A Proofs

A.1 Direct direction

Assumption A.1. (Assumption 2.1) Assume that $P_M(X) = P(X)$, and $P^d_M(Y|\boldsymbol{\theta}, X) \propto P(Y|\boldsymbol{\theta}, X)$ for $X \to Y \leftarrow \boldsymbol{\theta}$.

Proposition A.2. *(Proposition 2.2) If task d follows the* $X \to Y \leftarrow \boldsymbol{\theta}$ *direction,* $\arg\max_{y\in\mathcal{Y}} P^d_M(Y = y|\theta^d, X)$ *is the Bayes optimal classifier.*

Proof. Since the data generation of the task d can be written as $Y = f(X, \theta^d, \boldsymbol{\epsilon})$, we have

$$P^d(Y|X) = P(Y|\theta^d, X).$$

And by Assumption A.1, we have

$$\arg\max_{y\in\mathcal{Y}} P^d_M(Y = y|\theta^d, X) = \arg\max_{y\in\mathcal{Y}} P(Y = y|\theta^d, X).$$

Thus $\arg\max_{y\in\mathcal{Y}} P^d_M(Y = y|\theta^d, X)$ is the Bayes optimal classifier. □

Theorem A.3. *(Theorem 2.3) If task d follows the* $X \to Y \leftarrow \boldsymbol{\theta}$ *direction, then the in-context learning classifier*

$$\arg\max_{y\in\mathcal{Y}} P^d_M(Y = y|X^d_1, Y^d_1, ..., X^d_k, Y^d_k, X)$$

always has a higher or equal probability of misclassification to the Bayes optimal classifier $\arg\max_{y\in\mathcal{Y}} P^d_M(Y = y|\theta^d, X)$*. Equality only takes when*

$$\forall x \in \mathcal{X},\ P^d_M(\theta^d|X^d_1, Y^d_1, ..., X^d_k, Y^d_k, X = x) = 1.$$

Proof. Recall that in Equation (1), we have

$$P^d_M(Y|X^d_1, Y^d_1, ..., X^d_k, Y^d_k, X) = \int_\Theta P^d_M(Y|\boldsymbol{\theta}, X) P^d_M(\boldsymbol{\theta}|X^d_1, Y^d_1, ..., X^d_k, Y^d_k, X) d\boldsymbol{\theta}.$$

By Proposition A.2, $\arg\max_{y\in\mathcal{Y}} P^d_M(Y = y|\theta^d, X)$ is the Bayes optimal classifier. Let $C_{\boldsymbol{\theta}}(X) = \arg\max_{y\in\mathcal{Y}} P^d_M(Y = y|\boldsymbol{\theta}, X)$, then the risk is defined as the probability of misclassification

$$R(C_{\boldsymbol{\theta}}) = P(C_{\boldsymbol{\theta}}(X) \neq Y) = \mathbb{E}_{XY}[\mathbb{1}_{C_{\boldsymbol{\theta}}(X)\neq Y}].$$

Denote the in-context learning classifier $\arg\max_{y\in\mathcal{Y}} P^d_M(Y = y|X^d_1, Y^d_1, ..., X^d_k, Y^d_k, X)$ by $C_k(X)$. We then have

$$R(C_k) = \mathbb{E}_{XY}[\mathbb{1}_{C_k(X)\neq Y}] = \mathbb{E}_X[\sum_{y\in\mathcal{Y}} (1 - P^d_M(Y = y|\theta^d, X))\mathbb{1}_{C_k(X)=y}].$$

Such risk is minimized if and only if $C_k(X) = C_{\theta^d}(X)$, which only holds when $P^d_M(\theta^d|X^d_1, Y^d_1, ..., X^d_k, Y^d_k, X = x) = 1$ for all $x \in \mathcal{X}$. □

A.2 Channel direction

Assumption A.4. Assume that $P_M(X) = P(X)$, and $P^d_M(X|\boldsymbol{\theta}, Y) \propto P(X|\boldsymbol{\theta}, Y)$ for the $Y \to X \leftarrow \boldsymbol{\theta}$ direction.

Proposition A.5. *If task d follows the* $Y \to X \leftarrow \boldsymbol{\theta}$ *causal direction,* $\arg\max_{y\in\mathcal{Y}} P^d_M(X|\theta^d, Y = y)$ *is the Bayes optimal classifier when the label assignment is balanced.*

Proof. Since the data generation of the task d can be written as $X = g(Y, \theta^d, \boldsymbol{\epsilon})$, we have

$$P^d(X|Y) = P(X|\theta^d, Y)$$

When the label is balanced, i.e. $P^d(Y) = \frac{1}{|\mathcal{Y}|}$, we have

$$P^d(Y|X) = \frac{P^d(X|Y)P^d(Y)}{P(X)} \propto P^d(X|Y)$$

And by Assumption A.4, we have

$$\arg\max_{y\in\mathcal{Y}} P_M^d(X|\theta^d, Y=y) = \arg\max_{y\in\mathcal{Y}} P(X|\theta^d, Y=y).$$

Thus $\arg\max_{y\in\mathcal{Y}} P_M^d(X|\theta^d, Y=y) = \arg\max_{y\in\mathcal{Y}} P^d(Y=y|X)$ is the Bayes optimal classifier. $\square$

Theorem A.6. *If task d follows the $Y \to X \leftarrow \boldsymbol{\theta}$ direction, then the in-context learning classifier*

$$\arg\max_{y\in\mathcal{Y}} P_M^d(X|Y_1^d, X_1^d, ..., Y_k^d, X_k^d, Y=y)$$

always has a higher or equal probability of misclassification to the Bayes optimal classifier $\arg\max_{y\in\mathcal{Y}} P_M^d(X|\theta^d, Y=y)$. Equality only takes when

$$\forall y \in \mathcal{Y},\ P_M^d(\theta^d|Y_1^d, X_1^d, ..., Y_k^d, X_k^d, Y=y) = 1.$$

Proof. This theorem can be proved similarly as Theorem A.3. Recall that in Equation (2), we have

$$P_M^d(X|Y_1^d, X_1^d, ..., Y_k^d, X_k^d, Y) = \int_\Theta P_M^d(X|\boldsymbol{\theta}, Y)P_M^d(\boldsymbol{\theta}|Y_1^d, X_1^d, ..., Y_k^d, X_k^d, Y)d\boldsymbol{\theta}.$$

