kNN PROMPTING: LEARNING BEYOND THE CONTEXT WITH NEAREST NEIGHBOR INFERENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

In-Context Learning, which formulates target tasks as prompt completion conditioned on in-context demonstrations, has become the prevailing and standard utilization of large language models. In this paper, we disclose an actual predicament for this typical usage that it can not scale up with training data due to context length restrictions. We then advocate a simple and effective solution, kNN Prompting, which not only outperforms In-Context Learning under few shot scenarios, but more importantly, can scale up with as many training data as are available. kNN Prompting queries LLM with training data for distributed representations and caches them locally as anchors. At inference time, it predicts by simply aggregating nearest neighbors. We conduct comprehensive experiments and ablations across different scales of LLMs to demonstrate its substantial improvements, as well as other appealing aspects such as robustness and explainability. The proposed approach successfully bridges data scaling into model scaling, and brings new potentials for the gradient-free paradigm of LLM deployment.

1 INTRODUCTION

Large language models (LLMs), when scale up to billions of parameters, have demonstrated remarkable capabilities in a wide range of NLP tasks (Radford et al., 2019; Brown et al., 2020). However, such models are prohibitively expensive to train with most of the research- or consumer-level computational devices, though some of them are already publicly available (Zhang et al., 2022). As a result, it is now an emerging paradigm that LLMs are hosted in a remote data center while accessed by end users or applications via simple API requests[1][2]. The typical usage of LLM under such paradigm is In-Context Learning (ICL), where LLM reads and completes a prompt sequence as how it is pretrained on massive text corpora. The prompt is constructed by concatenation of several training examples and a test instance, and the prediction is obtained by mapping the LLM word continuations back to label space.

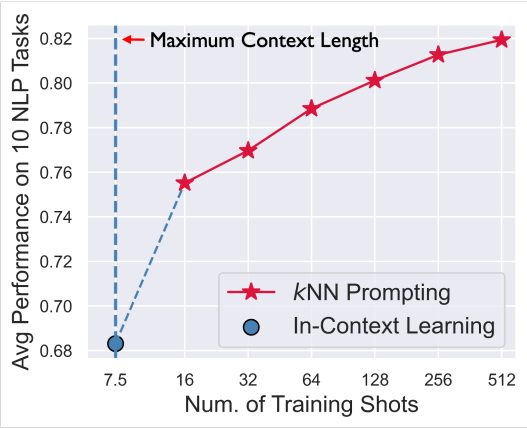


Figure 1: kNN Prompting demonstrates substantial improvements over standard ICL, and can continually scale up beyond the context. Conducted with GPT XL.

It is widely investigated and acknowledged that modern neural networks generally perform better w.r.t. increased training data. Specifically, there exists a power law between expected model performance and available data scale, which holds across various models, tasks and many orders of data scale magnitude (Hestness et al., 2017; Rosenfeld et al., 2020). For ICL, it is also empirically observed that the performance continually improves

[1] https://openai.com/api/
[2] https://gpt3demo.com/

when more training examples are prepended into the prompt Brown et al. (2020). However, such improvements are quickly prevented by the predicament of context length restriction, as language models are designed and trained to only process sequences that are no longer than a fixed length, which is in fact 1024 or 2048 tokens. For certain tasks that involve extensive passages or premises (Zhang et al., 2015; De Marneffe et al., 2019), only a handful of training examples will be allowed. This situation raises a serious challenge that for most scenarios where more than a few training data are available, we would expect further improvements, but they simply can not be put into usage, and the power law is bounded by how many examples can such a context actually contain.

In this paper, we advocate a simple and effective solution, kNN Prompting that scales ICL beyond the context length restriction and to as many training examples as are available. Specifically, we assign training examples into two sets, a demonstration set as is used in vanilla ICL, and an anchor set that exploits training data as reference. We append each anchor example into a respective prompt and query the LLM, then instead of decoding word continuations into labels, we collect the language modeling probability of LLM as distributed representation and cache it into a local datastore. At inference time, for each test instance, we similarly obtain its representation and match it against the maintained datastore to search for its k nearest neighbors.

On 10 established text classification tasks, kNN Prompting demonstrates substantial improvements over standard ICL (Figure 1): 1) Under few shot scenario where training data is very limited and fits in the context, kNN-Prompting significantly outperforms ICL already, which indicates that the anchor set yields superior data utility than in-context concatenation. 2) Under low resource or fully supervised scenario where training data can not fit in the context, kNN Prompting demonstrates its major advantage. With only 32 shots training data, it dramatically improves ICL by +13.58 in average score at its most, and the performance continual to scale up with as many training data as are available, achieving absolute improvements up to +16.96 with 128 shots training data and +18.84 under fully supervised setting.

We provide comprehensive ablation to analyze the choice of design and reasons for effectiveness. Although standard ICL has been proven poorly calibrated (Zhao et al., 2021; Jiang et al., 2021) and very unstable (Lu et al., 2022), we show that kNN Prompting is free from these risks as it no longer forces alignment between the pretrained language modeling objective and target label space. As a retrieval augmented method, kNN Prompting also offers explainability to an extent, but it still preserves efficiency. Accompanied with these appealing aspects, kNN Prompting is in general a promising solution that bridges the benefits of data scaling into model scaling to well suit the overall gradient-free paradigm of LLM deployment.

2 Background: In-Context Learning

In this section, we formulate the task and recap the ICL baseline. Assuming a target task with training data set $\mathcal{T} = \{(x_i, y_i)\}$, and Y as its categorical label space. At inference time, the model is asked to predict y_{test} given test instance x_{test}. We then denote an LLM $\boldsymbol{\theta}$ that is pretrained with a standard language modeling objective. At employment, it samely predicts a probability distribution $p(w_t|\boldsymbol{w}_{<t}, \boldsymbol{\theta})$ for the next token at t-th position conditioned on previous context $\boldsymbol{w}_{<t}$.

In-context learning first formulates training examples $\{(x_i, y_i)\}$ in the format of input-label mapping via intuitive templates (see Appendix E for illustration), and concatenates them into a natural language sequence along with the test instance to construct a prompt:

$$P = \pi(x_1, y_1) \oplus \pi(x_2, y_2) \oplus ... \oplus \pi(x_{|\mathcal{T}|}, y_{|\mathcal{T}|}) \oplus \pi(x_{\text{test}}, *) \tag{1}$$

where π denotes template-based transformation (see Appendix for illustration) and $\oplus$ denotes concatenation operation. Note that π implies a verbalization process that maps label space Y to corresponding tokens V picked from the LM vocabulary V_{LM}. When queried by the prompt P, LLM will try to mimic the prepended training examples in the context and predict a probability distribution $p(v|P, \boldsymbol{\theta})$ for the next token v. We then map it back to label space Y as prediction:

$$\hat{y}_{\text{test}} = \arg\max_{y \in Y} p(v|P, \boldsymbol{\theta}), \quad y \xrightarrow{\pi} v \tag{2}$$

We first conduct a pilot study to verify its performance with increased training examples. Figure 2 shows that when prompt P includes more available training data as demonstrations, the performance consistently improves, starting from one shot to the maximum allowed number of examples for each

	SST2	SUBJ	MPQA	AGNews	CB	CR	DBPedia	MR	RTE	TREC	AVG
Num. of Shots	53.08	28.88	98.25	8.36	7.18	34.78	1.97	30.85	12.34	19.26	**12.88**

Table 1: Maximum number of training examples (per class) allowed by In-Context Learning. Statistics are calculated using the GPT tokenizer, with context length restrictions of 2048 tokens.

task. This is in accord with the power law of data scaling investigated in many existing studies (Hestness et al., 2017; Rosenfeld et al., 2020). However, this trend is then prevented by the context length restrictions. As illustrated in the figure, GPT2 can only utilize 8 shots of training examples for SUBJ dataset, and 16 shots for SST2. We provide more comprehensive statistics in Table 4, the longest context of current LLMs, i.e., 2048 tokens, can maximumly contain an average of 12.88 training shots on the investigated 10 tasks. Although very few LM (Beltagy et al., 2020) supports even longer context up to 4096 tokens, the limitation is ultimately not removed, and such model has not been scaled up to enable a wide range of downstream applications as LLMs do. In conclusion, this situation poses an actual challenge in many scenarios where one would further collect training examples from few-shots to dozens and expect improved performance, but the power law fails.

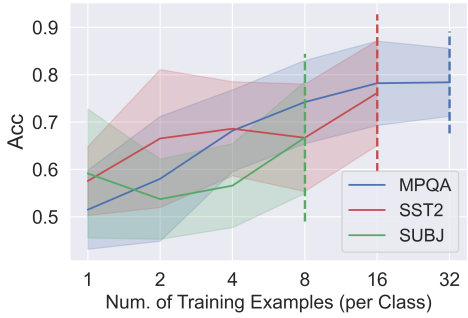


Figure 2: ICL performs better with more training examples. Dashed lines indicate the maximum number of examples allowed in the context.

3 kNN PROMPTING

In this section, we introduce the kNN Prompting framework. For a training data set $\mathcal{T}$, we split and exploit it in two respective usage: $\mathcal{T} = \mathcal{D} \cup \mathcal{A}$, i.e., a demonstration set $\mathcal{D} = \{(x_{d_i}, y_{d_i})\}$ and an anchor set $\mathcal{A} = \{(x_{a_i}, y_{a_i})\}$. kNN Prompting consists of two stages namely meta test and formal test, the overall framework is illustrated in Figure 3.

Meta Test We build a datastore that caches all anchor examples in $\mathcal{A}$ for later inference time usage. For each x_{a_i}, we respectively concatenate it into prompt P, where the prompt prefix is constructed using the demonstration set the same as Equation 1:

$$P_{a_i} = \pi(x_{d_1}, y_{d_1}) \oplus \pi(x_{d_1}, y_{d_2}) \oplus ... \oplus \pi(x_{d_{|\mathcal{D}|}}, y_{d_{|\mathcal{D}|}}) \oplus \pi(x_{a_i}, *) \tag{3}$$

By querying LLM using P_{a_i}, we similarly obtain the distribution $p(v|P_{a_i}, \boldsymbol{\theta})$. Instead of mapping it back to label space V, we cache the entire distribution over V_{LM} as the key representation for a_i:

$$\boldsymbol{k}_i = p(v|P_{a_i}, \boldsymbol{\theta}) \tag{4}$$

Then we use its label y_{a_i} as value, and construct the datastore accordingly. We denote the entire datastore as $\mathcal{M}$ consisting of paired $\mathcal{K}$ and $\mathcal{V}$.

