.3 ± 2.9	**92.6 ± 2.6**	91.4 ± 3.1	**92.4 ± 2.4**
CR	2	82.1 ± 5.9	**84.1 ± 5.7**	81.7 ± 4.7	**82.4 ± 5.1**	82.1 ± 2.9	**82.7 ± 3.0**	**82.0 ± 2.4**	81.7 ± 2.5
SUBJ	2	**68.1 ± 11.9**	63.1 ± 10.5	66.5 ± 10.3	**68.9 ± 10.5**	63.9 ± 6.0	**76.2 ± 6.7**	59.3 ± 5.2	**79.3 ± 5.5***
CB	2	44.0 ± 3.4	**44.7 ± 4.3**	42.8 ± 2.0	**43.8 ± 2.8**	**53.9 ± 6.2**	53.8 ± 9.1	51.1 ± 4.4	**56.7 ± 7.7***
RTE	2	**53.5 ± 1.5***	52.1 ± 3.0	54.0 ± 1.2	53.7 ± 2.2	**55.3 ± 1.1**	54.7 ± 3.0	54.9 ± 1.7	**55.7 ± 1.7**
AGNews	4	**69.2 ± 9.6**	68.1 ± 12.5	67.9 ± 8.1	**70.7 ± 8.4**	**80.5 ± 3.3***	72.5 ± 8.8	**80.0 ± 2.5***	73.0 ± 6.7
SST-5	5	40.1 ± 4.0	**42.4 ± 1.7***	40.4 ± 3.9	**42.6 ± 1.6**	**41.5 ± 4.2***	38.5 ± 5.7	39.2 ± 4.4	**41.7 ± 5.8***
TREC	6	57.4 ± 4.1	**64.8 ± 4.0***	55.3 ± 4.0	**64.6 ± 4.8***	48.9 ± 3.4	**51.6 ± 3.7**	48.1 ± 2.2	**53.0 ± 2.7***
DBPedia	14	**80.7 ± 5.0***	74.8 ± 12.1	**79.3 ± 4.4**	76.5 ± 8.4	**88.5 ± 3.3**	87.5 ± 4.7	**89.8 ± 3.2**	89.1 ± 3.6
NLU Scenario	18	27.8 ± 3.6	**46.6 ± 7.4**	27.5 ± 3.3	**44.4 ± 6.5**	49.7 ± 5.7	**51.7 ± 7.6**	48.7 ± 6.0	**52.8 ± 5.5***
TREC Fine	50	32.4 ± 5.1	**37.4 ± 4.8***	31.2 ± 4.1	**39.9 ± 3.6***	38.6 ± 3.1	**39.8 ± 6.1**	37.2 ± 2.3	**41.6 ± 3.8***
NLU Intent	68	24.3 ± 4.7	**26.0 ± 5.6**	20.3 ± 5.4	**27.3 ± 4.4**	34.8 ± 5.1	**35.9 ± 5.2**	37.1 ± 5.1	**38.6 ± 3.3***
BANKING77	77	26.6 ± 3.2	**35.2 ± 3.8***	25.5 ± 3.2	**36.0 ± 3.8***	31.0 ± 3.5	**35.4 ± 3.2***	29.6 ± 2.8	**37.7 ± 2.6***
CLINIC150	150	43.2 ± 1.8	**48.1 ± 1.9***	41.6 ± 2.2	**49.4 ± 2.0***	45.9 ± 2.9	**49.3 ± 2.3***	43.0 ± 2.4	**50.3 ± 2.5***

Table 13: The comparative results of classification tasks, quantified in terms of accuracy (in %), for both GPT2-Large and GPT2-XLarge models using different context windows (B = 4 and B = 5). The best scores for each model and dataset are highlighted in bold, while an asterisk (*) denotes statistical significance (as determined by a t-test with a p-value < 0.05).