By Proposition A.5, $\arg\max_{y\in\mathcal{Y}} P_M^d(X|\theta^d, Y=y)$ is the Bayes optimal classifier. Let $C_{\boldsymbol{\theta}}(X) = \arg\max_{y\in\mathcal{Y}} P_M^d(X|\boldsymbol{\theta}, Y=y)$, then the risk is defined as the probability of misclassification

$$R(C_{\boldsymbol{\theta}}) = P(C_{\boldsymbol{\theta}}(X) \neq Y) = \mathbb{E}_{XY}[\mathbb{1}_{C_{\boldsymbol{\theta}}(X)\neq Y}].$$

Denote the in-context learning classifier $\arg\max_{y\in\mathcal{Y}} P_M^d(X|Y_1^d, X_1^d, ..., Y_k^d, X_k^d, Y=y)$ by $C_k(X)$. We then have

$$R(C_k) = \mathbb{E}_{XY}[\mathbb{1}_{C_k(X)\neq Y}] = \mathbb{E}_X[\sum_{y\in\mathcal{Y}}(1 - P_M^d(X|\theta^d, Y=y))\mathbb{1}_{C_k(X)=y}].$$

Such risk is minimized if and only if $C_k(X) = C_{\theta^d}(X)$, which only holds when $P_M^d(\theta^d|Y_1^d, X_1^d, ..., Y_k^d, X_k^d, Y=y) = 1$ for all $y \in \mathcal{Y}$. $\square$

A.3 Method

Proposition A.7. *(Proposition 3.1) When $\mathcal{L}(\hat{\theta}^d)$ is minimized, $P_M^d(Y|\hat{\theta}^d, X) = P(Y|\theta^d, X)$ for $X \to Y \leftarrow \boldsymbol{\theta}$, and $P_M^d(X|\hat{\theta}^d, Y) = P(X|\theta^d, Y)$ for $Y \to X \leftarrow \boldsymbol{\theta}$. If the LLM M is invertible, then $\hat{\theta}^d = \theta^d$.*

Proof. The proof of this proposition is straightforward.

Since

$$\mathcal{L}(\hat{\theta}^d) = H(P(Y|\theta^d, X)) + KL(P(Y|\theta^d, X)||P_M^d(Y|\hat{\theta}^d, X))$$

when $\mathcal{L}(\hat{\theta}^d)$ is minimized, we have $P_M^d(Y|\hat{\theta}^d, X) = P(Y|\theta^d, X)$ for $X \to Y \leftarrow \boldsymbol{\theta}$, and $P_M^d(X|\hat{\theta}^d, Y) = P(X|\theta^d, Y)$ for $Y \to X \leftarrow \boldsymbol{\theta}$.

If M is invertible, since the embedding matrix is invertible with or without new concept tokens, $P_M^d(Y|\hat{\theta}, X) = P_M^d(Y|\hat{\theta}', X)$ implies that $\hat{\theta} = \hat{\theta}'$. Thus $\boldsymbol{\theta}$ is identifiable, which means $\hat{\theta}^d = \theta^d$. $\square$

Table 1: Prompt template and label mapping for the datasets we use. Since almost all sentences from ETHOS contain offensive content, we mask out the key offensive words in the examples below.

Dataset	Prompt	Label Mapping
SST-2	sentence: well worth revisiting as many times positive	negative/positive
FPB	The company anticipates its turnover for the whole 2010 to surpass that of the previous year when it was EUR 67.1 million . positive	negative/neutral/positive
COLA	It is this hat that I know the boy who is wearing. unacceptable	acceptable/unacceptable
DBPedia	The Nucet River is a tributary of the Chiojdeanca River in Romania. NaturalPlace	Album/Animal/Artist/ Athlete/Building/Company/ EducationalInstitution/Film/ MeanOfTransportation/ NaturalPlace/OfficeHolder/ Plant/Village/WrittenWork
EmoC	fast i mean fastingis a way of skipping meals i mena you move on too fast others	angry/happy/others/sad
EmoS	i feel this place was tragic sadness	anger/fear/joy/love/ sadness/surprise
ETHOS-SO	[Masked] should be removed from the face of the earth true	false/true
ETHOS-R	I hate being a [Masked], wish I was a [Masked] and no [Masked] on earth existed false	false/true

B Experiments

Dateset. In Table 1, we show how we process the text classification datasets into prompts. For each dataset, we take at most 16384 examples from the training set for training, and uniformly sample at most 1000 examples from the test set to test the in-context learning performance. In Table 2, we show the train size and test size we used for each dataset. We also list the set of diverse tasks trained with each dataset, which are denoted by their name in Huggingface datasets.[7] The license for SST2, ETHOS-SO and ETHOS-R is GNU General Public License v3. FPB is under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License. Note that these two datasets are hate speech detection datasets for different kinds of hate speech and contain many offensive texts. COLA is excerpted from the published works available on the website, and the copyright (where applicable) remains with the original authors or publishers. DBpedia is under a Creative Commons Attribution-ShareAlike License and the GNU Free Documentation License. EmoC and EmoS should be used for educational and research purposes only.

Experiment details. We run our experiments on A100, V100, and A6000 GPUs. We adopt a large portion of the code from the MetaICL repository [20][8]. The training takes around 20 to 40 hours on a single GPU. We use a learning rate of 1e-4 and a batch size of 16, and train for 10k steps in total.

Main results. In Table 3, we list the detailed results of our method and baselines with different LLMs on different datasets in Figure 2.

Causal direction results. The detailed results with anti-causal direction (the opposite direction to what we described in Section 4 are in Table 6) are shown in Table 6, corresponding to Figure 6 in the main text.

Other LLMs results. The detailed results with other LLMs are shown in Table 5, corresponding to Figure 3a in the main text.

Random token results. The detailed results with random tokens are shown in Table 4, corresponding to Figure 3b in the main text.

