GISTSCORE: LEARNING BETTER REPRESENTATIONS FOR IN-CONTEXT EXAMPLE SELECTION WITH GIST BOTTLENECKS

Shivanshu Gupta[1*] **Clemens Rosenbaum**[*] **Ethan R. Elenberg**[3*]
[1]University of California Irvine [3]Permanence AI
shivag5@uci.edu, cgbr@cs.umass.edu, ethan@permanence.ai

ABSTRACT

In-context Learning (ICL) is the ability of Large Language Models (LLMs) to perform new tasks when conditioned on prompts comprising a few task examples. However, ICL performance can be critically sensitive to the choice of examples. To dynamically select the best examples for every test input, we propose *Example Gisting*, a novel approach for training example encoders via supervised fine-tuning with an attention bottleneck between inputs and outputs. These *gist models* form the basis for *GistScore*, a novel metric for scoring and selecting informative examples. Further, in addition to fine-tuning gist models on each dataset, we also experiment with training a single model on a large multi-task corpus that can then be used for new tasks out-of-the-box, ensuring a training-free ICL pipeline. Evaluation with 21 datasets spanning 9 tasks and 8 diverse LLMs shows that our fine-tuned models yield state-of-the-art ICL performance with over 20% absolute gain over off-the-shelf retrievers and 5% over the best prior methods. Further, our multi-task model generalizes to new tasks and datasets and is on par or better than all baselines while being three orders faster than the strongest training-free baseline.

1 INTRODUCTION

In-context Learning (Brown et al., 2020) is a training-free approach for leveraging large language models (LLMs) for new tasks by conditioning them on a prompt comprising a few task demonstrations. However, it is highly sensitive to the choice of in-context examples (Zhao et al., 2021; Liu et al., 2022b; Lu et al., 2022). Despite extensive prior work on better example selection methods (Rubin et al., 2022; Ye et al., 2023; Mualem et al., 2024; Gupta et al., 2023), the standard approach remains to use off-the-shelf retrievers like BM25 or cosine similarity between general-purpose encoder representations Reimers & Gurevych (2019). This is because the more effective prior approaches require task or even LLM-specific training Rubin et al. (2022); Ye et al. (2023); Hu et al. (2022), eliminating the key advantage of in-context learning. More recently, Gupta et al. (2023) proposed training-free approaches based on BERTScore-Recall (BSR, Zhang et al. (2020)). However, BSR is computationally expensive, especially for long-text tasks, and is also limited by its use of general-purpose encoders.

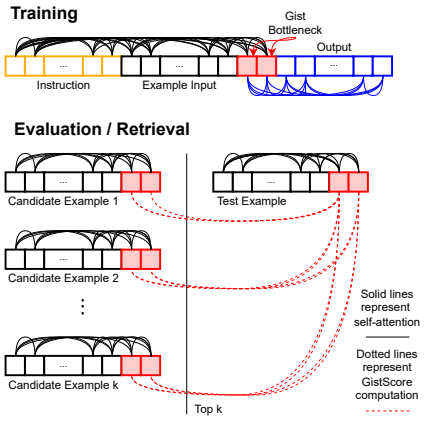


Figure 1: **Top** Example Gisting training with an attention masking bottleneck where the output can only attend to the inputs indirectly via gist tokens. **Bottom** Retrieval of the candidate examples with the highest GistScore with the test input.

This work proposes *Example Gisting*, a novel approach for training encoders for ICL example selection without feedback from a larger LLM. Based on Gisting, a recent technique by Mu et al. (2023) for compressing prompts, Example Gisting induces an attention masking bottleneck between example inputs and

[*] Work done at ASAPP

outputs (Figure 1, Top). Training with this bottleneck comprising a few *gist* tokens forces the model to store task-specific salient input information into those tokens' activations. Subsequently, the trained gist model maps both candidate examples and new test inputs into sequences of gist token embeddings that can be used with *GistScore*, a novel metric for scoring the informativeness of candidate examples (Figure 1, Bottom). By sharing BSR's functional form but operating on far fewer tokens, GistScore can be significantly faster while also being amenable to Gupta et al. (2023)'s extension to a set-level metric that can be used to find optimal sets of examples. Finally we experiment with two variations: (1) fine-tuning a gist model on each dataset for optimal performance and (2) multi-task training a single gist model on a large collection of datasets that can then be used to select in-context examples for new tasks and datasets out-of-the-box enabling a training-free ICL pipeline.

Evaluating on 21 diverse datasets spanning 9 task categories and 8 diverse LLMs, we find that example selection using GistScore dramatically improves ICL. With fine-tuning, it consistently outperforms all prior methods, including ones that leverage task or LLM-specific training, beating off-the-shelf retrievers by up to 21 points and the best trained method by 5 points on average. Further, our multi-task trained gist model recovers much of this performance gain. Applied out-of-the-box, it matches or outperforms all baselines even on held-out datasets while also being thousands of times faster than BSR. Finally, congruent to Gupta et al. (2023), we find that the set-extension of GistScore is highly effective for Semantic Parsing tasks and compositional generalization. Overall, our multi-task gist model presents the best tradeoff of performance, ease of use, and selection speed making it a promising alternative to general-purpose retrievers.

2 PRELIMINARIES

In-context Learning (ICL) Given a set of (input, output) pairs $\{(x_i, y_i)\}_{i=1}^{k}$, prompt template $\mathcal{T}$, and the test input $\mathbf{x}_{\text{test}}$, ICL using an *Inference LLM* involves prompting it with the context $\mathcal{T}(\mathbf{x}_1, \mathbf{y}_1, \ldots, \mathbf{x}_k, \mathbf{y}_k, \mathbf{x}_{\text{test}})$ to conditionally generate the test output $\mathbf{y}_{\text{test}}$. Often it is necessary to select the k examples from a larger pool of candidates $\{(x_i, y_i)\}_{i=1}^{N}$. This may be due to context-length limitation, computational efficiency, or sensitivity of LLMs to order (Liu et al., 2022b) and position of relevant in-context examples (Liu et al., 2023). The goal of in-context example selection is thus to select the $k \ll N$ most relevant examples that are most helpful for solving the test input.