Formal Test At inference time, for each test instance x_{test}, we construct the same prompt as Equation 1, and obtain distribution $\boldsymbol{p}_{\text{infer}} = p(v|P_{test}, \boldsymbol{\theta})$. We then match the distribution against cached $\mathcal{K}$ in the datastore, where standard KL divergence is used to measure the distance:

$$D_{KL}(\boldsymbol{p}_{\text{infer}}||\boldsymbol{k}_i) = \sum_{v \in V_{LM}} p(v|P_{test}, \boldsymbol{\theta}) \log \frac{p(v|P_{test}, \boldsymbol{\theta})}{p(v|P_{a_i}, \boldsymbol{\theta})} \tag{5}$$

The predictions are calculated by aggregating its k nearest neighbors:

$$\hat{y} = \arg\max_{y \in Y} \sum_{i \in \text{NN}^k(\boldsymbol{p}_{\text{infer}}, \mathcal{K})} \mathbb{1}(y_{a_i} = y) \tag{6}$$

where $\text{NN}^k(*, \mathcal{K})$ denotes the set of k nearest neighbors in $\mathcal{K}$.

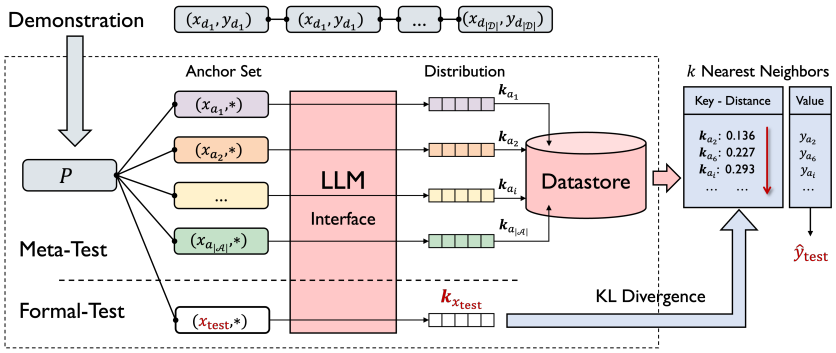


Figure 3: The overall framework of kNN Prompting

4 EXPERIMENTS

4.1 SETUP

Dataset and LLM We use 10 established text classification datasets, respectively SST2 (Socher et al., 2013), SUBJ (Pang & Lee, 2004), MPQA (Wiebe et al., 2005), AGNews (Zhang et al., 2015), CB (De Marneffe et al., 2019), CR (Hu & Liu, 2004), DBPedia (Zhang et al., 2015), MR (Pang & Lee, 2005), RTE (Dagan et al., 2005) and TREC (Voorhees & Tice, 2000). For each dataset, we devise intuitive template for ICL (Appendix E), and list other regarding statistics listed in Appendix A. We employ the most prevalent and extensively benchmarked GPT series model for experiments. We investigate a wide range of LLM scales, including GPT2 Large (0.8B), XL (1.5B) (Radford et al., 2019) and the OPT (Zhang et al., 2022) series (3B-30B) which is accessible and comparable to GPT3 (Brown et al., 2020). We use GPT XL for most analyses unless explicitly indicated.

Implementation We invariantly set the number of neighbors k to 3. There are no other hyperparameters as the entire framework is training-free. We run with 5 different random seeds for Table 2, Figure 5 and Table 3, and 10 random seeds for all other results. Mean and standard deviation are reported for robust conclusions. The largest 30B model requires 2×40GB memory for inference with model parallelism and FP16 mode. Our code will be available at `https://github.com/anonymous/knnprompting`.

4.2 DATA UTILITY

In this paper, we refer to data utility as whether and how performance scales up with increased training data. Generally, data utility can be formally depicted by the power law (Hestness et al., 2017):

$$\varepsilon(m) \propto \alpha m^{\beta} \tag{7}$$

where m is the training data size, α is a constant scaling factor, β describes the exponential steepness, and $\varepsilon(*)$ calculates the generalization error, i.e., test time performance. We refer m to training shots, i.e., number of examples per class throughout the paper. In the following section, we use data utility as the main measurement for the proposed method and investigate it with various settings of m, respectively few shot ($m \leq M_T$), low resource ($m = 32/128$), and the overall scaling law ($m \leq 1024$). We refer M_T to the maximum training shots allowed in context for each task T, and the specific statistics can be found in appendix A.

4.2.1 DATA UTILITY UNDER FEW SHOT SCENARIO

We first compare kNN Prompting to ICL under strictly comparable few shot settings, i.e., $m \leq M_T$. For kNN Prompting, we set $|\mathcal{D}| = 1$ and use all other examples for $\mathcal{A}$ to compare the data utility between $\mathcal{A}$ and P context. We leave better split strategies between $\mathcal{D}$ and $\mathcal{A}$ to Section 4.3.3.

Results in Figure 4 show that both methods can maintain data utility, but kNN prompting demonstrates a better scaling factor α. Therefore it is generally a better choice to employ training data as

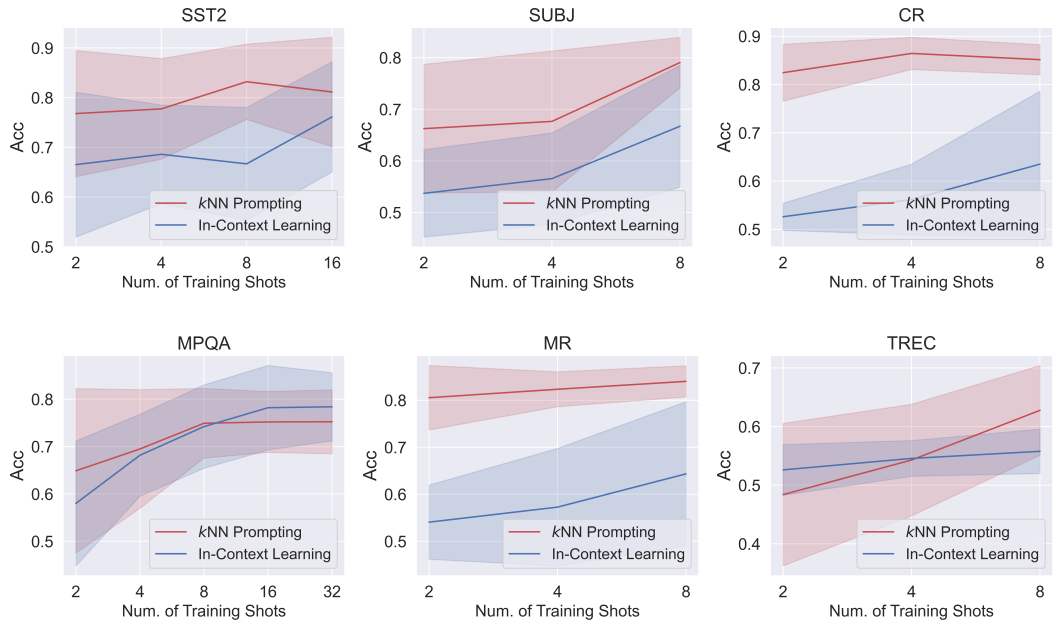


Figure 4: Results under few shot setting. kNN Prompting significantly outperforms ICL. 6 out of 10 tasks where $M_T \geq 8$ were selected to obtain more data points as well as observe the scaling trends. The curve stops at M_T for strictly fair comparison though kNN Prompting can continue to scale up.

anchor set using kNN Prompting rather than directly prepending them into context P. We discuss the reasons of effectiveness later in detail (Section 4.3.5).

Another advantage relates to computational cost. Consider the practical deployment of LLM service, the demonstration examples in P will be repeatedly computed for every test instance inference. While the anchor set is queried for only once, which is stationary and negligible for large scale inference. As kNN Prompting works well with $|\mathcal{D}| = 1$, P in ICL should be $m - 1$ demonstrations longer than P in kNN Prompting, resulting in linearly more monetary costs if measured by token numbers[3] or super-linearly more computational costs if measured by FLOPS.

4.2.2 Data Utility Beyond the Context

We then investigate a major advantage of kNN Prompting, which is scaling up to more training examples that otherwise would not fit in the context. We append M_T examples into prompt P for both methods, and see whether kNN prompting can continually scale up beyond the context via anchor set. In general, we observe significant and consistent improvements over the ICL baseline 2. By simply utilizing 32 shot of training examples, kNN Prompting brings substantial improvements of **+13.58** at its most (0.8B model, average score). The power law states that the marginal benefits of increasing data are much more lucrative when m is small, thus kNN Prompting makes great differences by re-enabling data utility from M_T to 32.

We then further increase m to 128, under this setting, the improvements over standard ICL continue to scale from +13.58 to **+16.96** at its most (0.8B model, average score). As ICL already reaches its best at $m = M_T$, we further include two baselines, respectively Finetuning and ICL Ensemble. With this amount of data, it is feasible to train a standard PLM such as BERT Large or GPT Large and produce meaningful performance. We see that kNN Prompting outperforms finetuning of standard PLM when the adopted LLM scales above **6B** (82.73). By contrast, it would require LLM to scale above **30B** (82.45) using ICL Ensemble and even larger using standard ICL.

ICL Ensemble is an intuitive alternative to scale ICL up, and has been adopted in previous works (Jiang et al., 2020). We split $\mathcal{T}$ into multiple demonstration sets $\mathcal{T} = \mathcal{D}_1 \cup \mathcal{D}_2 \cup ... \cup \mathcal{D}_N$

[3] API services like GPT3 is charged by the number of all tokens in both prompt P and generated response.