[7] `https://huggingface.co/docs/datasets/index`
[8] `https://github.com/facebookresearch/MetaICL`

datset d	train size	test size	task set $\mathcal{S}$
SST2 (glue-sst2)	16384	1000	glue-cola/glue-mnli/glue-qqp/ glue-mrpc/glue-qnli/glue-rte/glue-sst2/glue-wnli
FPB (financial_phrasebank)	1811	453	glue-sst2/glue-mnli/math_qa/sciq/ social_i_qa/wino_grande/glue-qqp/ ag_news/financial_phrasebank/ poem_sentiment/anli/quarel/quartz/ medical_questions_pairs/paws/dbpedia_14
COLA (cola-sst2)	8551	1000	glue-cola/glue-mnli/glue-qqp/glue-mrpc/ glue-qnli/glue-rte/glue-sst2/glue-wnli
DBpedia (dbpedia_14)	16384	1000	glue-sst2/glue-mnli/math_qa/sciq/ social_i_qa/wino_grande/glue-qqp/ ag_news/financial_phrasebank/ poem_sentiment/anli/quarel/quartz/ medical_questions_pairs/paws/dbpedia_14
EmoC (emo)	16384	1000	glue-sst2/amazon_polarity/ financial_phrasebank/poem_sentiment/ yelp_polarity/glue-cola/blimp/ag_news/ dbpedia_14/ethos/emo/emotion
EmoS (emotion)	16000	1000	glue-sst2/amazon_polarity/ financial_phrasebank/poem_sentiment/ yelp_polarity/glue-cola/blimp/ag_news/ dbpedia_14/ethos/emo/emotion
ETHOS-SO (ethos-sexual_orientation)	346	87	glue-sst2/amazon_polarity/ financial_phrasebank/poem_sentiment/ yelp_polarity/glue-cola/blimp/ag_news/ dbpedia_14/ethos/emo/emotion
ETHOS-R (ethos-religion)	346	87	glue-sst2/amazon_polarity/ financial_phrasebank/poem_sentiment/ yelp_polarity/glue-cola/blimp/ag_news/ dbpedia_14/ethos/emo/emotion

Table 2: Dataset details

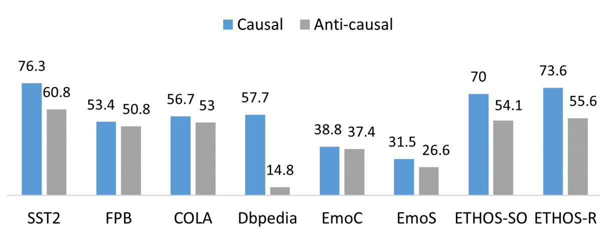


Figure 6: Accuracy of randomly selected demonstrations averaged over seven different LLMs except for GPT3-davinci, using the adopted *causal* direction and the *anti-causal* direction.

k-ablation study results. The detailed results of k ablation study are shown in Table 9, corresponding to Figure 4a in the main text. In this experiment, we do not reorder the selected demonstrations according to Equation (3), as we need to use GPT2-large for the reordering, and it cannot fit in all the demonstrations. Instead, we order the selected demonstrations from the largest $\hat{P}^d_M(\theta^d|X^d, Y^d)$ to the smallest.

c-ablation study results. The detailed results of c ablation study are shown in Table 10, corresponding to Figure 4b in the main text.

Effect of using ground truth labels. According to [21], the ground truth label is not necessary for demonstrations to have a good in-context learning performance, which we found is not entirely true for all the tasks. We compare our method with the randomly selected demonstration baseline under three scenarios: (a) **Original**: demonstrations with the correct labels; (b) **Random words**: using a random label projection map τ^d instead of a meaningful one. i.e., map each label to a fixed random word. In this case, the mapping from the input tokens X to the labels Y is still preserved; (c) **Random labels**: assign a random label to each demonstration, with the original label projection map τ^d. As shown in Figure 7, by using a random label projection map or randomly assigning the labels, the performance of the randomly selected demonstration baseline drops considerably. And randomize the label assignment gives a larger performance drop than only using a random label projection map,

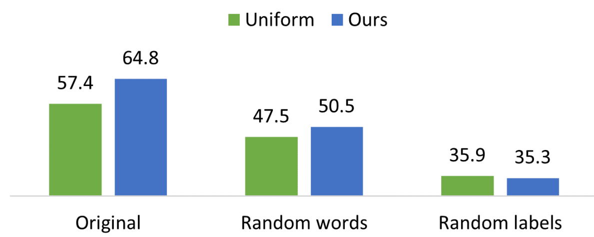


Figure 7: In-context learning accuracy of our method versus random selection baseline, with (a) ground truth labels (*original*), (b) random label mapping (*random words*), or random label assignments (*random label*), averaged over all eight datasets. Numbers are obtained with GPT2-large.

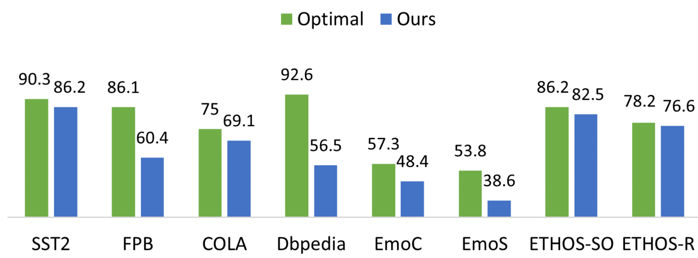


Figure 8: Accuracy of in-context learning using our method versus the theoretical maximum accuracy obtained using the learned concept tokens as prefixes. Numbers are obtained with GPT2-large.

which shows that the mapping between X and Y in the demonstrations matters. This indicates that in-context learning infers the mapping between X and Y from the demonstrations instead of merely invoking some learned function stored in the LLM parameters based on the appearance of X and Y. We also show that the demonstrations selected by our method represent the $X - Y$ mapping better, as under the **Random words** condition, our method performs better than the random selection baseline, while our method does not improve the random selection baseline under the **Random labels** condition. The detailed results with random words and random labels are shown in Table 7

Optimal performance As stated in Theorem 2.3, the optimal performance of an in-context learning classifier is the Bayes optimal classifier $\arg\max_{y \in \mathcal{Y}} P_M^d(Y = y|\theta^d, X)$, which is approximated by using the learned concept tokens as prefixes. Note that this approximated Bayes optimal classifier cannot be transferred across different LLMs, as the learned concept tokens embeddings are aligned with a specific LLM. The advantage of in-context learning with our method is that the demonstrations can be transferred to any LLMs without training. Here we only compare the accuracy of in-context learning with our method and the approximated Bayes optimal classifier using GPT2-large, as it is the LLM that concept tokens are fine-tuned with. As shown in Figure 8, our method comes close to the optimal accuracy on many datasets, while there are some datasets that our method is lagging. This indicates that there are two ways to improve our method: the first is to improve the performance of the optimal classifier, by introducing a better latent concept learning algorithm. The other way is to reduce the performance gap between our method and the optimal classifier, by improving the demonstration selection algorithm. The detailed results using the learned concept tokens as prefixes are shown in Table 8.