Instruction Gisting was proposed by Mu et al. (2023) for compressing instruction-following prompts into shorter *gists* for efficient inference. It involves training a gist model, to simultaneously compress prompts comprising task instructions into a few gist tokens and to follow instructions encoded in those gist tokens. This is achieved by masking attention such that any attention to/from the task instruction goes through the gist tokens. Specifically, given an instruction tuning dataset $\mathcal{D}_I = \{(t_i, x_i, y_i)\}$ of instruction, (optional) input, and target tuples, the model is trained to predict y from the sequence $[t, G, x]$, where G is the sequence of special "gist" tokens added to the model vocabulary. Attention masking forces the model to predict based on the information of t encoded in the activations above G. The trained gisting model can be used to compress new instructions by feeding it the sequence $[t, G]$, precomputing the activations above G, and then prompting it with those activations instead of t.

3 METHOD

Example Gisting While Mu et al. (2023) used Gisting for efficient prompting, we hypothesise that it can also be used to extract salient information from example inputs into compressed encodings that can then be used to retrieve the most relevant examples for a given test input. Specifically, we propose *Example Gisting* which involves supervised training with a gist-bottleneck between inputs and outputs. Given a labeled dataset for target task t, $\mathcal{D}_t = \{(x_i, y_i)\}$, it finetunes a gist model to predict y_i given the inputs $[x_i, G]$, where G is the attention bottleneck comprising l gist tokens. This encourages the activations of the gist tokens to encode task-specific salient information.

Example Selection The trained example gisting model is used to select examples as follows[1]: given the gists $G(x_{\text{test}})$ of the test input and $G(z)$ for each candidate z, we use the final layer gist activations

[1]Unlike Instruction Gisting, example gists are only used to select examples for ICL, which can then be performed with any Inference LLM with the full text of the selected examples.

as *gist embeddings*, *i.e.* $\mathbf{z} = \mathbf{z}_1, \ldots \mathbf{z}_l = G(z)[-1]$ and $\mathbf{x} = \mathbf{x}_1, \ldots \mathbf{x}_l = G(x_{\text{test}})[-1]$. Then we use the following metric, called *GistScore*, to score each candidate with respect to the test input:

$$\text{GS}(x, z) = \frac{1}{l} \sum_{i=1}^{l} \max_{j=1,\ldots l} \frac{\mathbf{x}_i^T \mathbf{z}_j}{\|\mathbf{x}_i\| \|\mathbf{z}_j\|} \tag{1}$$

Finally, the top-k examples with the highest GistScore are selected for ICL. Note that GistScore shares the functional form of BERTScore-Recall (Zhang et al., 2020), and for $l = 1$, reduces to cosine similarity. $l > 1$ may be useful when a single embedding cannot encode all salient information. Further, GistScore can also be extended to a set-level metric that can be greedily optimized to select examples together as a set (see App B).

Multi-Task Training While task-specific finetuning can yield greater performance, the need for training can limit ease-of-use. Thus, we propose a multi-task training approach so the gisting model may work for new tasks without further training, preserving the training-free ICL pipeline. The key idea is to encode both the task instruction and the example input so that the model can distinguish the task and extract task-specific salient information from the input. Formally, given a collection of datasets $\mathcal{D}_M = \bigcup_{t \in T} \{(t, x, y) : (x, y) \in \mathcal{D}_t\}$ spanning tasks T, we train the model to predict y given the input sequence $[t, x, G]$ where t is the task instruction and G, the gist-bottleneck as before.

4 Experimental Setup

Methods As described in § 3, we experiment with both finetuning gist models for each dataset as well as multi-task training a single model on a large collection of datasets. We refer to GistScore-based selection using these as **GS[F, l]** and **GS[M, l]** and the set-extension as **Set-GS[F, l]** and **Set-GS[M, l]**, respectively. Here, l refers to the number of gist tokens ($l = 1$ unless specified otherwise). We use encoder-decoder LMs for both settings; `flan-t5-base` (Chung et al., 2022) for GS[F] and `flan-t5-large` for GS[M]. We compare with the following training-free baselines: (1) Rand which selects examples randomly, (2) dense retrieval using an encoder (`all-mpnet-base-v2`) from SentenceBERT (**SBERT**, Reimers & Gurevych (2019)), (3) sparse retrieval using BM25, and (4) BERTScore-Recall (**BSR**, Zhang et al. (2020)) and its set-extension (**Set-BSR**) proposed by Gupta et al. (2023) both with `deberta-large-mnli`. We additionally compare with three trained baselines: (1) **EPR** (Rubin et al., 2022), (2) **CEIL** (Ye et al., 2023), and (3) **LLM-R** (Wang et al., 2023a). EPR and LLM-R train a retrievers while CEIL trains a Determinantal Point Process (Kulesza, 2012) for set selection. See App. C.1 for details of all the methods and the multi-task collection.

Datasets We evaluate on 21 datasets spanning 9 diverse task categories and multiple languages as listed in Table 4. These include several datasets not in multi-task collection to evaluate the out-of-the-box generalization of our multi-task gist models to new tasks (*e.g.* Semantic Parsing), new datasets for seen tasks (*e.g.* WANLI), and domains (*e.g.* medical domain in MedNLI), etc. Further, in addition to IID splits, for the semantic parsing datasets (SMCalFlow and COGS), we also evaluate on compositional generalization (CG) splits. App. C.2 provides additional details about all the datasets, including splits, sample instances, selection and ICL templates, and evaluation metrics.

Inference LLMs We experiment with 8 Inference LLMs including: 6 base LLMs viz. **GPT-Neo-2.7B** (Black et al., 2021), **LLaMA-7B** and **LLaMA-13B** (Touvron et al., 2023), **Mistral** (Jiang et al., 2023), OpenAI's **Babbage** (`babbage-002`) and **Davinci** (`davinci-002`); **Zephyr** (Tunstall et al., 2023), an instruction-tuned LLM; and **StarCoder** (Li et al., 2023), a code-pretrained base LLM. We include additional details about LLMs and ICL prompt construction in App. C.3 and C.4.