Models & Methods		SST2	SUBJ	MPQA	AGNews	CB	CR	DBPedia	MR	RTE	TREC	AVG
BERT Large FT		$88.3_{\pm 1.4}$	$90.7_{\pm 0.6}$	$74.5_{\pm 5.0}$	$88.0_{\pm 1.0}$	$78.6_{\pm 3.6}$	$88.0_{\pm 2.6}$	$95.1_{\pm 1.8}$	$83.0_{\pm 3.1}$	$58.1_{\pm 1.5}$	$78.8_{\pm 5.3}$	82.31
GPT Large FT		$90.7_{\pm 1.3}$	$86.1_{\pm 1.7}$	$87.6_{\pm 0.9}$	$88.3_{\pm 1.5}$	$70.0_{\pm 2.0}$	$86.7_{\pm 13.2}$	$96.5_{\pm 1.2}$	$86.2_{\pm 1.0}$	$55.4_{\pm 3.8}$	$71.2_{\pm 2.2}$	81.88
$m = 128$												
0.8B	**ICL Ensemble**	$63.0_{\pm 6.4}$	$57.7_{\pm 10.3}$	$69.1_{\pm 6.2}$	$67.4_{\pm 2.9}$	$41.1_{\pm 3.1}$	$83.8_{\pm 11.5}$	$67.8_{\pm 3.7}$	$72.7_{\pm 12.3}$	$55.1_{\pm 3.8}$	$59.0_{\pm 3.7}$	63.65
	***k*NN Prompting**	$84.5_{\pm 5.3}$	$85.8_{\pm 1.6}$	$83.1_{\pm 0.8}$	$84.5_{\pm 1.3}$	$62.1_{\pm 3.4}$	$89.7_{\pm 0.6}$	$95.8_{\pm 0.5}$	$84.0_{\pm 1.8}$	$53.6_{\pm 3.2}$	$74.2_{\pm 4.4}$	**79.73**
1.5B	**ICL Ensemble**	$83.4_{\pm 1.9}$	$63.4_{\pm 11.5}$	$75.0_{\pm 7.2}$	$81.1_{\pm 0.7}$	$52.9_{\pm 8.1}$	$63.4_{\pm 17.2}$	$83.7_{\pm 1.3}$	$73.8_{\pm 15.1}$	$55.9_{\pm 1.8}$	$59.6_{\pm 2.3}$	69.21
	***k*NN Prompting**	$86.3_{\pm 2.9}$	$83.8_{\pm 2.1}$	$82.3_{\pm 1.5}$	$87.2_{\pm 0.4}$	$64.6_{\pm 7.9}$	$88.9_{\pm 2.4}$	$96.5_{\pm 0.7}$	$86.4_{\pm 0.8}$	$51.1_{\pm 1.7}$	$74.0_{\pm 2.9}$	**80.12**
2.7B	**ICL Ensemble**	$90.2_{\pm 3.7}$	$77.7_{\pm 3.4}$	$86.6_{\pm 1.9}$	$78.4_{\pm 1.8}$	$54.6_{\pm 1.0}$	$93.4_{\pm 1.1}$	$89.3_{\pm 0.9}$	$92.4_{\pm 0.8}$	$54.3_{\pm 2.3}$	$78.8_{\pm 2.3}$	79.57
	***k*NN Prompting**	$93.4_{\pm 1.3}$	$87.5_{\pm 2.0}$	$83.3_{\pm 3.9}$	$86.7_{\pm 2.4}$	$57.1_{\pm 3.8}$	$90.2_{\pm 2.0}$	$98.6_{\pm 0.6}$	$91.4_{\pm 1.5}$	$53.4_{\pm 5.6}$	$80.8_{\pm 2.0}$	**82.25**
6B	**ICL Ensemble**	$93.2_{\pm 0.9}$	$85.6_{\pm 3.8}$	$87.7_{\pm 2.9}$	$71.1_{\pm 5.5}$	$45.7_{\pm 2.0}$	$93.4_{\pm 1.5}$	$90.9_{\pm 1.0}$	$92.0_{\pm 0.2}$	$54.8_{\pm 1.3}$	$70.3_{\pm 1.8}$	78.47
	***k*NN Prompting**	$92.2_{\pm 1.2}$	$87.4_{\pm 1.7}$	$85.2_{\pm 2.5}$	$87.4_{\pm 1.6}$	$63.6_{\pm 4.1}$	$90.6_{\pm 2.2}$	$98.5_{\pm 0.4}$	$91.0_{\pm 1.5}$	$55.6_{\pm 2.0}$	$75.8_{\pm 3.1}$	**82.73**
13B	**ICL Ensemble**	$88.2_{\pm 4.7}$	$90.7_{\pm 1.2}$	$78.4_{\pm 4.8}$	$82.6_{\pm 1.5}$	$62.1_{\pm 3.9}$	$94.0_{\pm 0.7}$	$94.6_{\pm 1.0}$	$89.0_{\pm 2.5}$	$56.2_{\pm 2.0}$	$59.8_{\pm 3.7}$	79.56
	***k*NN Prompting**	$94.8_{\pm 0.6}$	$90.1_{\pm 1.5}$	$86.2_{\pm 3.8}$	$87.4_{\pm 1.7}$	$78.9_{\pm 4.8}$	$89.6_{\pm 1.1}$	$98.9_{\pm 0.5}$	$92.0_{\pm 0.5}$	$58.2_{\pm 4.4}$	$73.1_{\pm 2.9}$	**84.93**
30B	**ICL Ensemble**	$92.6_{\pm 2.3}$	$84.1_{\pm 4.6}$	$79.9_{\pm 1.9}$	$78.3_{\pm 2.4}$	$67.1_{\pm 4.8}$	$88.1_{\pm 2.9}$	$93.8_{\pm 1.0}$	$93.4_{\pm 1.0}$	$65.1_{\pm 4.6}$	$82.0_{\pm 2.5}$	82.45
	***k*NN Prompting**	$94.3_{\pm 0.9}$	$92.7_{\pm 1.7}$	$84.5_{\pm 0.9}$	$87.1_{\pm 1.4}$	$70.7_{\pm 8.2}$	$91.0_{\pm 1.2}$	$98.8_{\pm 0.5}$	$93.1_{\pm 1.5}$	$61.7_{\pm 4.8}$	$79.8_{\pm 2.0}$	**85.38**
$m = 32$												
0.8B	**ICL**	$63.4_{\pm 7.3}$	$58.9_{\pm 8.7}$	$70.5_{\pm 5.2}$	$61.7_{\pm 15.4}$	$45.0_{\pm 9.1}$	$83.3_{\pm 13.7}$	$59.9_{\pm 11.5}$	$77.0_{\pm 15.7}$	$53.6_{\pm 3.1}$	$54.4_{\pm 1.7}$	62.77
	***k*NN Prompting**	$82.0_{\pm 6.9}$	$82.8_{\pm 1.2}$	$70.5_{\pm 5.2}$	$82.2_{\pm 0.9}$	$56.8_{\pm 7.3}$	$87.5_{\pm 3.0}$	$93.7_{\pm 1.1}$	$83.0_{\pm 2.5}$	$53.6_{\pm 4.0}$	$71.5_{\pm 6.2}$	**76.35**
1.5B	**ICL**	$81.3_{\pm 5.4}$	$64.1_{\pm 11.3}$	$75.2_{\pm 8.8}$	$72.7_{\pm 18.5}$	$60.7_{\pm 2.8}$	$66.2_{\pm 16.7}$	$83.5_{\pm 3.8}$	$72.2_{\pm 13.9}$	$53.0_{\pm 1.7}$	$54.2_{\pm 4.9}$	68.31
	***k*NN Prompting**	$87.7_{\pm 3.5}$	$77.0_{\pm 3.5}$	$75.2_{\pm 8.8}$	$86.2_{\pm 1.8}$	$58.9_{\pm 2.2}$	$88.2_{\pm 3.5}$	$94.1_{\pm 2.3}$	$83.9_{\pm 2.4}$	$53.6_{\pm 3.0}$	$64.8_{\pm 4.2}$	**76.97**
2.7B	**ICL**	$89.9_{\pm 4.5}$	$77.7_{\pm 5.8}$	$84.5_{\pm 2.1}$	$78.8_{\pm 4.0}$	$51.1_{\pm 6.1}$	$92.6_{\pm 0.6}$	$88.8_{\pm 1.3}$	$92.5_{\pm 1.0}$	$52.4_{\pm 4.2}$	$64.6_{\pm 3.0}$	77.29
	***k*NN Prompting**	$89.9_{\pm 4.5}$	$82.3_{\pm 4.4}$	$84.5_{\pm 2.1}$	$84.4_{\pm 1.2}$	$50.0_{\pm 5.4}$	$89.9_{\pm 3.2}$	$97.0_{\pm 0.7}$	$91.2_{\pm 2.6}$	$52.8_{\pm 1.9}$	$75.1_{\pm 5.5}$	**79.72**
6B	**ICL**	$92.7_{\pm 1.8}$	$81.6_{\pm 4.6}$	$86.2_{\pm 1.9}$	$68.8_{\pm 8.2}$	$52.5_{\pm 9.7}$	$92.0_{\pm 2.6}$	$89.8_{\pm 0.9}$	$91.6_{\pm 1.1}$	$55.0_{\pm 1.6}$	$62.0_{\pm 7.3}$	77.19
	***k*NN Prompting**	$92.7_{\pm 1.8}$	$85.5_{\pm 2.9}$	$86.2_{\pm 1.9}$	$83.0_{\pm 1.5}$	$65.0_{\pm 1.6}$	$91.0_{\pm 2.5}$	$97.4_{\pm 0.6}$	$91.8_{\pm 0.7}$	$55.2_{\pm 3.2}$	$71.2_{\pm 5.2}$	**81.90**
13B	**ICL**	$89.0_{\pm 4.3}$	$91.3_{\pm 2.4}$	$78.4_{\pm 7.2}$	$78.1_{\pm 5.6}$	$53.2_{\pm 4.4}$	$93.4_{\pm 1.1}$	$92.2_{\pm 2.4}$	$89.9_{\pm 2.2}$	$55.8_{\pm 3.0}$	$55.5_{\pm 6.1}$	77.68
	***k*NN Prompting**	$89.0_{\pm 4.3}$	$91.1_{\pm 1.1}$	$78.4_{\pm 7.2}$	$84.5_{\pm 2.2}$	$77.9_{\pm 4.1}$	$89.5_{\pm 4.6}$	$97.3_{\pm 0.7}$	$91.1_{\pm 1.4}$	$56.8_{\pm 3.2}$	$71.4_{\pm 4.1}$	**82.69**
30B	**ICL**	$90.8_{\pm 4.1}$	$83.5_{\pm 8.7}$	$80.7_{\pm 1.9}$	$74.8_{\pm 4.6}$	$64.6_{\pm 8.3}$	$87.7_{\pm 3.9}$	$93.3_{\pm 0.4}$	$93.4_{\pm 1.0}$	$61.6_{\pm 2.7}$	$71.7_{\pm 2.7}$	80.21
	***k*NN Prompting**	$90.8_{\pm 4.1}$	$91.3_{\pm 0.6}$	$80.7_{\pm 1.9}$	$84.1_{\pm 1.2}$	$70.0_{\pm 9.9}$	$89.0_{\pm 3.5}$	$98.3_{\pm 0.2}$	$94.5_{\pm 1.1}$	$59.5_{\pm 3.6}$	$77.6_{\pm 2.3}$	**83.56**

Table 2: Results under the low resource scenario ($m = \{32, 128\}$). We also include Finetuning (FT) and ICL Ensemble as baselines for $m = 128$. Best performance under each setting are **bolded**.

to construct multiple different prompts, and ensemble their predictions. This baseline provides only marginal improvements and falls behind the proposed method, leaving a performance gap of **+16.08** at worst case (0.8B model, average score). The computational costs are also a severe issue. Assuming $m = 128$, $M_T = 8$, we need to query 16 times for every test instance, and this keeps growing linearly with more training data, soon becomes prohibitively inefficient and can not scale up either.