Reordering results. The detailed results with and without reordering are shown in Table 11, corresponding to Figure 9.

Similar tokens. We show the top ten similar tokens to some learned concept tokens in Table 12, as summarized in Figure 5 in the main text.

Likelihood histogram. We also show histograms of the probability of each example predicting corresponding concept tokens in different datasets. We can see that the probability of prediction concept tokens can well differentiate examples in a dataset.

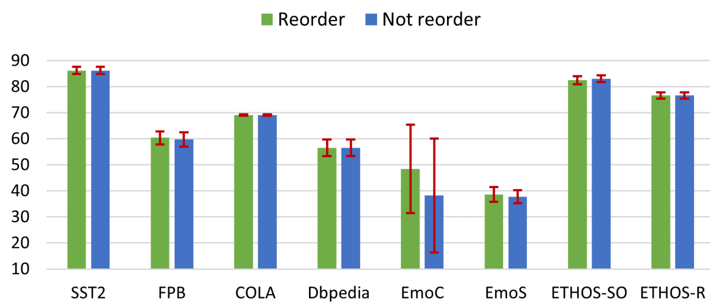


Figure 9: In-context learning accuracy of our method versus random selection baseline, with and without reordering. The red error bars represent the standard deviation across five runs. Numbers are obtained with GPT2-large.

Table 3: Accuracy of selected demonstration. Our demonstrations are selected using GPT2-large, and the same set of demonstrations is applied to all different LLMs. All LLMs are pre-trained only with the language modeling objective, while the pre-training data size of GPT2s is much smaller than GPT3s.

LLM	Method	SST2	FPB	COLA	DBpedia	EmoC	EmoS	ETHOS-SO	ETHOS-R	Avg
GPT2	Uniform	69.7 $\pm$ 1.8	52.9 $\pm$ 2.3	61.9 $\pm$ 1.4	48.0 $\pm$ 0.7	35.3 $\pm$ 1.7	26.4 $\pm$ 1.0	64.1 $\pm$ 4.8	71.0 $\pm$ 1.8	53.7
(124M)	Similar	69.5 $\pm$ 0.6	55.9 $\pm$ 1.7	63.2 $\pm$ 1.2	44.7 $\pm$ 3.1	36.4 $\pm$ 2.0	26.6 $\pm$ 1.3	77.7 $\pm$ 2.7	80.0 $\pm$ 3.7	56.8
	Ours	76.8 $\pm$ 2.9	64.5 $\pm$ 3.2	69.1 $\pm$ 0.2	53.5 $\pm$ 2.95	37.2 $\pm$ 11.1	30.6 $\pm$ 4.8	80.9 $\pm$ 1.9	76.8 $\pm$ 2.6	61.2
GPT2-m	Uniform	70.8 $\pm$ 1.3	52.0 $\pm$ 1.7	57.8 $\pm$ 1.3	49.3 $\pm$ 2.0	34.2 $\pm$ 1.8	34.2 $\pm$ 1.8	76.3 $\pm$ 4.9	74.7 $\pm$ 2.2	56.2
(355M)	Similar	75.0 $\pm$ 1.9	57.7 $\pm$ 2.0	57.5 $\pm$ 2.2	47.9 $\pm$ 6.0	37.2 $\pm$ 3.6	35.2 $\pm$ 1.8	86.9 $\pm$ 2.9	84.6 $\pm$ 4.3	60.3
	Ours	81.2 $\pm$ 1.3	59.3 $\pm$ 4.3	69.0 $\pm$ 0.2	52.9 $\pm$ 2.3	40.4 $\pm$ 21.5	37.2 $\pm$ 2.4	83.7 $\pm$ 1.1	76.8 $\pm$ 1.1	62.6
GPT2-l	Uniform	77.1 $\pm$ 1.2	51.3 $\pm$ 2.4	62.7 $\pm$ 0.8	54.4 $\pm$ 0.9	38.7 $\pm$ 2.1	34.5 $\pm$ 1.2	67.6 $\pm$ 4.3	72.9 $\pm$ 2.8	57.4
(774M)	Similar	80.7 $\pm$ 1.6	54.8 $\pm$ 3.8	50.9 $\pm$ 1.4	51.1 $\pm$ 5.2	39.9 $\pm$ 2.6	35.1 $\pm$ 2.1	80.9 $\pm$ 2.8	84.4 $\pm$ 2.6	59.7
	Ours	86.2 $\pm$ 1.4	60.4 $\pm$ 2.5	69.1 $\pm$ 0.2	56.5 $\pm$ 3.2	48.4 $\pm$ 17.0	38.6 $\pm$ 2.8	82.5 $\pm$ 1.5	76.6 $\pm$ 1.2	64.8
GPT2-xl	Uniform	74.7 $\pm$ 0.9	53.2 $\pm$ 1.9	55.8 $\pm$ 1.6	53.0 $\pm$ 1.9	38.2 $\pm$ 1.5	38.2 $\pm$ 1.5	67.8 $\pm$ 6.4	72.6 $\pm$ 4.1	56.7
(1.5B)	Similar	80.6 $\pm$ 1.3	53.0 $\pm$ 2.5	55.0 $\pm$ 2.5	51.6 $\pm$ 5.9	39.9 $\pm$ 2.0	32.9 $\pm$ 2.1	82.8 $\pm$ 2.2	83.9 $\pm$ 4.5	60
	Ours	83.1 $\pm$ 3.6	62.0 $\pm$ 2.5	68.9 $\pm$ 0.2	58.6 $\pm$ 3.3	43.6 $\pm$ 16.4	43.6 $\pm$ 16.4	83.0 $\pm$ 1.3	77.9 $\pm$ 1.3	65.1
GPT3-a	Uniform	76.9 $\pm$ 0.7	56.6 $\pm$ 1.1	53.1 $\pm$ 1.8	62.1 $\pm$ 1.4	38.6 $\pm$ 1.4	27.7 $\pm$ 1.3	65.5 $\pm$ 5.7	74.0 $\pm$ 3.0	56.8
(350M)	Similar	78.7 $\pm$ 1.0	52.2 $\pm$ 2.7	53.1 $\pm$ 1.8	54.6 $\pm$ 1.7	42.4 $\pm$ 3.5	37.2 $\pm$ 1.1	84.1 $\pm$ 2.2	87.8 $\pm$ 3.5	61.3
	Ours	85.4 $\pm$ 1.7	61.9 $\pm$ 10.5	58.2 $\pm$ 7.0	64.0 $\pm$ 4.4	43.0 $\pm$ 7.2	37.9 $\pm$ 2.3	84.4 $\pm$ 1.4	78.9 $\pm$ 0.9	64.2
GPT3-b	Uniform	80.8 $\pm$ 0.6	55.2 $\pm$ 3.3	46.8 $\pm$ 2.0	66.5 $\pm$ 1.4	42.0 $\pm$ 0.7	27.0 $\pm$ 1.2	71.0 $\pm$ 4.6	72.6 $\pm$ 3.1	57.7
(1.3B)	Similar	83.9 $\pm$ 1.3	56.2 $\pm$ 2.3	45.1 $\pm$ 1.8	59.8 $\pm$ 1.8	42.9 $\pm$ 3.5	38.1 $\pm$ 1.7	86.7 $\pm$ 3.0	86.4 $\pm$ 3.0	62.4
	Ours	87.3 $\pm$ 2.0	64.3 $\pm$ 5.9	67.2 $\pm$ 0.9	70.2 $\pm$ 3.2	43.6 $\pm$ 13.0	38.9 $\pm$ 5.0	84.6 $\pm$ 0.9	78.9 $\pm$ 1.2	66.9
GPT3-c	Uniform	84.2 $\pm$ 1.4	52.6 $\pm$ 1.8	59.1 $\pm$ 1.5	70.6 $\pm$ 0.8	44.3 $\pm$ 2.5	32.3 $\pm$ 1.9	77.5 $\pm$ 4.7	77.5 $\pm$ 0.6	62.3
(6.7B)	Similar	85.7 $\pm$ 1.4	62.2 $\pm$ 0.9	58.0 $\pm$ 1.7	62.2 $\pm$ 2.0	47.4 $\pm$ 4.3	39.8 $\pm$ 1.7	89.2 $\pm$ 1.4	89.7 $\pm$ 1.9	66.8
	Ours	88.8 $\pm$ 0.7	64.1 $\pm$ 5.7	69.0 $\pm$ 0.3	73.6 $\pm$ 2.9	50.3 $\pm$ 11.9	43.1 $\pm$ 4.6	86.2 $\pm$ 0.0	78.2 $\pm$ 0.0	69.2
GPT3-d	Uniform	86.5 $\pm$ 0.9	59.2 $\pm$ 2.4	45.5 $\pm$ 2.8	73.6 $\pm$ 1.9	39.4 $\pm$ 0.7	40.6 $\pm$ 1.7	77.2 $\pm$ 2.6	76.8 $\pm$ 3.5	62.4
(175B)	Similar	88.5 $\pm$ 0.8	55.4 $\pm$ 3.3	45.4 $\pm$ 1.5	67.2 $\pm$ 1.8	37.6 $\pm$ 1.6	39.8 $\pm$ 1.4	86.9 $\pm$ 2.4	89.0 $\pm$ 3.8	63.7
	Ours	87.8 $\pm$ 3.4	62.7 $\pm$ 3.3	58.5 $\pm$ 8.2	75.5 $\pm$ 2.4	41.3 $\pm$ 3.6	42.7 $\pm$ 3.9	85.1 $\pm$ 0.0	79.3 $\pm$ 0.0	66.6
Avg	Uniform	77.6	54.1	55.3	59.7	38.8	32.6	70.9	74.0	57.9
	Similar	80.3	55.9	53.5	54.9	40.5	35.6	84.4	85.7	61.4
	Ours	84.6	62.4	66.1	63.1	43.5	39.1	83.8	77.9	65.0