5 Results

Table 1 and Figures 3 (Top) and 6a compare the performance of ICL example selection using single-token GistScore with prior training-free and trained approaches for a variety of datasets and Inference LLMs. With the exception of Semantic Parsing datasets, GS[F, 1] consistently and dramatically outperforms all baselines, beating the training-free SBERT and BSR by up to 21 and 11 points and the trained baselines, CEIL and LLM-R, by 5 and 8 points on average, respectively. In fact, even our multi-task model (GS[M, 1]) used without task-specific finetuning, matches or outperforms all baselines. Further, Figure 3 (Bottom) shows that it is also able to generalize out-of-the-box to

Selector	Neo	L7B	L13B	Mis.	Zeph.	Bab.	Dav.
RAND	38.0	46.3	48.9	56.4	58.8	39.9	52.4
BM25	46.2	53.6	57.3	64.0	65.1	45.4	57.4
SBERT	46.5	53.7	57.7	64.6	65.5	47.3	58.1
BSR	57.1	60.8	64.6	70.9	70.1	57.3	65.4
GS[M, 1]	63.5	65.8	68.1	73.6	71.7	63.1	68.4
GS[F, 1]	**68.1**	**70.1**	**71.8**	**76.5**	**74.9**	**67.3**	**71.0**

Table 1: Average 8-shot ICL with single-token GistScore v/s training-free baselines for different LLMs. See App. D for complete results for each dataset and LLM. While fine-tuning (GS[F]) yields the best performance, GS[M] also outperforms the baselines and recovers much of GS[F]'s performance despite requiring no finetuning.

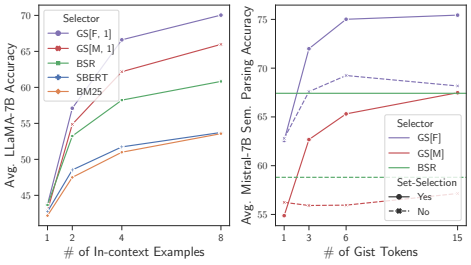


Figure 2: **Left** GS[F] and GS[M] consistently outperform baselines across varying number of in-context examples, requiring just 2 examples to surpass 8-shot ICL using SBERT and BM25. **Right** Due to their complex compositional nature, Semantic Parsing datasets benefit from additional gist tokens and set-selection. With 15 tokens, SET-GS[M] matches the average 8-shot semantic parsing ICL performance of SET-BSR, while SET-GS[F] vastly outperforms it. See Table 2 for trained baselines and Table 9 for complete results.

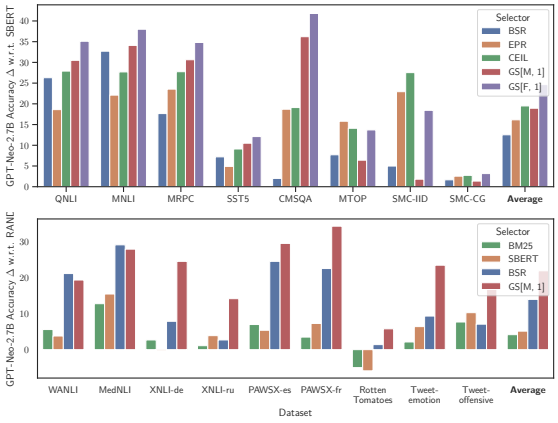


Figure 3: **Top** Single-token GistScore v/s BSR and trained baselines, EPR and CEIL with GPT-Neo-2.7B. All numbers are absolute gain in 8-shot ICL performance over SBERT except EPR and CEIL on MNLI, SST5, MRPC, and CMSQA which are with 50 examples. **Bottom** Comparison of training-free methods on held-out datasets. GS[M] generalizes without training to held-out datasets, significantly outperforming both off-the-shelf retrievers as well as the stronger but slower BSR.

Selector	SMC IID	SMC CG	COGS IID	COGS CG	MTOP	AVG
BSR	65.3	18.6	91.8	78.0	68.0	64.3
EPR	69.8	17.3			72.6	
GS[M, 1]	58.2	16.0	88.4	70.8	68.5	60.4
GS[F, 1]	69.0	14.6	89.0	75.0	71.0	63.7
SET-BSR	69.6	51.4	92.4	77.1	70.0	72.1
CEIL	71.0	31.8			73.7	
SET-GS[M, 15]	69.2	52.3	91.7	71.6	71.7	71.3
SET-GS[F, 15]	**73.7**	**53.1**	**94.7**	**81.4**	**75.5**	**75.7**

Table 2: Average 8-shot ICL using StarCoder for Semantic Parsing datasets with independent ranking (**top**) and set selection (**bottom**) methods.

held-out datasets, significantly outperforming off-the-shelf retrievers (BM25 and SBERT) as well as BSR. Additional results for all datasets and LLMs are provided in App. D. In particular, Figure 7 shows that the improvements from GistScore persist across varying number of shots and Figure 6b shows that GistScore-based selection is thousands of times faster than BSR.

While a single gist token worked well for most tasks, it may not suffice to capture all the salient information in complex compositional semantic parsing instances. Moreover, it is also known to require set-selection as opposed to independent ranking-based selection (Gupta et al., 2023). Indeed, as shown in Figure 8, set-selection of examples using the set-extension of GistScore with additional gist-tokens leads to dramatic gains for these datasets for both variants of gist models. In fact, with 15 tokens, SET-GS[F] outperforms all prior methods on semantic parsing as well (see Table 2).

6 ANALYSIS

We now analyze the improvements from GistScore. For classification tasks, we found ICL accuracy to be strongly correlated with the precision of selected examples' labels (see Figure 4). In particular, this suggests that GistScore improves ICL performance by selecting examples that share the test input's class labels. This is possible because, as shown in Figure 5, for classification tasks like QNLI

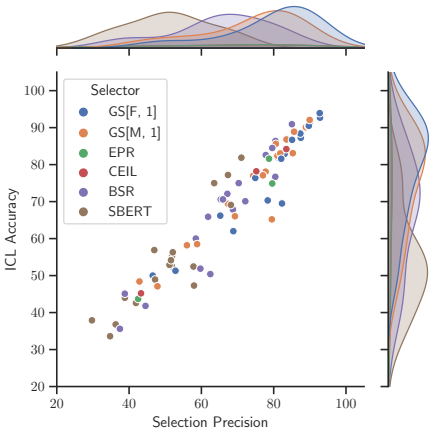


Figure 4: GPT-Neo-2.7B ICL accuracy across all classification tasks is strongly correlated with the precision of the various selectors, *i.e.* per-dataset-average of the fraction of in-context examples matching the test input's label. This suggests that retrieving such examples is the primary driver of ICL performance for these datasets.