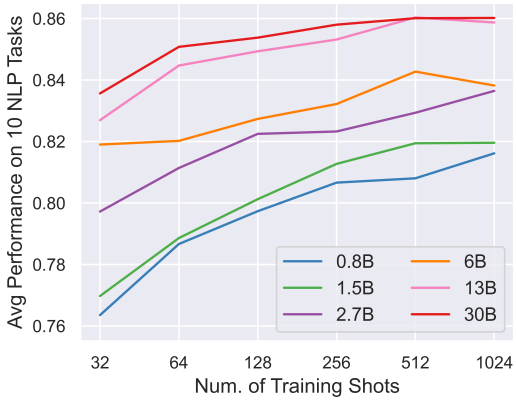


Figure 5: Results under fully supervised scenarios, scaling trends are consistently maintained across a wide range of data scales as well as LLM scales.

4.2.3 Continually Scaling Up to Thousands of Training Data

We now fully scale data up to thousands of training examples and provide extensive results across model scales to observe their scaling performance. Figure 5 shows kNN prompting can continually generalize across the tested range to provide effective data utility, re-enabling the power law under gradient-free paradigm of LLM deployment. The improvements over standard ICL further scale

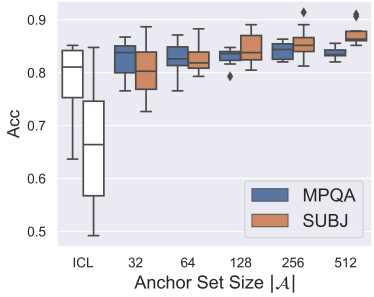


Figure 6: Robustness. White box for ICL.

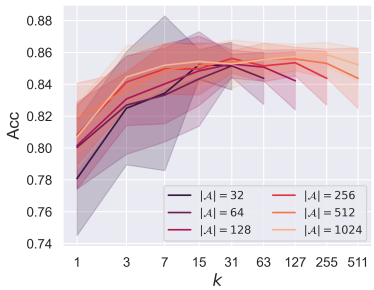


Figure 7: Performances w.r.t. different k.

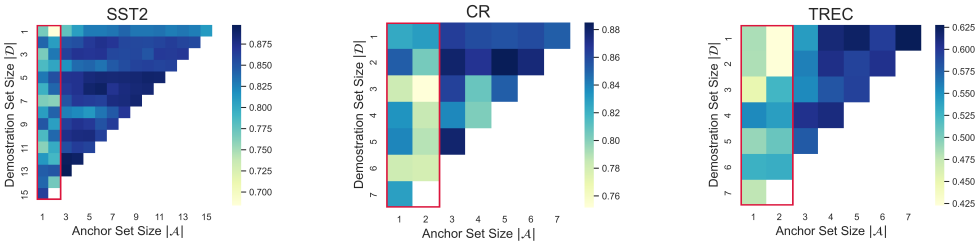


Figure 8: Split of Demonstration and Anchor Set. $|\mathcal{A}| + |\mathcal{D}| \leq M_T$.

from +16.96 to **+18.84** at its most (0.8B model, average score), and the largest model, OPT30B, achieves a best performance of 86.02. Full results can be found in Appendix B.

4.3 Ablations and Analyses

4.3.1 Robustness

It is previously found that vanilla In-Context Learning suffers from severe instability resulting from various sources of bias such as majority label, context recency, common token etc (Zhao et al., 2021). In Figure 6 we show that kNN Prompting significantly improves robustness. Besides, the results are also more robust with increasing anchor set size, as we are able to find more similar anchors and produce more reliable predictions. More broadly, we can infer from Table 2 that the standard deviation of kNN Prompting (3.83) is significantly smaller than ICL (9.14)[4].

4.3.2 Empirical Choice of k

In Figure 7 we search for different choices of k on MPQA, and found that it is generally a rather robust choice within the wide range $[3, |\mathcal{A}|/2 - 1]$.

4.3.3 Split of Demonstration Set and Anchor Set

For fully supervised scenario where $m \gg M_T$, we can simply set $|\mathcal{D}|$ to its maximum M_T. Otherwise, we might need to deliberate the trade-off between the demonstration set and anchor set. In Figure 8 we search through all possible combinations given that $|\mathcal{A}|+|\mathcal{D}| \leq M_T$. We see the left part of the heatmap generally performs inferior, i.e., $|\mathcal{A}| \in \{1, 2\}$. This means a larger $|\mathcal{A}|$ contributes more to the performance while the choice for $|\mathcal{D}|$ is relatively more robust. The conclusion also corresponds to few shot results that anchor set yields better data utility than context concatenation.

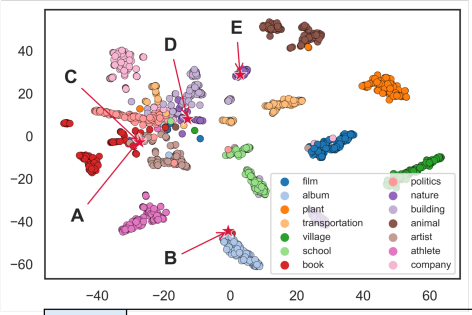
4.3.4 Distance Measurement

We investigate euclidean distance as an alternative distance measurement, which has also been explored in Khandelwal et al. (2020). We take the contextual representation $\boldsymbol{h}$ of LLM, and denotes

[4]Calculated with 32 shot setting, 0.8B model (GPT Large).

Measurements	SST2	SUBJ	AGNews	CB	CR	DBPedia	MPQA	MR	RTE	TREC	AVG
In-Context Learning	$81.3_{\pm 5.4}$	$64.1_{\pm 11.3}$	$72.7_{\pm 18.5}$	$\mathbf{60.7}_{\pm 2.8}$	$66.2_{\pm 16.7}$	$83.5_{\pm 3.8}$	$75.2_{\pm 8.8}$	$72.2_{\pm 13.9}$	$53.0_{\pm 1.7}$	$54.2_{\pm 4.9}$	68.31
Contextual Repr + D_{L2}	$84.1_{\pm 8.5}$	$73.0_{\pm 8.4}$	$82.0_{\pm 0.8}$	$60.4_{\pm 8.1}$	$78.0_{\pm 11.7}$	$\mathbf{95.1}_{\pm 1.0}$	$75.2_{\pm 8.8}$	$82.6_{\pm 5.0}$	$53.3_{\pm 2.4}$	$\mathbf{70.3}_{\pm 4.4}$	75.40 (+7.09)
LM Distribution + D_{KL}	$\mathbf{87.7}_{\pm 3.5}$	$\mathbf{77.0}_{\pm 3.5}$	$\mathbf{86.2}_{\pm 1.8}$	$58.9_{\pm 2.2}$	$\mathbf{88.2}_{\pm 3.5}$	$94.1_{\pm 2.3}$	$75.2_{\pm 8.8}$	$\mathbf{83.9}_{\pm 2.4}$	$\mathbf{53.6}_{\pm 3.0}$	$64.8_{\pm 4.2}$	**76.97** (+8.66)

Table 3: Comparison for alternative distance measurement.



	ICL	kNN Prompting
A	book (✗)	artist
B	film (✗)	album
C	book (✗)	artist
D	building (✗)	nature
E	building (✗)	nature

A	Derek Landy (born October 23 1974) is an Irish author and screenwriter best known for the Skulduggery Pleasant series of children's books.	artist	distance↓
kNN	Adrian Dawson (born 26 January 1971) is a British author of thriller and horror fiction currently best known for his 2010 debut novel Codex ... ... non-fiction rather than fiction for such subject matter.	artist	0.1355
	Michelle Harrison (born 1979 Grays Essex England) is a British author whose debut novel 13 Treasures won the Waterstone's Children's Book Prize … … Young-adult fiction book is Unrest and was released in 2012.	artist	0.2267
	Barry Kingham Oakley (born 24 February 1931) is an Australian writer.	artist	0.2932

Figure 9: t-SNE (van der Maaten & Hinton, 2008) for anchors and test cases. Cases are randomly selected given that kNN Prompting outperforms ICL. DBPedia is an ontology classification task.

their distance as $D_{L2}(\boldsymbol{h}_{\text{infer}}, \boldsymbol{h}_i)$. Table 3 shows that both methods are effective but D_{KL} (Equation 5) based on LM distribution $\boldsymbol{p}$ is a superior measurement. Actually, $\boldsymbol{p}$ is a projection of $\boldsymbol{h}$ through the word embedding, we think this procedure exploits the well-structured word embeddings of LLM to provide more disentangled representations, thus can better serves as the anchor space.

4.3.5 Qualitative Analyses: Reasons for Effectiveness and Explanability

We explain the effectiveness using visualized case study (Figure 9). Generally the representations exhibit partially clustered pattern, indicating LLM can indeed complete specific tasks using ICL though not explicitly trained on them. However, we can also identify proportional examples that get entangled with different categories and crowded together (Case A, C, D), these confusing cases are likely to cause erroneous predictions in ICL. Essentially this is because the pretraining objective is not well aligned with the label space of downstream tasks, it always tries to predict the original language modeling distribution. While kNN Prompting can alleviate the reliance on such alignment by exploiting the entire distribution. It provides fine-grained matching towards anchors and are less disturbed by the clustering quality. For a specific case, A is an abstract about a novelist and should belong to category "*artist*", but it is easily confused with category "*book*" because the context did mention books. By contrast, kNN Prompting can correctly predict by referring to similar anchors that are also about novelists and their books (as listed in the table). It does not matter whether LLM gives "*book*" or "*artist*" as completion, as long as the language modeling representations are similarly distributed with anchors, in other words, instances of same category can align with themselves. The illustration also exhibits the explainability of kNN Prompting as a further advantage.