C Limitations and Future Work

While the assumption that a large language model captures the true distribution of language is fairly common in the literature studying LLMs [44, 29], this assumption is not entirely accurate in practice. According to [12], LLMs systematically underestimate rare text sequences, which constitute a significant portion of the long-tail distribution of language. Although this assumption is adequate to achieve favorable empirical results, it is expected that more accurate language models will, in theory, lead to improved outcomes.

The selection of the accompanying diverse tasks $\mathcal{S}$ is currently left to the user's discretion. A better approach to constructing such a task set is needed to gain a deeper understanding of latent concept variables and to improve the latent concept learning algorithm.

Our algorithm currently only applies to classification tasks. More complex latent variables could be designed to improve the in-context learning performance of more complex tasks like math word questions and logical reasoning problems.

Table 4: Accuracy of selected demonstration. Our demonstrations are selected using GPT2-large, and the same set of demonstrations is applied to all different LLMs. All LLMs are pre-trained only with the language modeling objective, while the pre-training data size of GPT2s is much smaller than GPT3s.

LLM	Method	SST2	FPB	COLA	DBpedia	EmoC	EmoS	ETHOS-SO	ETHOS-R	Avg
GPT2	Uniform	69.7 $\pm$ 1.8	52.9 $\pm$ 2.3	61.9 $\pm$ 1.4	48.0 $\pm$ 0.7	35.3 $\pm$ 1.7	26.4 $\pm$ 1.0	64.1 $\pm$ 4.8	71.0 $\pm$ 1.8	53.7
(124M)	Random	69.8 $\pm$ 3.3	51.1 $\pm$ 1.7	69.0 $\pm$ 0.1	49.0 $\pm$ 4.5	33.7 $\pm$ 15.5	24.2 $\pm$ 7.6	66.4 $\pm$ 17.5	66.2 $\pm$ 16.2	53.7
	Ours	76.8 $\pm$ 2.9	64.5 $\pm$ 3.2	69.1 $\pm$ 0.2	53.5 $\pm$ 2.95	37.2 $\pm$ 11.1	30.6 $\pm$ 4.8	80.9 $\pm$ 1.9	76.8 $\pm$ 2.6	61.2
GPT2-l	Uniform	77.1 $\pm$ 1.2	51.3 $\pm$ 2.4	62.7 $\pm$ 0.8	54.4 $\pm$ 0.9	38.7 $\pm$ 2.1	34.5 $\pm$ 1.2	67.6 $\pm$ 4.3	72.9 $\pm$ 2.8	57.4
(774M)	Random	81.9 $\pm$ 4.5	46.5 $\pm$ 4.7	64.9 $\pm$ 7.8	50.3 $\pm$ 4.3	42.5 $\pm$ 16.7	36.1 $\pm$ 6.5	67.6 $\pm$ 20.4	67.8 $\pm$ 15.0	57.2
	Ours	86.2 $\pm$ 1.4	60.4 $\pm$ 2.5	69.1 $\pm$ 0.2	56.5 $\pm$ 3.2	48.4 $\pm$ 17.0	38.6 $\pm$ 2.8	82.5 $\pm$ 1.5	76.6 $\pm$ 1.2	64.8

Table 5: We test our method on other similar sizes (6-7B) LLMs.