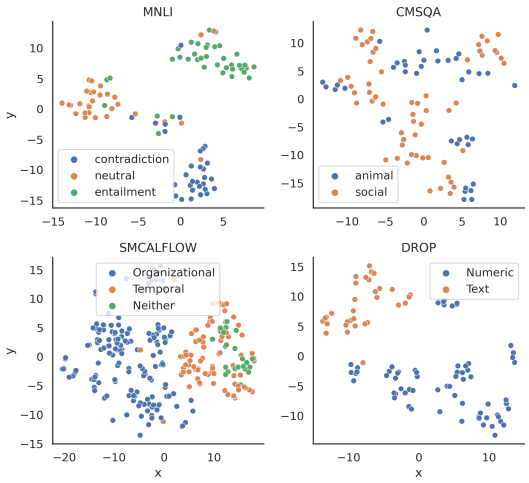


Figure 5: t-SNE plots of GS[M,1] gist embeddings. They encode task-specific information such as class labels for MNLI; whether the question is about an animal or an action (*e.g. driving car*) for CMSQA; whether the input pertains to organizational hierarchy (*e.g.* `Who is Bill's manager?`), contains temporal information (*e.g.* `Dinner at 3pm today`), or neither, for SMCalFlow; and whether the answer is numeric or textual for DROP.

and MNLI, the gist embeddings contain information of the correct label. Note that this does not necessarily mean that ICL performance for classification tasks is bounded by selection precision – as shown in Figure 9, stronger LLMs are less reliant on accurate retrieval and can improve ICL performance beyond it especially when the selector is weak.

As gist models are trained to perform the tasks, it is possible to compare against their performance directly. Table 3 shows that ICL with GistScore-based selection can yield performance exceeding that of the underlying gist model itself, especially when using stronger LLMs. This is best exemplified on tasks requiring compositional generalization and chain-of-thought reasoning (GSM8K), a known emergent capability Wei et al. (2022). This is because, as shown in Figure 5 for SMCalFlow, in these settings, gists can encode abstract task-specific salient aspects (Gupta et al., 2023) useful for selecting informative examples. We share additional analyses of gist embeddings in App. E.

	Method	SST5	QNLI	CMSQA	SMC		COGS		GSM	DROP
					CG	IID	CG	IID		
	GM[F]	53.7	85.6	64.6	0.0	64.7	45.7	99.0	0.0	32.5
Neo	**RAND**	13.0	41.9	19.0	0.0	3.3	3.8	8.1	1.7	7.7
	SBERT	37.9	44.0	18.1	1.1	31.6	26.0	34.7	2.0	12.6
	GS[F, 1]	50.0	82.0	59.9	4.2	50.0	56.3	62.4	3.1	25.4
Zephyr	**RAND**	52.3	73.4	72.5	0.0	5.9	15.4	17.7	37.9	37.0
	SBERT	51.2	72.1	71.6	13.4	50.8	39.7	55.4	35.9	46.3
	GS[F, 1]	56.1	85.2	73.0	16.1	66.8	68.5	78.0	39.0	53.6

Table 3: ICL performance v/s the performance of the gist models trained on various tasks. Here, the gist model means the full encoder-decoder model with the gist bottleneck. GistScore-based selection can improve ICL performance beyond that of the underlying gist model (GM) itself.

7 CONCLUSION

We presented Example Gisting, a novel approach for training retrievers for in-context learning through supervised fine-tuning of encoder-decoder models with a bottleneck that is forced to encode the salient information in inputs into a few tokens. We further proposed GistScore, a novel metric that uses gist encodings to evaluate informativeness of candidates for a given test input. Evaluation on diverse tasks and LLMs shows that finetuned gist models yield state-of-the-art ICL performance. Further, with multi-task training, gist models can be used out-of-the-box for new tasks and datasets, enabling am improved yet training-free ICL pipeline.

REFERENCES

Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. In-context examples selection for machine translation, 2022.

Jacob Andreas, John Bufe, David Burkett, Charles Chen, Josh Clausman, Jean Crawford, Kate Crim, Jordan DeLoach, Leah Dorner, Jason Eisner, Hao Fang, Alan Guo, David Hall, Kristin Hayes, Kellie Hill, Diana Ho, Wendy Iwaszuk, Smriti Jha, Dan Klein, Jayant Krishnamurthy, Theo Lanman, Percy Liang, Christopher H. Lin, Ilya Lintsbakh, Andy McGovern, Aleksandr Nisnevich, Adam Pauls, Dmitrij Petters, Brent Read, Dan Roth, Subhro Roy, Jesse Rusak, Beth Short, Div Slomin, Ben Snyder, Stephon Striplin, Yu Su, Zachary Tellman, Sam Thomson, Andrei Vorobev, Izabela Witoszko, Jason Wolfe, Abby Wray, Yuchen Zhang, and Alexander Zotov. Task-oriented dialogue as dataflow synthesis. *Transactions of the Association for Computational Linguistics*, 8: 556–571, 2020. doi: 10.1162/tacl_a_00333. URL https://aclanthology.org/2020.tacl-1.36.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In Trevor Cohn, Yulan He, and Yang Liu (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1644–1650, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.148. URL https://aclanthology.org/2020.findings-emnlp.148.

Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The Long-Document transformer, 2020.

Luisa Bentivogli, Bernardo Magnini, Ido Dagan, Hoa Trang Dang, and Danilo Giampiccolo. The fifth PASCAL recognizing textual entailment challenge. In *Proceedings of the Second Text Analysis Conference, TAC 2009, Gaithersburg, Maryland, USA, November 16-17, 2009*. NIST, 2009.

Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. GPT-Neo: Large scale autoregressive language modeling with meshtensorflow, 2021.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers, 2019.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models, 2022.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2924–2936, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1300. URL https://aclanthology.org/N19-1300.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. XNLI: Evaluating cross-lingual sentence representations. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2475–2485, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1269. URL https://aclanthology.org/D18-1269.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2978–2988, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1285.

William B. Dolan and Chris Brockett. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*, 2005. URL https://aclanthology.org/I05-5002.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2368–2378, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1246. URL https://aclanthology.org/N19-1246.

Mandy Guo, Joshua Ainslie, David Uthus, Santiago Ontanon, Jianmo Ni, Yun-Hsuan Sung, and Yinfei Yang. LongT5: Efficient text-to-text transformer for long sequences. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pp. 724–736, Seattle, United States, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-naacl.55.