5 Related Works

Large language models, since firstly scaled up to hundreds of billions parameters by Brown et al. (2020) and followed by several others (Rae et al., 2021; Zhang et al., 2022; Chowdhery et al., 2022; Cohen et al., 2022; Smith et al., 2022), have become the most prominent direction of NLP area. Although these models exhibit surprisingly powerful and even emergent capabilities in a wide

range of NLP tasks (Wei et al., 2022b; Hendrycks et al., 2020; Srivastava et al., 2022), they are prohibitively expensive for most researchers or users to train or even hold. In-Context Learning, which suits LLM to various tasks while requires no training, therefore becomes the typical usage as is popularized by Brown et al. (2020). Similar ideas of formulating target tasks into natural language sequences can also be found in earlier works (Trinh & Le, 2018; Raffel et al., 2020).

To better exploit LLM for various scenarios, it becomes a crucial problem to develop further augmented methods built upon ICL. Xie et al. (2022) provide theoretical explanations that formalize ICL as Bayesian inference. Sun et al. (2022) formally conceptualize the gradient-free utilization of LLM as Language-Model-as-a-Service (LMaaS). Wei et al. (2022a) and Sanh et al. (2022) propose instruction tuning, which further pretrains LLM with a collection of downstream tasks in a shared prompting format. Min et al. (2022) and Chen et al. (2022b) introduce meta-learning to better adapt LMs to ICL. Wei et al. (2022c) and Kojima et al. (2022) propose to augment the demonstrations with human-aided reasoning steps or hints, which surprisingly improved the performance for arithmetics and other reasoning tasks. Closely related to this work, Liu et al. (2022) and Rubin et al. (2022) propose to compose prompt P by retrieving training data that are most similar to the test instance, although such selection strategy still only exploits limited in-context training data at one time.

kNN is a classical machine learning algorithm (Fix & Hodges, 1989) well known for its being simple and effective, and has been applied to various scenarios (Papernot & McDaniel, 2018; Orhan, 2018). In the field of NLP, Kaiser et al. (2017) constructed a differentiable memory module for nearest neighbor searching and improved generalization for rare events. The general idea can also be applied to generation tasks Guu et al. (2018), such as dialog generation Weston et al. (2018), machine translation (Khandelwal et al., 2021), etc. Wang et al. (2022) and Chen et al. (2022a) propose to retrieve similar training examples and incorporate them into the input to jointly train the model. Besides retrieval of training data, Khandelwal et al. (2020) and Shi et al. (2022) consider unsupervised corpus as datastore for language modeling probabilities, and interpolate them with the current step distribution. The retrieved corpus can also serve as references for knowledge intensive tasks, but the retriever would need explicitly training for such purpose Lewis et al. (2020); Borgeaud et al. (2022); Izacard et al. (2022). Different from these works, kNN Prompting is suitably situated in the gradient-free paradigm of LLM deployment, bridging data scaling into model scaling.

6 Discussion

Under the existing ICL paradigm, it is often impossible to take advantage of both the capability of LLM and the data utility of finetuning, i.e., model scaling and data scaling. kNN Prompting finds an effective solution to promise them both. Nevertheless, we assume its data utility should still be inferior to the specialized finetuning of LLMs, if given sufficient computation resources in an ideal setting. We believe that it is a very important and promising direction to further approach this upper-bound and expect to raise more interests in future works.

A potential concern for retrieval-based models is their efficiency, especially when corpus level datastore is utilized. kNN Prompting is free of such concerns as it considers training data, which is a manageable scale. Besides, anchor representations neither participate in the within-model computation nor trained jointly on-the-fly. kNN Prompting even reduces computational costs at deployment time compared to standard ICL, and the efficiency advantage is more considerable if compared to ICL Ensemble, as discussed earlier in Section 4.2.1 and Section 4.2.2.

7 Conclusion

In this paper, we reveal the context length predicament of ICL and propose kNN Prompting as a significantly effective solution that scales ICL up to as many training data as are available, and successfully bridges the utility of data scaling into model scaling under the gradient-free paradigm of LLM deployment. We demonstrate the superiority of kNN Prompting under various scenarios, it not only significantly outperforms ICL in strictly comparable few-shot setting, but also brings more substantial improvements by scaling up with training data. We also analyze other appealing aspects including its robustness, efficiency and explainability. The proposed framework endeavors to realize more effective, efficient and applicable utilization of large language models in realistic scenarios, and hopefully could inspire further research interests toward the same goal.

ETHICAL CONSIDERATIONS

This work is built upon the ICL paradigm and involves querying LLM for responses. These models might generate contents with potential ethical risks regarding fairness and bias (Bommasani et al., 2021; Blodgett et al., 2020), depending on specific downstream tasks. Although the scope of this paper remains on how to better exploit LLM for task performance, it is worth further discussion to combine the proposed framework in conjunction with well-established methods that can measure (Nadeem et al., 2021) and mitigate (Nadeem et al., 2021; Gupta et al., 2022) such ethical risks.

REFERENCES

Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv:2004.05150*, 2020.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5454–5476, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.485. URL https://aclanthology.org/2020.acl-main.485.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen Creel, Jared Quincy Davis, Dorottya Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, and et al. On the opportunities and risks of foundation models. *CoRR*, abs/2108.07258, 2021. URL https://arxiv.org/abs/2108.07258.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego De Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack Rae, Erich Elsen, and Laurent Sifre. Improving language models by retrieving from trillions of tokens. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 2206–2240. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/borgeaud22a.html.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Xiang Chen, Lei Li, Ningyu Zhang, Xiaozhuan Liang, Shumin Deng, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. Decoupling knowledge from memorization: Retrieval-augmented prompt learning. *arXiv preprint arXiv:2205.14704*, 2022a.

Yanda Chen, Ruiqi Zhong, Sheng Zha, George Karypis, and He He. 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)*, pp. 719–730, Dublin, Ireland, May

2022b. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.53. URL https://aclanthology.org/2022.acl-long.53.

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

Aaron Daniel Cohen, Adam Roberts, Alejandra Molina, Alena Butryna, Alicia Jin, Apoorv Kulshreshtha, Ben Hutchinson, Ben Zevenbergen, Blaise Hilary Aguera-Arcas, Chung-ching Chang, et al. Lamda: Language models for dialog applications. 2022.

Ido Dagan, Oren Glickman, and Bernardo Magnini. The pascal recognising textual entailment challenge. In *Proceedings of the First International Conference on Machine Learning Challenges: Evaluating Predictive Uncertainty Visual Object Classification, and Recognizing Textual Entailment*, MLCW'05, pp. 177–190, Berlin, Heidelberg, 2005. Springer-Verlag. ISBN 3540334270. doi: 10.1007/11736790_9. URL https://doi.org/10.1007/11736790_9.

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

Evelyn Fix and Joseph Lawson Hodges. Discriminatory analysis. nonparametric discrimination: Consistency properties. *International Statistical Review/Revue Internationale de Statistique*, 57 (3):238–247, 1989.

Umang Gupta, Jwala Dhamala, Varun Kumar, Apurv Verma, Yada Pruksachatkun, Satyapriya Krishna, Rahul Gupta, Kai-Wei Chang, Greg Ver Steeg, and Aram Galstyan. Mitigating gender bias in distilled language models via counterfactual role reversal. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 658–678, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.55. URL https://aclanthology.org/2022.findings-acl.55.

Kelvin Guu, Tatsunori B. Hashimoto, Yonatan Oren, and Percy Liang. Generating sentences by editing prototypes. *Transactions of the Association for Computational Linguistics*, 6:437–450, 2018. doi: 10.1162/tacl_a_00030. URL https://aclanthology.org/Q18-1031.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md Patwary, Mostofa Ali, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*, 2017.

Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pp. 168–177, New York, NY, USA, 2004. Association for Computing Machinery. ISBN 1581138881. doi: 10.1145/1014052.1014073. URL https://doi.org/10.1145/1014052.1014073.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. Few-shot learning with retrieval augmented language models. *arXiv preprint arXiv:2208.03299*, 2022.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438, 2020. doi: 10.1162/tacl_a_00324. URL https://aclanthology.org/2020.tacl-1.28.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977, 2021. doi: 10.1162/tacl_a_00407. URL https://aclanthology.org/2021.tacl-1.57.

Lukasz Kaiser, Ofir Nachum, Aurko Roy, and Samy Bengio. Learning to remember rare events. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=SJTQLdqlg.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Generalization through memorization: Nearest neighbor language models. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HklBjCEKvH.

Urvashi Khandelwal, Angela Fan, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Nearest neighbor machine translation. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=7wCBOfJ8hJM.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *ICML 2022 Workshop on Knowledge Retrieval and Language Models*, 2022. URL https://openreview.net/forum?id=6p3AuaHAFiN.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pp. 100–114, Dublin, Ireland and Online, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.deelio-1.10. URL https://aclanthology.org/2022.deelio-1.10.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 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)*, pp. 8086–8098, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.556. URL https://aclanthology.org/2022.acl-long.556.

Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. 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*, pp. 2791–2809, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.201. URL https://aclanthology.org/2022.naacl-main.201.

Moin Nadeem, Anna Bethke, and Siva Reddy. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 5356–5371, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.416. URL https://aclanthology.org/2021.acl-long.416.

Emin Orhan. A simple cache model for image recognition. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/6e0917469214d8fbd8c517dcdc6b8dcf-Paper.pdf.

Bo Pang and Lillian Lee. 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 (ACL-04)*, pp. 271–278, Barcelona, Spain, July 2004. doi: 10.3115/1218955.1218990. URL https://aclanthology.org/P04-1035.

Bo Pang and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pp. 115–124, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. doi: 10.3115/1219840.1219855. URL https://aclanthology.org/P05-1015.

Nicolas Papernot and Patrick McDaniel. Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning. *arXiv preprint arXiv:1803.04765*, 2018.

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

Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*, 2021.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

Jonathan S. Rosenfeld, Amir Rosenfeld, Yonatan Belinkov, and Nir Shavit. A constructive prediction of the generalization error across scales. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=ryenvpEKDr.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 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*, pp. 2655–2671, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.191. URL https://aclanthology.org/2022.naacl-main.191.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M Rush. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=9Vrb9D0WI4.