LLM	Method	SST2	FPB	COLA	DBpedia	EmoC	EmoS	ETHOS-SO	ETHOS-R	Avg
GPT2-l	Random	77.1 $\pm$ 1.2	51.3 $\pm$ 2.4	62.7 $\pm$ 0.8	54.4 $\pm$ 0.9	38.7 $\pm$ 2.1	34.5 $\pm$ 1.2	67.6 $\pm$ 4.3	72.9 $\pm$ 2.8	57.4
	Ours	86.2 $\pm$ 1.4	60.4 $\pm$ 2.5	69.1 $\pm$ 0.2	56.5 $\pm$ 3.2	48.4 $\pm$ 17.0	38.6 $\pm$ 2.8	82.5 $\pm$ 1.5	76.6 $\pm$ 1.2	64.8
GPT3-c	Random	84.2 $\pm$ 1.4	52.6 $\pm$ 1.8	59.1 $\pm$ 1.5	70.6 $\pm$ 0.8	44.3 $\pm$ 2.5	32.3 $\pm$ 1.9	77.5 $\pm$ 4.7	77.5 $\pm$ 0.6	62.3
	Ours	88.8 $\pm$ 0.7	64.1 $\pm$ 5.7	69.0 $\pm$ 0.3	73.6 $\pm$ 2.9	50.3 $\pm$ 11.9	43.1 $\pm$ 4.6	86.2 $\pm$ 0.0	78.2 $\pm$ 0.0	69.2
GPT-J	Random	78.5 $\pm$ 1.0	53.1 $\pm$ 1.7	58.3 $\pm$ 2.2	55.6 $\pm$ 1.2	38.5 $\pm$ 2.0	33.3 $\pm$ 1.5	76.6 $\pm$ 3.7	76.6 $\pm$ 1.4	58.8
	Ours	87.8 $\pm$ 1.9	56.7 $\pm$ 4.3	69.1 $\pm$ 0.2	60.0 $\pm$ 3.6	32.5 $\pm$ 16.1	33.2 $\pm$ 2.8	85.3 $\pm$ 0.5	77.0 $\pm$ 0.0	62.7
OPT	Random	72.4 $\pm$ 0.8	32.8 $\pm$ 0.3	34.8 $\pm$ 0.6	29.4 $\pm$ 1.4	67.1 $\pm$ 1.8	36.9 $\pm$ 0.6	86.2 $\pm$ 0.0	78.2 $\pm$ 0.0	54.7
	Ours	74.2 $\pm$ 3.0	34.1 $\pm$ 6.1	35.7 $\pm$ 3.1	28.8 $\pm$ 2.1	76.7 $\pm$ 4.1	39.0 $\pm$ 3.4	86.2 $\pm$ 0.0	78.2 $\pm$ 0.0	56.6
LLaMA	Random	57.7 $\pm$ 1.5	23.7 $\pm$ 1.3	30.8 $\pm$ 0.2	15.8 $\pm$ 0.8	4.4 $\pm$ 0.7	35.2 $\pm$ 0.7	66.2 $\pm$ 5.8	57.2 $\pm$ 5.1	36.4
	Ours	60.5 $\pm$ 4.7	19.1 $\pm$ 1.9	30.8 $\pm$ 0.2	16.9 $\pm$ 1.3	4.3 $\pm$ 0.7	35.3 $\pm$ 0.6	77.2 $\pm$ 13.6	56.3 $\pm$ 10.8	37.6

D Broader Impact

The utilization of language models (LLMs) for specific tasks is often hindered by the high cost associated with training or fine-tuning them. However, the in-context learning paradigm offers a cost-effective and convenient alternative for utilizing the power of pre-trained LLMs. Our work has demonstrated a significant improvement in the performance of in-context learning through a relatively low-cost and simple approach, thus making the use of LLMs more accessible for individuals with limited resources.

However, it is important to consider the broader implications of the increasing use of LLMs. As LLMs are not infallible and may make mistakes, it is crucial to explicitly warn users of the potential for misleading output and to regulate the distribution of LLMs in order to prevent any negative societal impact. Additionally, it is possible that LLMs could be intentionally misused, thus it is important to consider the ethical implications of their use and to take appropriate measures to mitigate any potential negative effects. We posit that these regulations and measures should be put in place at the time of distributing LLMs to ensure the safe and responsible use of these models. Furthermore, as we publicly release our code, we will also provide clear warnings and guidelines to users to ensure that the potential risks associated with the use of our method are fully understood and addressed.

Table 6: We test random selection baseline with anti-causal direction.

LLM	SST2	FPB	COLA	DBpedia	EmoC	EmoS	ETHOS-SO	ETHOS-R
GPT2	57.4 ± 1.9	56.6 ± 2.1	55.9 ± 1.7	11.3 ± 1.0	24.6 ± 2.4	22.1 ± 1.1	64.1 ± 4.8	58.6 ± 5.5
GPT2-m	56.7 ± 1.6	48.7 ± 2.1	55.3 ± 1.8	13.9 ± 1.2	22.4 ± 1.9	24.9 ± 2.3	44.8 ± 1.9	45.5 ± 3.5
GPT2-l	58.7 ± 0.7	33.7 ± 1.3	50.8 ± 1.6	13.6 ± 1.3	28.2 ± 3.6	26.2 ± 2.7	48.7 ± 3.7	53.6 ± 5.3
GPT2-xl	54.2 ± 0.5	46.8 ± 1.2	50.6 ± 1.1	12.6 ± 1.5	31.4 ± 2.8	25.9 ± 3.2	65.5 ± 4.9	61.8 ± 1.5
GPT3-a	55.8 ± 0.9	58.9 ± 2.1	51.6 ± 1.4	14.3 ± 0.8	54.2 ± 3.1	27.7 ± 1.3	49.2 ± 3.3	54.9 ± 6.4
GPT3-b	64.4 ± 1.6	58.9 ± 2.6	53.4 ± 1.1	14.6 ± 1.1	52.0 ± 2.5	27.0 ± 1.3	48.3 ± 2.7	51.0 ± 4.0
GPT3-c	78.2 ± 1.6	52.3 ± 2.3	53.7 ± 0.7	23.0 ± 2.5	49.1 ± 2.6	32.2 ± 1.9	57.9 ± 2.7	64.1 ± 5.0
Avg	60.8	50.8	53	14.8	37.4	26.6	54.1	55.6

Table 7: We test our method with random words and random labels using GPT2-large.