Shivanshu Gupta, Matt Gardner, and Sameer Singh. Coverage-based example selection for in-context learning, 2023.

Christine Herlihy and Rachel Rudinger. MedNLI is not immune: Natural language inference artifacts in the clinical domain. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 1020–1027, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-short.129. URL https://aclanthology.org/2021.acl-short.129.

Yushi Hu, Chia-Hsuan Lee, Tianbao Xie, Tao Yu, Noah A. Smith, and Mari Ostendorf. In-context learning for few-shot dialogue state tracking. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 2627–2643, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023.

Najoung Kim and Tal Linzen. COGS: A compositional generalization challenge based on semantic interpretation. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9087–9105, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.731. URL https://aclanthology.org/2020.emnlp-main.731.

Alex Kulesza. Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning*, 5(2-3):123–286, 2012. doi: 10.1561/2200000044.

Itay Levy, Ben Bogin, and Jonathan Berant. Diverse demonstrations improve in-context compositional generalization, 2022.

Haoran Li, Abhinav Arora, Shuohui Chen, Anchit Gupta, Sonal Gupta, and Yashar Mehdad. MTOP: A comprehensive multilingual task-oriented semantic parsing benchmark. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty (eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 2950–2962, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.257. URL https://aclanthology.org/2021.eacl-main.257.

Raymond Li, Loubna Ben Allal, Yangtian Zi, Niklas Muennighoff, Denis Kocetkov, Chenghao Mou, Marc Marone, Christopher Akiki, Jia Li, Jenny Chim, Qian Liu, Evgenii Zheltonozhskii, Terry Yue Zhuo, Thomas Wang, Olivier Dehaene, Mishig Davaadorj, Joel Lamy-Poirier, João Monteiro, Oleh Shliazhko, Nicolas Gontier, Nicholas Meade, Armel Zebaze, Ming-Ho Yee, Logesh Kumar Umapathi, Jian Zhu, Benjamin Lipkin, Muhtasham Oblokulov, Zhiruo Wang, Rudra Murthy, Jason Stillerman, Siva Sankalp Patel, Dmitry Abulkhanov, Marco Zocca, Manan Dey, Zhihan Zhang, Nour Fahmy, Urvashi Bhattacharyya, Wenhao Yu, Swayam Singh, Sasha Luccioni, Paulo Villegas, Maxim Kunakov, Fedor Zhdanov, Manuel Romero, Tony Lee, Nadav Timor, Jennifer Ding, Claire Schlesinger, Hailey Schoelkopf, Jan Ebert, Tri Dao, Mayank Mishra, Alex Gu, Jennifer Robinson, Carolyn Jane Anderson, Brendan Dolan-Gavitt, Danish Contractor, Siva Reddy, Daniel Fried, Dzmitry Bahdanau, Yacine Jernite, Carlos Muñoz Ferrandis, Sean Hughes, Thomas Wolf, Arjun Guha, Leandro von Werra, and Harm de Vries. Starcoder: may the source be with you!, 2023.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013.

Alisa Liu, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. WANLI: Worker and AI collaboration for natural language inference dataset creation. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 6826–6847, Abu Dhabi, United Arab Emirates, December 2022a. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.508. URL https://aclanthology.org/2022.findings-emnlp.508.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for GPT-3? In Eneko Agirre, Marianna Apidianaki, and Ivan Vulić (eds.), *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 2022b. Association for Computational Linguistics. doi: 10.18653/v1/2022.deelio-1.10. URL https://aclanthology.org/2022.deelio-1.10.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts, 2023.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V. Le, Barret Zoph, Jason Wei, and Adam Roberts. The flan collection: Designing data and methods for effective instruction tuning, 2023.

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 Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *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.

Joram Meron. Simplifying semantic annotations of SMCalFlow. In Harry Bunt (ed.), *Proceedings of the 18th Joint ACL - ISO Workshop on Interoperable Semantic Annotation within LREC2022*, pp. 81–85, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.isa-1.11.

Jesse Mu, Xiang Lisa Li, and Noah Goodman. Learning to compress prompts with gist tokens. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Loay Mualem, Ethan R. Elenberg, Moran Feldman, and Amin Karbasi. Submodular minimax optimization: Finding effective sets. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, 2024.

Bo Pang and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In Kevin Knight, Hwee Tou Ng, and Kemal Oflazer (eds.), *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.

Jack W. Rae, Anna Potapenko, Siddhant M. Jayakumar, Chloe Hillier, and Timothy P. Lillicrap. Compressive transformers for long-range sequence modelling. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410.

Stephen Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. Okapi at trec. In *Text Retrieval Conference*, pp. 109–123. National Institute of Standards and Technology, 1993.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. Learning to retrieve prompts for in-context learning. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), *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.

Noam Shazeer and Mitchell Stern. Adafactor: Adaptive learning rates with sublinear memory cost. In Jennifer G. Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4603–4611. PMLR, 2018.

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 David Yarowsky, Timothy Baldwin, Anna Korhonen, Karen Livescu, and Steven Bethard (eds.), *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.

Hongjin Su, Jungo Kasai, Chen Henry Wu, Weijia Shi, Tianlu Wang, Jiayi Xin, Rui Zhang, Mari Ostendorf, Luke Zettlemoyer, Noah A. Smith, and Tao Yu. Selective annotation makes language models better few-shot learners, 2022.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4149–4158, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1421. URL https://aclanthology.org/N19-1421.

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

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. Zephyr: Direct distillation of lm alignment, 2023.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In Tal Linzen, Grzegorz Chrupała, and Afra Alishahi (eds.), *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 353–355, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5446. URL https://aclanthology.org/W18-5446.

Liang Wang, Nan Yang, and Furu Wei. Learning to retrieve in-context examples for large language models, 2023a.

Xinyi Wang, Wanrong Zhu, Michael Saxon, Mark Steyvers, and William Yang Wang. Large language models are implicitly topic models: Explaining and finding good demonstrations for in-context learning, 2023b.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641, 2019. doi: 10.1162/tacl_a_00290. URL https://aclanthology.org/Q19-1040.

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*, 2022. ISSN 2835-8856. Survey Certification.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023.

Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In Marilyn Walker, Heng Ji, and Amanda Stent (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1112–1122, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1101. URL https://aclanthology.org/N18-1101.

Zhiyong Wu, Yaoxiang Wang, Jiacheng Ye, and Lingpeng Kong. Self-adaptive in-context learning: An information compression perspective for in-context example selection and ordering, 2023.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks, 2023.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3687–3692, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1382. URL https://aclanthology.org/D19-1382.

Jiacheng Ye, Zhiyong Wu, Jiangtao Feng, Tao Yu, and Lingpeng Kong. Compositional exemplars for in-context learning, 2023.

Xi Ye, Srinivasan Iyer, Asli Celikyilmaz, Ves Stoyanov, Greg Durrett, and Ramakanth Pasunuru. Complementary explanations for effective in-context learning, 2022.

Pengcheng Yin, Hao Fang, Graham Neubig, Adam Pauls, Emmanouil Antonios Platanios, Yu Su, Sam Thomson, and Jacob Andreas. Compositional generalization for neural semantic parsing via span-level supervised attention. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou

(eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2810–2823, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.225. URL https://aclanthology.org/2021.naacl-main.225.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pp. 649–657, 2015.

Yuan Zhang, Jason Baldridge, and Luheng He. PAWS: Paraphrase adversaries from word scrambling. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 1298–1308, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1131. URL https://aclanthology.org/N19-1131.

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, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12697–12706. PMLR, 2021.

A Additional Related Work

In-Context Learning Numerous approaches have been proposed for selecting in-context examples: (1) training retrievers with feedback from a larger LLM Rubin et al. (2022); Wang et al. (2023a), (2) selecting diverse examples with reduced redundancy (Su et al., 2022; Levy et al., 2022; Agrawal et al., 2022; Ye et al., 2022), (3) selecting examples that minimize the entropy of the LLM's output distribution for the test input (Lu et al., 2022; Wu et al., 2023), (4) Bayesian inference (Wang et al., 2023b), and (5) selecting examples as a set (Gupta et al., 2023; Ye et al., 2023; Mualem et al., 2024). The most effective of these either require task and/or LLM-specific training Rubin et al. (2022); Wang et al. (2023a); Ye et al. (2023) or are computationally inefficient Gupta et al. (2023).

Attention and Memory Both Example and Instruction Gisting Mu et al. (2023) leverage attention bottlenecks to encode pertinent information in a few tokens, thereby acting as a memory. This is related to past work on improving memory and long-range sequence modeling with Transformers (Dai et al., 2019; Child et al., 2019; Beltagy et al., 2020; Rae et al., 2020). In particular, similar to the specialization of gist tokens, Guo et al. (2022) and Xiao et al. (2023) model long sequence dependencies using specific tokens that act as a shared global memory, rather than passthrough tokens. Additionally, the sparsity induced by attention-masking in gisting is related to various sparse attention methods that have been proposed to improve Transformer efficiency. For example, Dai et al. (2019) use block-wise dense local attention combined with recursive attention to the previous attention block. Child et al. (2019) and Beltagy et al. (2020) use different forms of sliding (and strided) attention.

B Set Extension

Gupta et al. (2023) proposed a class of metrics called Coverage Measures for evaluating the relevance of a candidate example z with respect to the test input x_{test} as a recall of salient aspects with the following form,

$$\text{cover}\left(x_{\text{test}}, z\right) = \sum_{s \in S_{x_{\text{test}}}} \text{c}(s, z) \tag{2}$$

where the set of salient aspects $S_{x_{\text{test}}}$ and the coverage of individual aspects $\text{c}(s, z)$ would be defined differently for every metric. Such metrics can be extended to a sub-modular, and hence greedily optimizable, set-level metrics for evaluating sets of examples Z as follows:

$$\text{setcov}\left(x_{\text{test}}, Z\right) = \sum_{s \in S_{x_{\text{test}}}} \max_{z \in Z} \text{c}(s, z) \tag{3}$$

For $l > 1$ GistScore, as defined in Eq. 1, has the form of Eq. 2 for $S_{x_{\text{test}}} = \{1, \ldots, L\}$ and $c(s, z) = \frac{1}{l} \max_{j=1,\ldots l} \frac{\mathbf{x}_s^T \mathbf{z}_j}{\|\mathbf{x}_s\| \|\mathbf{z}_j\|}$. Thus, its set-extension can be defined as:

$$\text{Set-GS}_{l>1}(x, Z) = \frac{1}{l} \sum_{i=1}^{l} \max_{z \in Z} \max_{j=1,\ldots l} \frac{\mathbf{x}_i^T \mathbf{z}_j}{\|\mathbf{x}_i\| \|\mathbf{z}_j\|} \tag{4}$$

For $l = 1$, GistScore reduces to cosine similarity. Hence, we use Gupta et al. (2023)'s extension for cosine similarity in this case which assumes $S_{x_{\text{test}}} = \{1, \ldots, d\}$ where d is the embedding size and $c(s, z) = \frac{\mathbf{x}_1[i] \mathbf{z}_1[i]}{\|\mathbf{x}_1\| \|\mathbf{z}_1\|}$:

$$\text{Set-GS}_{l=1}(x, Z) = \sum_{i=1}^{d} \max_{z \in Z} \frac{\mathbf{x}_1[i] \mathbf{z}_1[i]}{\|\mathbf{x}_1\| \|\mathbf{z}_1\|} \tag{5}$$

C Experimental Setup

C.1 Methods

C.1.1 GistScore

We use encoder-decoder models for both task fine-tuned and multi-task pretrained gist models. This means that after training, we can drop the decoder and only keep the encoder for computing exmaple gists. We experiment with the following different variants of Gist LM-based retrievers:

Finetuned Gisting models (GS[F]) In this setting, we fine-tune Flan-T5-base Chung et al. (2022) models to produce gists of varying lengths on each individual dataset using the procedure described in § 3. For each dataset, we use the entire train set with instances longer than 500 tokens filtered out for computational efficiency. For early stopping, we compute Rouge-L (Lin, 2004) for DROP and GSM8K and Exact-Match Accuracy for the remaining datasets on up to 1000 random instances from the validation set. All training was done with batch size 36 for up to 40000 steps with early stopping with the Adafactor optimizer (Shazeer & Stern, 2018) and a constant learning rate of 5e-5.