Timo Schick and Hinrich Schütze. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 255–269, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.20. URL https://aclanthology.org/2021.eacl-main.20.

Weijia Shi, Julian Michael, Suchin Gururangan, and Luke Zettlemoyer. Nearest neighbor zero-shot inference. *arXiv preprint arXiv:2205.13792*, 2022.

Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, et al. Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model. *arXiv preprint arXiv:2201.11990*, 2022.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1631–1642, Seattle, Washington, USA, October 2013. Association for Computational Linguistics. URL https://aclanthology.org/D13-1170.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022.

Tianxiang Sun, Yunfan Shao, Hong Qian, Xuanjing Huang, and Xipeng Qiu. Black-box tuning for language-model-as-a-service. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 20841–20855. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/sun22e.html.

Trieu H Trinh and Quoc V Le. A simple method for commonsense reasoning. *arXiv preprint arXiv:1806.02847*, 2018.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. URL http://jmlr.org/papers/v9/vandermaaten08a.html.

Ellen M. Voorhees and Dawn M. Tice. Building a question answering test collection. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '00, pp. 200–207, New York, NY, USA, 2000. Association for Computing Machinery. ISBN 1581132263. doi: 10.1145/345508.345577. URL https://doi.org/10.1145/345508.345577.

Shuohang Wang, Yichong Xu, Yuwei Fang, Yang Liu, Siqi Sun, Ruochen Xu, Chenguang Zhu, and Michael Zeng. Training data is more valuable than you think: A simple and effective method by retrieving from training data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3170–3179, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.226. URL https://aclanthology.org/2022.acl-long.226.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*, 2022a. URL https://openreview.net/forum?id=gEZrGCozdqR.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022b. URL https://openreview.net/forum?id=yzkSU5zdwD. Survey Certification.

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

Jason Weston, Emily Dinan, and Alexander Miller. Retrieve and refine: Improved sequence generation models for dialogue. In *Proceedings of the 2018 EMNLP Workshop SCAI: The 2nd International Workshop on Search-Oriented Conversational AI*, pp. 87–92, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5713. URL https://www.aclweb.org/anthology/W18-5713.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2):165–210, 2005.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit bayesian inference. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=RdJVFCHjUMI.

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

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper/2015/file/250cf8b51c773f3f8dc8b4be867a9a02-Paper.pdf.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12697–12706. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/zhao21c.html.

A Dataset Statistics: Max Shot In Context

We provide detailed statistics about the number of maximum shots, i.e., M_T for each task (Table 4).

	SST2	**SUBJ**	**MPQA**	**AGNews**	**CB**	**CR**	**DBPedia**	**MR**	**RTE**	**TREC**
Num. of Classes	2	2	2	4	3	2	14	2	2	6
Average length of Templates	19.1	34.9	10.4	59.5	90.8	29.0	71.6	32.7	79.8	17.6
Num. of Shots (1024 tok)	16	8	32	2	2	8	1	8	4	8
Num. of Shots (2048 tok)	32	16	32	4	4	16	1	16	8	16

Table 4: Settings for In-Context Learning in this paper. This is the maximum shots (per class) that a context of 1024 tokens or 2048 tokens can allow. We set them from $\{1, 2, 4, 8, 16, 32\}$ for simplicity.

B Full Results for Data Scaling

We provide comprehensive results of kNN Prompting across data scales and LLM scales in 5, corresponding to Figure 5 in the main manuscript.

C Comparison to FT

We compare kNN Prompting to standard PLM finetuning in a more extensive data scale (Table 5). For finetuning baselines in both Table 5 and Table 2, we set hyper-parameters following previous works (Schick & Schütze, 2021). We set learning rate to 1e-5, batch size to 16, and training steps to $\{125, 250, 500\}$ respectively for $m \in 32, 64, 128, 256, 512, 1024$. For CB, AGNews and RTE, batch size is adjusted to 8, for DBPedia, batch size is adjusted to 4 to avoid OOM. We observe that with the same model scale, kNN Prompting is superior than finetuning under the low resource setting, and inferior under fully supervised setting. This indicates its data utility factor α is still smaller than finetuning. However, the main advantage of kNN Prompting comes with LLM, which significantly outperforms the finetuning baseline. We have also discussed this in Section 6.

D More Cases

We provide more case study, respectively on SST2 (Figure 11), AGNews (Figure 12), and TREC (Figure 13).

E Prompt Template

See Table 6 for the used templates (Adopted from Lu et al. (2022)). They are intuitively designed and the proposed method should be robust with choices of templates.

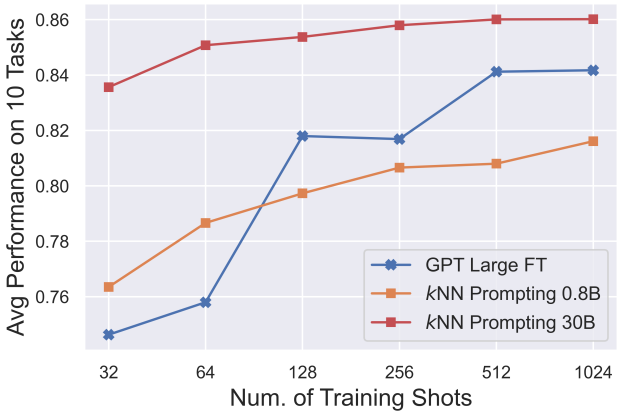
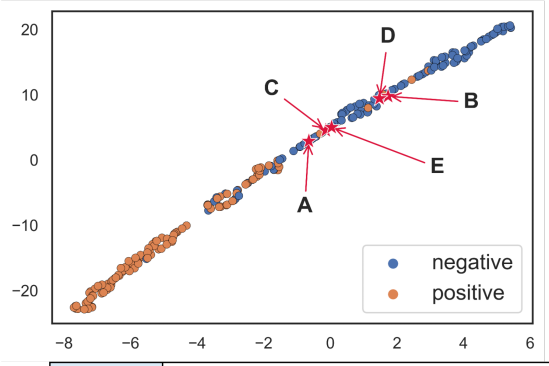


Figure 10: Comparison to finetuning baseline.