	Method	SST2	FPB	COLA	DBpedia	EmoC	EmoS	ETHOS-SO	ETHOS-R	Avg
R words	Random	54.1 ± 4.2	43.4 ± 1.9	62.2 ± 4.9	11.2 ± 0.9	32.4 ± 5.2	19.1 ± 1.8	80.7 ± 4.8	77.0 ± 3.6	47.5
	Ours	50.3 ± 1.3	44.9 ± 4.2	69.2 ± 0.2	13.9±1.2	37.8 ± 12.1	23.5 ± 7.4	86.0 ± 0.5	77.9 ± 0.5	50.5
R labels	Random	51.5 ± 0.9	32.5 ± 1.2	49.3 ± 3.0	6.7 ± 1.0	25.1 ± 0.6	17.2 ± 0.9	48.0 ± 2.5	56.8 ± 3.1	35.9
	Ours	49.6 ± 0.9	36.2 ± 2.5	49.3 ± 1.6	6.6± 0.2	24.7 ± 0.6	16.6 ± 1.0	51.0 ± 4.9	48.7 ± 3.5	35.3

Table 8: Accuracy using concept tokens as prefixes.

SST2	FPB	COLA	DBpedia	EmoC	EmoS	ETHOS-SO	ETHOS-R
90.3 ± 0.0	86.1 ± 0.0	75.0 ± 0.1	92.6 ± 0.6	57.3 ± 1.8	53.8 ± 0.7	86.2 ± 0.0	78.2 ± 0.0

Table 9: k ablation study using GPT2-large, without reordering.

	Method	SST2	FPB	COLA	DBpedia	EmoC	EmoS	ETHOS-SO	ETHOS-R	Avg
$k = 2$	Random	74.4 ± 1.0	48.5 ± 1.1	48.9 ± 1.6	52.9 ± 2.0	42.8 ± 0.6	37.1 ± 1.2	66.9 ± 4.7	66.4 ± 6.8	54.7
	Ours	78.1 ± 4.5	50.1 ± 2.9	54.3 ± 8.8	57.3 ± 5.1	41.1 ± 9.8	36.1 ± 2.6	84.6 ± 1.6	76.8 ± 4.5	59.8
$k = 4$	Random	76.9 ± 0.7	56.6 ± 1.1	53.1 ± 1.8	62.1 ± 1.4	38.6 ± 1.4	27.7 ± 1.3	65.5 ± 5.7	74.0 ± 3.0	56.8
	Ours	86.2 ± 1.4	59.7 ± 2.8	69.1 ± 0.2	56.5 ± 3.2	38.2 ± 21.8	37.7 ± 2.5	83.0 ± 1.3	76.6 ± 1.2	63.4
$k = 8$	Random	79.9 ± 0.2	57.1 ± 1.6	51.3 ± 1.0	66.5 ± 1.2	37.6 ± 1.5	36.2 ± 0.6	68.5 ± 3.5	72.9 ± 3.3	58.8
	Ours	87.0 ± 2.4	59.9 ± 3.3	55.3 ± 9.7	67.0 ± 0.9	39.9 ± 5.3	38.8 ± 2.6	77.0 ± 11.1	78.9 ± 0.9	63
$k = 16$	Random	79.9 ± 1.1	54.9 ± 2.7	54.5 ± 2.8	69.1 ± 1.1	33.7 ± 2.2	33.5 ± 1.4	64.8 ± 4.0	69.0 ± 3.2	57.4
	Ours	84.6 ± 1.9	60.4 ± 6.4	62.0 ± 7.0	71.0 ± 1.9	37.2 ± 6.1	37.1 ± 2.2	72.4 ± 7.6	74.7 ± 4.7	62.4

Table 10: c ablation study using GPT2-large

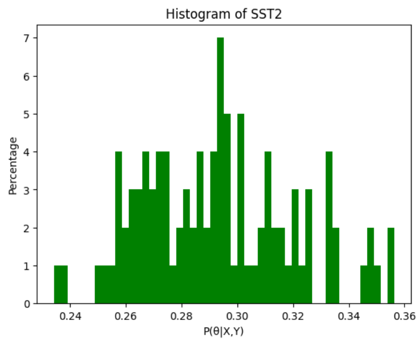
	SST2	FPB	COLA	DBpedia	EmoC	EmoS	ETHOS-SO	ETHOS-R	Avg
$c = 5$	78.9 ± 2.4	59.8 ± 10.8	34.3 ± 5.0	62.9 ± 2.4	44.9 ± 9.5	38.1 ± 2.4	71.7 ± 5.9	62.1 ± 19.7	56.6
$c = 10$	85.4 ± 1.7	61.9 ± 10.5	58.2 ± 7.0	64.0 ± 4.4	43.0 ± 7.2	37.9 ± 2.3	84.4 ± 1.4	78.9 ± 0.9	64.2
$c = 15$	80.1 ± 1.4	64.3 ± 7.7	63.1 ± 9.4	58.7 ± 3.2	36.4 ± 11.5	38.6 ± 1.9	80.9 ± 3.9	76.3 ± 5.9	62.3
$c = 20$	78.5 ± 4.1	51.8 ± 8.0	66.5 ± 2.3	58.0 ± 3.4	36.3 ± 4.3	41.8 ± 5.8	80.7 ± 4.5	73.8 ± 5.4	60.92

Table 11: Reorder versus not reorder using our method, with GPT2-large.