Multi-task Pre-trained Gist Model (GS[M]) For this setting, we train using a large multi-task collection of prompts subsampled from the FLAN 2022 collection (Longpre et al., 2023) of 15M prompts from over 473 datasets and 146 task categories. Specifically, we take zero-shot prompts at most 256 tokens long and further subsample at most 10,000 prompts for every task category, yielding roughly 5M prompts. We use 95% of this sub-collection for training and 1000 random instances from the remaining 5% for early stopping with Rouge-L (Lin, 2004) as the metric. To assess effect from varying gist lengths, we train four models that can gist to 1, 3, 6, and 15 tokens. Each model was trained using the Adafactor optimizer (Shazeer & Stern, 2018) on an NVIDIA A10G GPU with a batch size of 4 and 64 gradient accumulation steps for an effective batch size of 256. The learning rate was kept constant at 5e-4.

C.1.2 BASELINES

In addition to randomly selecting in-context examples (**RAND**), we compare with the following training-free rankind-based selection baselines: (1) dense retrieval using a general-purpose encoder (`all-mpnet-base-v2`) from SentenceBERT library (**SBERT**, Reimers & Gurevych (2019)), (2) sparse-retrieval using Okapi variant (Robertson et al., 1993) of **BM25** from the `rank_bm25`[2] library, and (3) BERTScore-Recall (**BSR**, Zhang et al. (2020)) using `deberta-large-mnli`(Williams et al., 2018) as encoder. We also compare with the set-extension of BSR (**SET-BSR**) proposed by Gupta et al. (2023) for selecting optimal sets of examples.

Further, we compare with three methods that leverage training with feedback from an Inference LLM: (1) **EPR** (Rubin et al., 2022) which uses LLM perplexity (GPT-Neo-2.7B) to train a dense retriever for each dataset, (2) **CEIL** (Ye et al., 2023) which uses EPR and feedback from an LLM to train a Determinantal Point Process (Kulesza, 2012) for each dataset that is used to select examples as a set, and (3) **LLM-R** (Wang et al., 2023a) which uses feedback from LLaMA-7B to train a reward model for evaluating candidate examples that is distilled into a dense retriever used for example selection. For EPR and CEIL, we compare with the 8-shot results reported in Gupta et al. (2023), if available, and the 50-shot results from Ye et al. (2023), otherwise. For LLM-R, we use their 8-shot ICL results with LLaMA-7B. Being multi-task trained, LLM-R can also be applied to held-out tasks; however, as Wang et al. (2023a)'s held-out tasks are included in our multi-task collection, we only compare with it on its held-in datasets.

C.2 DATASETS

We evaluate on 21 datasets spanning 9 diverse task categories and multiple languages as listed in Table 4. These include several datasets not in FLAN-2022 to evaluate the out-of-the-box generalization of our multi-task gist models to new tasks, datasets, domains, etc.[3] In particular, MedNLI (Herlihy & Rudinger, 2021) and TweetEval (Barbieri et al., 2020) evaluate on held-out domains (Medical and Tweets) while XNLI (Conneau et al., 2018) and PAWSX (Yang et al., 2019) evaluate generalization to non-English languages.

Splits For all datasets other than XNLI, PAWSX, COGS, and SMCalFlow, we use the standard IID splits. For XNLI which is a multilingual NLI dataset, we use the German and Russian splits. For PAWSX which is a multilingual paraphrase detection dataset, we use the French and Spanish splits. For COGS, we evaluate on the standard IID and compositional generalization evaluation sets. For SMCalFlow we evaluate on the IID and compositional generalization splits from Yin et al. (2021) as described below. Following prior work (Gupta et al., 2023; Rubin et al., 2022; Ye et al., 2023), for

[2] `https://github.com/dorianbrown/rank_bm25`

[3] Most of our held-in datasets also require the multi-task models to generalize to new prompt templates as our ICL prompt templates differ from FLAN-2022's.

each split, we use up to 44,000 random instances from the train set as the candidate pool and evaluate on up to 1000 instances from the validation set if available, and the test set otherwise.

SMCalFlow (Andreas et al., 2020) is a dataset of task-oriented natural language dialogs about calendars, weather, places, and people paired with executable dataflow programs. SMCalFlow-CS (Yin et al., 2021) is a subset of SMCalFlow containing single-turn dialogs involving two domains (organization structure and calendar event creation), each having its own set of program symbols with two types of test sets: a cross-domain (C) test set containing only instances where both domains appear and meant to test for compositional generalization, and a single-domain (S) test set contains instances with only single-domain for in-distribution evaluation. For compositional evaluation, we use the 32-C split, a few-shot cross-domain split where the training set includes 32 cross-domain examples. For our IID evaluation, following Levy et al. (2022), we use the 8-S split. Additionally, we use the programs with the simplified syntax provided by Meron (2022).

Templates Tables 5, 6, and 7 contain the textual templates we use to linearize the instances for example selection and ICL for the various datasets. The complete ICL prompt is constructed described in App. C.4.

Task Category	Dataset
Natural Language Inference	QNLI Wang et al. (2018) MNLI Williams et al. (2018) RTE Bentivogli et al. (2009) WANLI Liu et al. (2022a) XNLI Conneau et al. (2018) MedNLI Herlihy & Rudinger (2021)
Paraphrase Detection	MRPC Dolan & Brockett (2005) QQP Wang et al. (2018) PAWS Zhang et al. (2019) PAWSX Yang et al. (2019)
Question Answering	DROP Dua et al. (2019) BoolQ Clark et al. (2019)
Semantic Parsing	SMCalFlow (SMC, Andreas et al. (2020) MTOP Li et al. (2021) COGS Kim & Linzen (2020)
Sentiment Analysis	SST2 Socher et al. (2013) SST5 Socher et al. (2013) Rotten Tomatoes Pang & Lee (2005) TweetEval-emotion Barbieri et al. (2020)
Commonsense	CommonSenseQA (CMSQA, Talmor et al. (2019))
CoT	GSM8K Wei et al. (2023)
Summarization	AGNews Zhang et al. (2015)
Misc	TweetEval-offensive Barbieri et al. (2020) CoLA Warstadt et al. (2019)

Table 4: Datasets used in this work. Red highlights datasets held-out from our multi-task collection. We use the German and Russian splits of XNLI and Spanish and French of PAWSX, and both IID and Compositional Generalization (CG) splits of SMCalFlow and COGS.