Models & Methods		SST2	SUBJ	MPQA	AGNews	CB	CR	DBPedia	MR	RTE	TREC	AVG
0.8B	In-Context Learning	$63.4_{\pm 7.3}$	$58.9_{\pm 8.7}$	$70.5_{\pm 5.2}$	$61.7_{\pm 15.4}$	$45.0_{\pm 9.1}$	$83.3_{\pm 13.7}$	$59.9_{\pm 11.5}$	$77.0_{\pm 15.7}$	$53.6_{\pm 3.1}$	$54.4_{\pm 1.7}$	62.77
	$m = 32$	$82.0_{\pm 6.9}$	$82.8_{\pm 1.2}$	$70.5_{\pm 5.2}$	$82.2_{\pm 0.9}$	$56.8_{\pm 7.3}$	$87.5_{\pm 3.0}$	$93.7_{\pm 1.1}$	$83.0_{\pm 2.5}$	$53.6_{\pm 4.0}$	$71.5_{\pm 6.2}$	76.35
	$m = 64$	$84.2_{\pm 5.4}$	$84.6_{\pm 0.7}$	$81.2_{\pm 3.0}$	$83.3_{\pm 1.9}$	$59.3_{\pm 6.0}$	$90.3_{\pm 1.4}$	$95.2_{\pm 0.8}$	$82.2_{\pm 2.9}$	$53.8_{\pm 1.7}$	$72.4_{\pm 4.1}$	78.66
	$m = 128$	$84.5_{\pm 5.3}$	$85.8_{\pm 1.6}$	$83.1_{\pm 0.8}$	$84.5_{\pm 1.3}$	$62.1_{\pm 3.4}$	$89.7_{\pm 0.6}$	$95.8_{\pm 0.5}$	$84.0_{\pm 1.8}$	$53.6_{\pm 3.2}$	$74.2_{\pm 4.4}$	79.73
	$m = 256$	$85.8_{\pm 4.0}$	$86.0_{\pm 2.5}$	$83.2_{\pm 1.3}$	$86.5_{\pm 1.3}$	$62.1_{\pm 3.4}$	$89.5_{\pm 0.4}$	$96.1_{\pm 0.9}$	$82.8_{\pm 0.9}$	$53.5_{\pm 3.3}$	$81.0_{\pm 2.5}$	80.66
	$m = 512$	$85.5_{\pm 4.2}$	$86.7_{\pm 1.5}$	$83.0_{\pm 2.4}$	$86.3_{\pm 0.7}$	$62.1_{\pm 3.4}$	$89.1_{\pm 1.0}$	$96.4_{\pm 0.8}$	$83.5_{\pm 4.1}$	$52.8_{\pm 3.4}$	$82.4_{\pm 2.9}$	80.80
	$m = 1024$	$85.5_{\pm 4.5}$	$87.6_{\pm 1.4}$	$84.8_{\pm 1.0}$	$87.6_{\pm 0.6}$	$62.1_{\pm 3.4}$	$87.5_{\pm 1.0}$	$96.7_{\pm 0.7}$	$83.6_{\pm 2.3}$	$54.1_{\pm 2.6}$	$86.7_{\pm 1.5}$	81.61
1.5B	In-Context Learning	$81.3_{\pm 5.4}$	$64.1_{\pm 11.3}$	$75.2_{\pm 8.8}$	$72.7_{\pm 18.5}$	$60.7_{\pm 2.8}$	$66.2_{\pm 16.7}$	$83.5_{\pm 3.8}$	$72.2_{\pm 13.9}$	$53.0_{\pm 1.7}$	$54.2_{\pm 4.9}$	68.31
	$m = 32$	$87.7_{\pm 3.5}$	$77.0_{\pm 3.5}$	$75.2_{\pm 8.8}$	$86.2_{\pm 1.8}$	$58.9_{\pm 2.2}$	$88.2_{\pm 3.5}$	$94.1_{\pm 2.3}$	$83.9_{\pm 2.4}$	$53.6_{\pm 3.0}$	$64.8_{\pm 4.2}$	76.97
	$m = 64$	$87.2_{\pm 2.8}$	$82.5_{\pm 1.2}$	$83.0_{\pm 4.0}$	$86.2_{\pm 1.1}$	$62.9_{\pm 6.1}$	$89.1_{\pm 1.6}$	$95.5_{\pm 1.4}$	$84.3_{\pm 1.9}$	$49.1_{\pm 3.2}$	$68.9_{\pm 2.9}$	78.86
	$m = 128$	$86.3_{\pm 2.9}$	$83.8_{\pm 2.1}$	$82.3_{\pm 1.5}$	$87.2_{\pm 0.4}$	$64.6_{\pm 7.9}$	$88.9_{\pm 2.4}$	$96.5_{\pm 0.7}$	$86.4_{\pm 0.8}$	$51.1_{\pm 1.7}$	$74.0_{\pm 2.9}$	80.12
	$m = 256$	$89.1_{\pm 1.6}$	$85.0_{\pm 2.2}$	$83.4_{\pm 1.5}$	$87.5_{\pm 1.7}$	$64.6_{\pm 7.9}$	$89.3_{\pm 0.9}$	$97.3_{\pm 0.3}$	$85.6_{\pm 2.0}$	$52.7_{\pm 3.3}$	$78.1_{\pm 3.3}$	81.27
	$m = 512$	$89.0_{\pm 1.5}$	$87.6_{\pm 1.6}$	$83.7_{\pm 1.2}$	$88.4_{\pm 1.5}$	$64.6_{\pm 7.9}$	$90.1_{\pm 1.0}$	$97.6_{\pm 0.5}$	$85.1_{\pm 1.6}$	$52.9_{\pm 1.8}$	$80.5_{\pm 1.8}$	81.94
	$m = 1024$	$88.2_{\pm 1.0}$	$88.4_{\pm 1.8}$	$84.1_{\pm 1.2}$	$89.1_{\pm 1.2}$	$64.6_{\pm 7.9}$	$86.7_{\pm 1.3}$	$97.8_{\pm 0.6}$	$84.4_{\pm 1.5}$	$53.3_{\pm 4.2}$	$83.0_{\pm 1.4}$	81.96
2.7B	In-Context Learning	$89.9_{\pm 4.5}$	$77.7_{\pm 5.8}$	$84.5_{\pm 2.1}$	$78.8_{\pm 4.0}$	$51.1_{\pm 6.1}$	$92.6_{\pm 0.6}$	$88.8_{\pm 1.3}$	$92.5_{\pm 1.0}$	$52.4_{\pm 4.2}$	$64.6_{\pm 3.0}$	77.29
	$m = 32$	$89.9_{\pm 4.5}$	$82.3_{\pm 4.4}$	$84.5_{\pm 2.1}$	$84.4_{\pm 1.2}$	$50.0_{\pm 5.4}$	$89.9_{\pm 3.2}$	$97.0_{\pm 0.7}$	$91.2_{\pm 2.6}$	$52.8_{\pm 1.9}$	$75.1_{\pm 5.5}$	79.72
	$m = 64$	$91.4_{\pm 2.1}$	$84.9_{\pm 3.7}$	$80.8_{\pm 6.0}$	$86.2_{\pm 1.7}$	$57.1_{\pm 3.3}$	$91.0_{\pm 1.7}$	$97.7_{\pm 0.8}$	$91.5_{\pm 1.9}$	$51.6_{\pm 3.4}$	$79.1_{\pm 2.8}$	81.14
	$m = 128$	$93.4_{\pm 1.3}$	$87.5_{\pm 2.0}$	$83.3_{\pm 3.9}$	$86.7_{\pm 2.4}$	$57.1_{\pm 3.8}$	$90.2_{\pm 2.0}$	$98.6_{\pm 0.6}$	$91.4_{\pm 1.5}$	$53.4_{\pm 5.6}$	$80.8_{\pm 2.0}$	82.25
	$m = 256$	$92.9_{\pm 2.0}$	$87.7_{\pm 1.7}$	$83.3_{\pm 2.4}$	$86.4_{\pm 1.3}$	$57.1_{\pm 3.8}$	$91.6_{\pm 1.5}$	$99.0_{\pm 0.2}$	$90.5_{\pm 1.7}$	$53.5_{\pm 3.2}$	$81.2_{\pm 2.5}$	82.32
	$m = 512$	$91.9_{\pm 1.0}$	$89.4_{\pm 1.0}$	$83.8_{\pm 2.3}$	$88.1_{\pm 2.1}$	$57.1_{\pm 3.8}$	$91.3_{\pm 1.2}$	$99.1_{\pm 0.4}$	$91.2_{\pm 1.4}$	$54.3_{\pm 1.6}$	$83.1_{\pm 4.0}$	82.93
	$m = 1024$	$91.7_{\pm 2.3}$	$90.5_{\pm 1.5}$	$83.7_{\pm 2.2}$	$89.1_{\pm 1.2}$	$57.1_{\pm 3.8}$	$89.8_{\pm 1.0}$	$99.1_{\pm 0.3}$	$90.9_{\pm 1.2}$	$57.7_{\pm 3.0}$	$86.7_{\pm 2.3}$	83.64
6B	In-Context Learning	$92.7_{\pm 1.8}$	$81.6_{\pm 4.6}$	$86.2_{\pm 1.9}$	$68.8_{\pm 8.2}$	$52.5_{\pm 9.7}$	$92.0_{\pm 2.6}$	$89.8_{\pm 0.9}$	$91.6_{\pm 1.1}$	$55.0_{\pm 1.6}$	$62.0_{\pm 7.3}$	77.19
	$m = 32$	$92.7_{\pm 1.8}$	$85.5_{\pm 2.9}$	$86.2_{\pm 1.9}$	$83.0_{\pm 1.5}$	$65.0_{\pm 1.6}$	$91.0_{\pm 2.5}$	$97.4_{\pm 0.6}$	$91.8_{\pm 0.7}$	$55.2_{\pm 3.2}$	$71.2_{\pm 5.2}$	81.90
	$m = 64$	$92.7_{\pm 0.8}$	$87.2_{\pm 2.0}$	$85.2_{\pm 3.9}$	$85.1_{\pm 1.4}$	$66.4_{\pm 6.6}$	$90.5_{\pm 2.8}$	$98.1_{\pm 0.5}$	$91.1_{\pm 1.4}$	$51.6_{\pm 4.4}$	$72.3_{\pm 1.3}$	82.02
	$m = 128$	$92.2_{\pm 1.2}$	$87.4_{\pm 1.7}$	$85.2_{\pm 2.5}$	$87.4_{\pm 1.6}$	$63.6_{\pm 4.1}$	$90.6_{\pm 2.2}$	$98.5_{\pm 0.4}$	$91.0_{\pm 1.5}$	$55.6_{\pm 2.0}$	$75.8_{\pm 3.1}$	82.73
	$m = 256$	$92.9_{\pm 0.8}$	$90.5_{\pm 1.2}$	$84.9_{\pm 0.9}$	$86.4_{\pm 2.0}$	$63.6_{\pm 4.1}$	$90.7_{\pm 0.7}$	$99.0_{\pm 0.2}$	$91.4_{\pm 1.0}$	$56.2_{\pm 2.4}$	$76.6_{\pm 1.5}$	83.22
	$m = 512$	$93.5_{\pm 1.3}$	$90.8_{\pm 0.9}$	$85.5_{\pm 1.7}$	$88.6_{\pm 0.8}$	$63.6_{\pm 4.1}$	$90.4_{\pm 1.4}$	$99.0_{\pm 0.2}$	$90.3_{\pm 1.0}$	$59.4_{\pm 2.0}$	$81.6_{\pm 1.4}$	84.27
	$m = 1024$	$93.0_{\pm 1.1}$	$90.6_{\pm 1.6}$	$85.5_{\pm 1.4}$	$89.5_{\pm 1.7}$	$63.6_{\pm 4.1}$	$88.3_{\pm 1.6}$	$98.9_{\pm 0.5}$	$90.4_{\pm 1.5}$	$57.0_{\pm 2.7}$	$81.5_{\pm 0.9}$	83.82
13B	In-Context Learning	$89.0_{\pm 4.3}$	$91.3_{\pm 2.4}$	$78.4_{\pm 7.2}$	$78.1_{\pm 5.6}$	$53.2_{\pm 4.4}$	$93.4_{\pm 1.1}$	$92.2_{\pm 2.4}$	$89.9_{\pm 2.2}$	$55.8_{\pm 3.0}$	$55.5_{\pm 6.1}$	77.68
	$m = 32$	$89.0_{\pm 4.3}$	$91.1_{\pm 1.1}$	$78.4_{\pm 7.2}$	$84.5_{\pm 2.2}$	$77.9_{\pm 4.1}$	$89.5_{\pm 4.6}$	$97.3_{\pm 0.7}$	$91.1_{\pm 1.4}$	$56.8_{\pm 3.2}$	$71.4_{\pm 4.1}$	82.69
	$m = 64$	$94.8_{\pm 0.8}$	$90.8_{\pm 2.1}$	$86.3_{\pm 2.7}$	$86.0_{\pm 1.1}$	$77.5_{\pm 5.1}$	$91.8_{\pm 1.1}$	$98.1_{\pm 0.7}$	$91.1_{\pm 2.1}$	$57.7_{\pm 4.1}$	$70.5_{\pm 2.7}$	84.47
	$m = 128$	$94.8_{\pm 0.6}$	$90.1_{\pm 1.5}$	$86.2_{\pm 3.8}$	$87.4_{\pm 1.7}$	$78.9_{\pm 4.8}$	$89.6_{\pm 1.1}$	$98.9_{\pm 0.5}$	$92.0_{\pm 0.5}$	$58.2_{\pm 4.4}$	$73.1_{\pm 2.9}$	84.93
	$m = 256$	$94.5_{\pm 0.9}$	$90.5_{\pm 1.6}$	$85.3_{\pm 0.7}$	$87.2_{\pm 1.8}$	$78.9_{\pm 4.8}$	$91.2_{\pm 1.9}$	$99.5_{\pm 0.2}$	$92.0_{\pm 1.4}$	$59.3_{\pm 1.5}$	$74.8_{\pm 4.1}$	85.31
	$m = 512$	$94.5_{\pm 0.8}$	$91.2_{\pm 2.1}$	$86.3_{\pm 0.9}$	$86.9_{\pm 1.5}$	$78.9_{\pm 4.8}$	$90.5_{\pm 1.3}$	$99.3_{\pm 0.3}$	$92.7_{\pm 0.3}$	$64.1_{\pm 4.8}$	$75.9_{\pm 4.1}$	86.03
	$m = 1024$	$94.5_{\pm 0.4}$	$91.6_{\pm 2.4}$	$87.0_{\pm 2.1}$	$88.5_{\pm 1.0}$	$78.9_{\pm 4.8}$	$88.1_{\pm 1.3}$	$99.3_{\pm 0.2}$	$91.6_{\pm 0.3}$	$63.0_{\pm 3.6}$	$76.1_{\pm 6.2}$	85.87
30B	In-Context Learning	$90.8_{\pm 4.1}$	$83.5_{\pm 8.7}$	$80.7_{\pm 1.9}$	$74.8_{\pm 4.6}$	$64.6_{\pm 8.3}$	$87.7_{\pm 3.9}$	$93.3_{\pm 0.4}$	$93.4_{\pm 1.0}$	$61.6_{\pm 2.7}$	$71.7_{\pm 2.7}$	80.21
	$m = 32$	$90.8_{\pm 4.1}$	$91.3_{\pm 0.6}$	$80.7_{\pm 1.9}$	$84.1_{\pm 1.2}$	$70.0_{\pm 9.9}$	$89.0_{\pm 3.5}$	$98.3_{\pm 0.2}$	$94.5_{\pm 1.1}$	$59.5_{\pm 3.6}$	$77.6_{\pm 2.3}$	83.56
	$m = 64$	$93.6_{\pm 1.6}$	$91.9_{\pm 1.4}$	$85.5_{\pm 1.8}$	$85.3_{\pm 2.1}$	$69.3_{\pm 7.6}$	$91.6_{\pm 1.6}$	$98.3_{\pm 0.2}$	$94.1_{\pm 0.6}$	$60.2_{\pm 6.0}$	$81.0_{\pm 2.9}$	85.08
	$m = 128$	$94.3_{\pm 0.9}$	$92.7_{\pm 1.7}$	$84.5_{\pm 0.9}$	$87.1_{\pm 1.4}$	$70.7_{\pm 8.2}$	$91.0_{\pm 1.2}$	$98.8_{\pm 0.5}$	$93.1_{\pm 1.5}$	$61.7_{\pm 4.8}$	$79.8_{\pm 2.0}$	85.38
	$m = 256$	$94.4_{\pm 0.4}$	$93.8_{\pm 1.1}$	$84.1_{\pm 1.1}$	$87.6_{\pm 0.9}$	$70.7_{\pm 8.2}$	$90.3_{\pm 1.0}$	$99.1_{\pm 0.3}$	$93.7_{\pm 0.6}$	$62.1_{\pm 2.5}$	$82.1_{\pm 1.9}$	85.80
	$m = 512$	$94.1_{\pm 0.3}$	$94.1_{\pm 1.4}$	$85.6_{\pm 1.5}$	$87.9_{\pm 1.9}$	$70.7_{\pm 8.2}$	$89.8_{\pm 1.5}$	$99.1_{\pm 0.3}$	$93.2_{\pm 1.1}$	$61.8_{\pm 2.6}$	$83.8_{\pm 1.8}$	86.01
	$m = 1024$	$94.1_{\pm 0.6}$	$93.9_{\pm 0.8}$	$85.4_{\pm 0.7}$	$87.9_{\pm 2.3}$	$70.7_{\pm 8.2}$	$88.4_{\pm 1.1}$	$99.1_{\pm 0.2}$	$93.4_{\pm 1.6}$	$61.0_{\pm 4.0}$	$86.2_{\pm 1.5}$	86.02