Dataset	# Labels	OPT-1.3B				OPT-6.7B			
		B = 4		B = 5		B = 4		B = 5	
		PCW	NBCE	PCW	NBCE	PCW	NBCE	PCW	NBCE
SST-2	2	81.1 ± 7.7	**88.1 ± 5.7***	79.9 ± 9.8	**88.8 ± 5.2***	93.9 ± 2.7	**96.1 ± 0.9***	92.3 ± 4.2	**96.3 ± 0.9***
CR	2	88.5 ± 3.3	**88.8 ± 1.6**	85.6 ± 3.6	**89.1 ± 1.5***	90.8 ± 2.3	**91.9 ± 1.6***	90.0 ± 2.7	**91.5 ± 1.4***
SUBJ	2	68.5 ± 6.6	**70.5 ± 7.4**	65.2 ± 8.3	**70.9 ± 6.3***	**66.6 ± 13.2**	65.7 ± 9.7	67.3 ± 14.2	**68.4 ± 9.8**
CB	2	**51.6 ± 5.2**	51.5 ± 4.3	49.1 ± 1.0	**51.6 ± 3.6***	55.6 ± 10.4	**59.8 ± 12.0**	**60.7 ± 8.7**	56.1 ± 9.9
RTE	2	50.6 ± 3.1	**51.4 ± 2.9**	50.9 ± 2.1	**51.3 ± 2.5**	55.7 ± 1.6	**56.6 ± 2.0**	55.0 ± 1.4	**56.9 ± 1.9***
AGNews	4	**65.1 ± 5.9***	60.3 ± 9.0	**69.4 ± 5.0***	62.9 ± 6.7	75.7 ± 5.3	**73.0 ± 5.6**	**77.7 ± 3.9**	77.1 ± 5.1
SST-5	5	42.4 ± 4.0	**44.8 ± 2.2***	41.6 ± 4.3	**45.1 ± 2.0***	**44.3 ± 4.5***	41.3 ± 3.5	**46.3 ± 3.6***	42.8 ± 3.4
TREC	6	55.2 ± 3.2	**55.7 ± 4.3**	52.5 ± 2.8	**57.1 ± 3.9***	**72.1 ± 2.9**	72.0 ± 3.4	**73.6 ± 2.7**	72.9 ± 2.9
DBPedia	14	**88.1 ± 2.6**	87.5 ± 2.6	87.0 ± 3.1	**87.9 ± 2.6**	**94.4 ± 2.1**	93.4 ± 2.3	**94.7 ± 1.5***	93.7 ± 2.0
NLU Scenario	18	**69.9 ± 2.6**	69.3 ± 4.3	67.7 ± 4.0	**72.8 ± 3.8***	**76.8 ± 4.3***	**76.8 ± 3.1***	77.7 ± 3.8	**79.3 ± 2.1***
TREC Fine	50	40.5 ± 5.8	**43.1 ± 6.4**	35.3 ± 3.5	**42.0 ± 4.7***	48.2 ± 6.7	**49.4 ± 6.9**	**51.5 ± 6.9**	50.7 ± 5.2
NLU Intent	68	48.8 ± 4.2	**51.0 ± 4.7**	45.4 ± 3.2	**54.5 ± 3.3***	61.8 ± 3.6	**62.4 ± 3.9**	61.1 ± 3.7	**66.4 ± 2.3***
BANKING77	77	26.0 ± 3.5	**30.1 ± 3.5***	28.9 ± 3.1	**32.5 ± 3.5***	51.5 ± 3.2	53.8 ± 3.2	52.2 ± 2.0	**56.4 ± 2.6**
CLINIC150	150	52.6 ± 2.0	**57.2 ± 2.5***	49.3 ± 2.5	**58.4 ± 2.0***	62.7 ± 2.2	**65.5 ± 2.5***	61.9 ± 1.8	**67.1 ± 2.2***

Table 14: The comparative results of context extension, measured by accuracy (in %), for OPT models with windows (B = 4 and B = 5).

Dataset	Number of shots per window B k_{max} **GPT2**	k_{max} **LLAMA**	Prompt Example	Labels
SST-2	27	48	Sentence: {Sentence} Label: Label	[negative, positive]
CR	21	39	Review:{Sentence} Sentiment:{Label}	[negative, positive]
SUBJ	18	32	Input:{Sentence} Type:{Label}	[objective, subjective]
CB	5	10	Premise:{Sentence} Hypothesis:{ hypothesis} Prediction:{Label}	[true, false, neither]
RTE	5	10	Premise:{Sentence} Hypothesis:{ hypothesis} Prediction:{Label}	[True, False]
AGNews	11	20	Input:{Sentence} Type:{Label}	[world, sports, business, technology]
SST-5	20	36	Review:{Sentence} Sentiment:Sentiment	[terrible, bad, okay, good, great]
TREC	38	69	Question:{Sentence} Type:{Label}	[abbreviation, entity, description, human, location, numeric]
DBPedia	7	14	Input:{Sentence} Type:{Label}	[company, school, artist, athlete, politics, transportation, building, nature, village, animal, plant, album, film, book]
NLU Scenario	43	80	Utterance:{Sentence} Scenario:{Label}	[lists, weather, general, cooking, email, alarm, datetime, calendar, social, transport, iot, recommendation, takeaway, play, music, qa, news, audio]
TREC Fine	37	65	Question:{Sentence} Type:{Label}	[abbreviation abbreviation, abbreviation expansion, entity animal, entity body, entity color, entity creation, entity currency, entity disease, entity event, entity food...
NLU Intent	43	80	Utterance:{Sentence} Intent:{Label}	[alarm query, alarm remove, alarm set, audio volume down, audio volume mute, audio volume other, audio volume up, calendar query, calendar remove, calendar set...
BANKING77	27	51	Query:{Sentence} Intent:{Label}	[activate my card, age limit, apple pay or google pay, atm support, automatic top up, balance not updated after bank transfer, balance not updated after cheque or cash deposit...
CLINIC150	39	72	Sentence:{Sentence} Intent:{Label}	[restaurant reviews, nutrition info, account blocked, oil change how, time, weather, redeem rewards, interest rate, gas type...

Table 15: Classification datasets with used prompts and k_{max} for GPT2 and LLAMA. Note that OPT shares the same length of LLAMA (i.e., 2048)