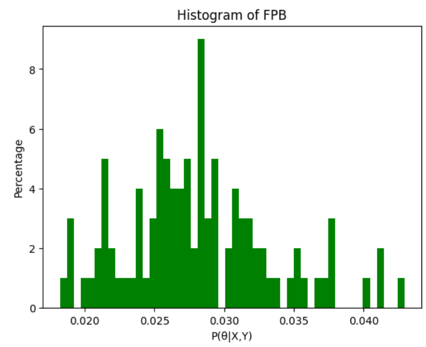
	SST2	FPB	COLA	DBpedia	EmoC	EmoS	ETHOS-SO	ETHOS-R	Avg
reorder	86.2 ± 1.4	60.4 ± 2.5	69.1 ± 0.2	56.5 ± 3.2	48.4 ± 17.0	38.6 ± 2.8	82.5 ± 1.5	76.6 ± 1.2	64.8
not reorder	86.2 ± 1.4	59.7 ± 2.8	69.1 ± 0.2	56.5 ± 3.2	38.2 ± 21.8	37.7 ± 2.5	83.0 ± 1.3	76.6 ± 1.2	63.4

Table 12: We list the top 10 similar words (tokens) to some of the learned concept tokens.

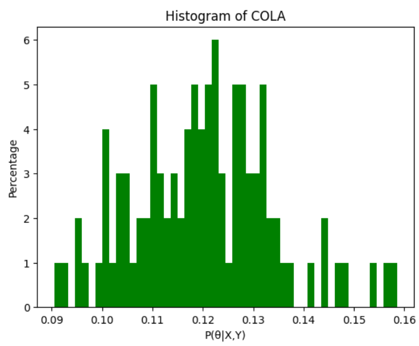
concept token	similar words
FPB-2	milo coordinate notify rendering benefiting routing EntityItem routed Messages Plot
FPB-3	unlocked updating deleting dropping damage updates drops Gained taken dropped
FPB-4	FX Safari Fixes advertisers Links Coins Operator marketers Guidelines
FPB-5	674 592 693 696 498 593 793 504 691 683
COLA-1	exha trunc curv fragmented elong iterator initialized bounds Iter filament
COLA-2	Sp spa contributed cerv borrower paper tiger Erica USH Schwartz
COLA-7	democr Barack WH ophobic neum Democrats Rachel WH Democrats
DBpedia-4	often impede blockade incarcerated LEASE pollutants pesticides uphe lawmakers fossils
DBpedia-5	categorized closes therapies antidepressant retrospective clinically physicians therapists randomized clinicians
DBpedia-7	JS provided Killed richness Compet Nevertheless Probably Proceedings horizontally
ETHOS-SO-3	Revolution Spread itu Million Pascal stabil Indy Georgian Figure resy
ETHOS-R-2	council Chocobo Shant uyomi aditional cumbers subur ThumbnailImage araoh Pharaoh
ETHOS-R-8	seems outlines emitted grin outline circuitry sized flips emits flipped
ETHOS-R-9	223 asel Cyrus Sith Scorpion Snape Jas Leia Ned Morty
EmoC-6	behavi checkpoints unintention crib eleph looph np mosquit blat pione
EmoC-8	depressed bullied choked stricken devastated unsuccessful cheated distraught troubled failing
EmoS-1	frightened rebellious depressed careless bullied restless reluctant distraught clumsy disgruntled
EmoS-5	obsessive crappy demonic delusions psychosis psychotic childish stupidity reckless insanity
EmoS-7	benevolent charismatic perfected volunte unintention pione innocuous fearless glamorous ruthless
EmoS-9	whispers pundits Sadly horribly curiously noticeably Sadly gaping painfully shockingly



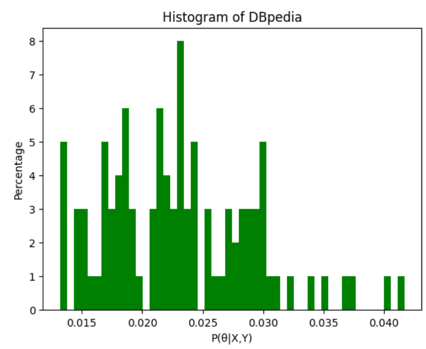
(a) SST2



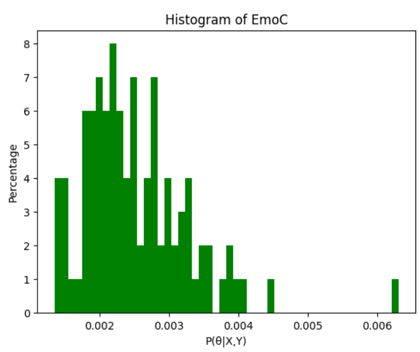
(b) FBP



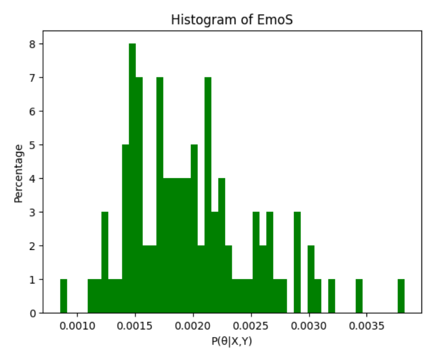
(c) COLA



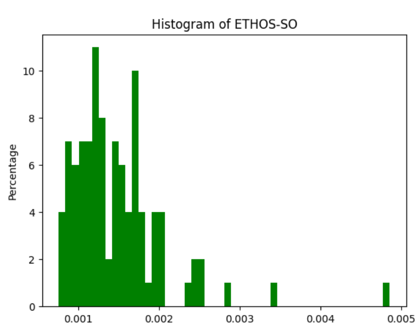
(d) DBpedia



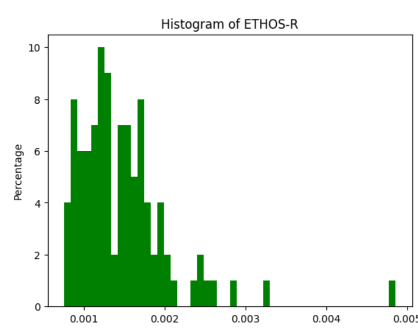
(e) EmoC



(f) EmoS



(g) ETHOS-SO



(h) RTHOS-R

Figure 10: Historgrams of the probability of train examples in each dataset predicting corresponding concept tokens.