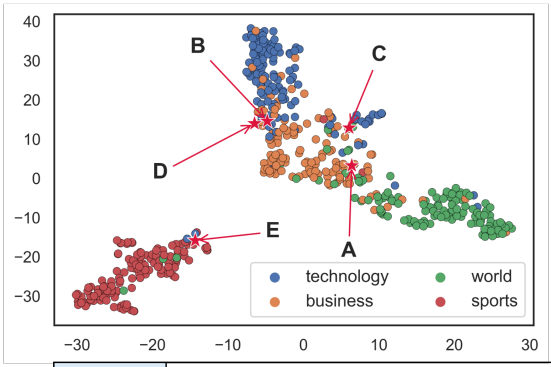
Table 5: Full results for data scaling, corresponds to Figure 5.



	ICL	kNN Prompting
A	positive (✗)	negative
B	positive (✗)	negative
C	positive (✗)	negative
D	positive (✗)	negative
E	positive (✗)	negative

A	looks and feels like a project better suited for the small screen .	negative	distance ↓
kNN	see `` simone , " and consider a dvd rental instead	negative	0.000348
	it 's a thin notion	negative	0.000395
	last-minute	negative	0.000430

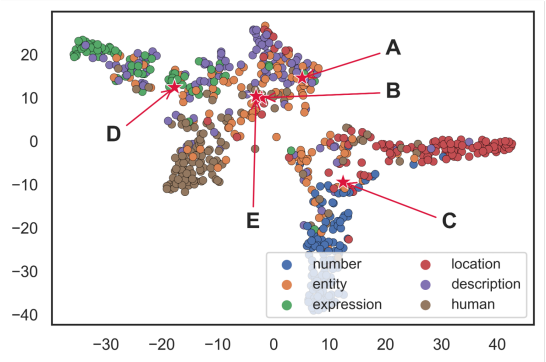
Figure 11: Case study on SST2, all 5 test instances are randomly selected given that kNN Prompting outperforms ICL.



	ICL	kNN Prompting
A	technology (✗)	business
B	technology (✗)	business
C	technology (✗)	business
D	technology (✗)	business
E	world (✗)	sports

A	Oil Falls From Record on Concern High Prices May Slow Growth. Crude oil fell from yesterday #39;s record of \\$54.88 a barrel in New York amid concern that sustained high prices may slow economies and reduce demand for energy.	business	distance ↓
kNN	Oil futures rise toward \\$US54. "Crude oil futures in New York rose to a record-high settlement overnight for the fifth straight session, closing just below \\$US54 a barrel, reflecting real and perceived threats to global oil supply."	business	0.0146
	Update 13: Oil Prices Little Changed on Norway News. Crude oil futures were little changed Tuesday as oil workers in Norway said they will end their four-month strike after the government ordered them back on the job.	business	0.0186
	Oil Rebounds After Iraq Pipeline Attack. Oil prices rose on Friday as a sabotage attack on Iraqi oil infrastructure reignited concern about the reliability of crude exports from the country despite a peace deal to end an uprising in Najaf.	business	0.0432

Figure 12: Case study on AGNews, all 5 test instances are randomly selected given that kNN Prompting outperforms ICL.



	ICL	kNN Prompting
A	entity (✗)	description
B	entity (✗)	human
C	number (✗)	entity
D	number (✗)	expression
E	entity (✗)	human

A	What is amitriptyline ?	description	distance ↓
kNN	What is digitalis ?	description	0.0687
	What is BPH ?	expression	0.0748
	What is agent orange ?	description	0.0938

Figure 13: Case study on TREC, all 5 test instances are randomly selected given that kNN Prompting outperforms ICL.

Task	Template	Label Space
SST2	Review: contains no wit , only labored gags Sentiment: negative Review: the film is powerful , accessible and funny . Sentiment:	negative, positive
SUBJ	Input: the script isn't very good ; not even someone as gifted as hoffman (the actor) can make it work . Type: subjective Input: he must do this in secret so that the parents and school personnel know nothing of his plan . Type:	subjective, objective
MPQA	Review: would not find it at all strange Sentiment: negative Review: as small (yet acceptable) as possible Sentiment:	negative, positive
AGNews	Input: Carlyle Looks Toward Commercial Aerospace (Reuters). "Reuters - Private investment firm Carlyle Group, which has a reputation for making well-timed and occasionally controversial plays in the defense industry, has quietly placed its bets on another part of the market. Type: technology Input: Superstar Kewell remains centre of attention. Socceroo forward Harry Kewell loosens up by tossing around a ball at Bondi beach yesterday. Photo: Craig Golding. There were half a dozen Socceroos standing on a raised platform in Sydney #39;s Type:	world, sports, business, technology
CB	Premise: It was a complex language. Not written down but handed down. One might say it was peeled down. Hypothesis: the language was peeled down Prediction: False Premise: A: so I don't know if I wasn't drug tested based on that or because the man who hired me didn't request the drug test, because I know that my company does drug testing on occasion. B: Right. Well, for instance, does the company you worked for before have the right or do they have the ability to say, hey, we've already drug tested her and she came up negative. A: Well, no, I don't think they can force another company to not drug test me just by saying that I didn't, I mean, Hypothesis: they can force another company to not drug test her Prediction:	False, True, Neither
CR	Review: it 's not as stylized as a sony or samsung . Sentiment: negative Review: i went out and got the canon today . Sentiment:	negative, positive
DBPedia	Input: Geoffrey D. Falksen (born July 31 1982) is an American steampunk writer. Type: artist Input: Monster Night is a 2006 film directed by Leslie Allen and Lorenzo Doumani. Type:	company, school, artist, athlete, politics, transportation, building, nature, village, animal, plant, album, film, book
MR	Review: "you might say tykwer has done all that heaven allows , if you wanted to make as anti-kieslowski a pun as possible . suffice to say its total promise is left slightly unfulfilled ." Sentiment: negative Review: an alternately raucous and sappy ethnic sitcom . . . you'd be wise to send your regrets . Sentiment:	negative, positive
RTE	Premise: A man is due in court later charged with the murder 26 years ago of a teenager whose case was the first to be featured on BBC One's Crimewatch. Colette Aram, 16, was walking to her boyfriend's house in Keyworth, Nottinghamshire, on 30 October 1983 when she disappeared. Her body was later found in a field close to her home. Paul Stewart Hutchinson, 50, has been charged with murder and is due before Nottingham magistrates later." Hypothesis: Paul Stewart Hutchinson is accused of having stabbed a girl. Prediction: false Premise: For women earning 22,000 a year, the total pay accumulated after six months maternity leave would be just 5,300 in the UK and 5,850 in Ireland. Entitlements in Germany would also be relatively low, at 5,900, along with those in France, Spain and the Netherlands, all at 6,750. At the other end of the scale, pay received after six months leave in Italy would be 9,150 while in Denmark and Norway it would be as much as 11,000. Hypothesis: Maternity leave varies in Europe. Prediction:	false, true
TREC	Question: How did serfdom develop in and then leave Russia ? Type: description Question: What is Shakespeare 's nickname ? Type:	description, entity, expression, human, location, number

Table 6: Templates for ICL. These are minimum cases with only one demonstration example for illustration.