Evaluation Metric We report Exact-Match Accuracy for all the Semantic Parsing datasets and Accuracy for the remaining datasets.

C.3 Inference LLMs

We experiment with eight diverse Inference LLMs including: 6 base LLMs viz. **GPT-Neo-2.7B** (Black et al., 2021), **LLaMA-7B** and **LLaMA-13B** (Touvron et al., 2023), **Mistral**[4] (Jiang et al., 2023), OpenAI's **Babbage** (`babbage-002`) and **Davinci** (`davinci-002`); **Zephyr**[5] (Tunstall et al., 2023), an instruction-tuned and aligned LLM; and **StarCoder**[6] (Li et al., 2023), a code-pretrained base LLM. GPT-Neo-2.7B, LLaMA-7B, and LLaMA-13B have context windows of 2048, StarCoder of 7000, Mistral and Zephyr of 8192, and Babbage and Davinci of 16384.

C.4 ICL Prompt Construction

Following prior work (Rubin et al., 2022; Gupta et al., 2023), for k-shot ($k = 8$ unless specified otherwise) ICL with any given dataset, example selection method, and LLM, we construct the ICL prompt by selecting k (or fewer depending on LLM context window) examples from the train split. (2) ordering the examples by increasing relevance so that the more relevant examples are closer to the test input, (3) linearizing the ordered examples and the test input using the dataset's ICL example template in Tables 5, 6, and 7, and (4) concatenating the linearizations using `\n\n` as the separator For set-selection methods (Set-BSR and Set-GS), the examples are ordered by their corresponding instance-level score.

[4] `https://hf.co/mistralai/Mistral-7B-v0.1`
[5] `https://hf.co/HuggingFaceH4/zephyr-7b-alpha`
[6] `https://hf.co/bigcode/starcoder`

Dataset	Selector Example Template	ICL Example Template
SMCalFlow	`Translate this sentence into a logical form representing its meaning: Great , thanks ! I am going to need a meeting with Karen , Jim , and Pam tomorrow before noon .` `Logical Form:`	`Great , thanks ! I am going to need a meeting with Karen , Jim , and Pam tomorrow before noon . CreateEvent(AND(with_attendee(" Pam "),with_attendee(" Karen "), with_attendee(" Jim "),starts_at( OnDateBeforeTime(date=Tomorrow(), time=Noon()))))`
MTOP	`Translate this sentence into a logical form representing its meaning: call Nicholas and Natasha` `Logical Form:`	`call Nicholas and Natasha [IN: CREATE_CALL [SL:CONTACT Nicholas ] [SL:CONTACT Natasha ] ]`
COGS	`Translate this sentence into a logical form representing its meaning: Liam hoped that a box was burned by a girl .` `Logical Form:`	`Liam hoped that a box was burned by a girl . hope ( agent = Liam , ccomp = burn ( theme = box , agent = girl ) )`

QNLI	`As of that day, the new constitution heralding the Second Republic came into force.`
	`Can we know "What came into force after the new constitution was herald?" given the above sentence (Yes or No)?`
		`Question: What came into force after the new constitution was herald?`
		`Sentence: As of that day, the new constitution heralding the Second Republic came into force.`
		`Answer: Yes`

MNLI	`Premise: The new rights are nice enough`
	`Does the above premise entail the hypothesis that "Everyone really likes the newest benefits " (Yes, Maybe, or No)?`
	`Answer:`
		`Premise: The new rights are nice enough`
		`Hypothesis: Everyone really likes the newest benefits`
		`Answer: Maybe`

RTE	`Dana Reeve, the widow of the actor Christopher Reeve, has died of lung cancer at age 44, according to the Christopher Reeve Foundation.`
	`Based on the above paragraph can we conclude that "Christopher Reeve had an accident." (Yes or No)?`
		`Premise: Dana Reeve, the widow of the actor Christopher Reeve, has died of lung cancer at age 44, according to the Christopher Reeve Foundation.`
		`Hypothesis: Christopher Reeve had an accident.?`
		`Answer: No`

MedNLI	`Premise: No history of blood clots or DVTs, has never had chest pain prior to one week ago.`
	`Is the hypothesis that "Patient has angina." an entailment, contradiction or neutral with respect to the above premise?`
	`Answer:`
		`Premise: No history of blood clots or DVTs, has never had chest pain prior to one week ago.`
		`Hypothesis: Patient has angina.`
		`Answer: Yes`

WANLI	`Premise: In the past, I have found that there is no point in making a speech unless you have prepared it .`
	`Is the hypothesis that "You should prepare a speech." an entailment, contradiction or neutral with respect to the above premise?`
	`Answer:`
		`Premise: In the past, I have found that there is no point in making a speech unless you have prepared it .`
		`Hypothesis: You should prepare a speech .`
		`Answer: Yes`

XNLI	`Premise: Et il a dit, maman, je suis à la maison.`
	`Is the hypothesis that "Il a appelé sa mère dès que le bus scolaire l'a déposé." an entailment, contradiction or neutral with respect to the above premise?`
	`Answer:`
		`Premise: Et il a dit, maman, je suis à la maison.`
		`Hypothesis: Il a appelé sa mère dès que le bus scolaire l'a déposé.`
		`Answer: No`

Table 5: The example templates we use for example selection and in-context learning for the various datasets. See also Tables 6, 7 and 8.

Dataset	Selector Example Template	ICL Example Template

SST5	`Review: in his first stab at the form , jacquot takes a slightly anarchic approach that works only sporadically .`
	`Does the review above see the movie as terrible, bad, OK, good, or great?`
	`Answer:`
		`Review: in his first stab at the form , jacquot takes a slightly anarchic approach that works only sporadically .`
		`Sentiment: OK`

SST2	`Review: it 's a charming and often affecting journey .`
	`Is the sentiment of the above review Negative or Positive?`
	`Answer:`
		`Review: it 's a charming and often affecting journey .`
		`Sentiment: Positive`

Rotten Tomatoes	`Review: compassionately explores the seemingly irreconcilable situation between conservative christian parents and their estranged gay and lesbian children .`
	`Is the sentiment of the above review Negative or Positive?`
	`Answer:`
		`Review: compassionately explores the seemingly irreconcilable situation between conservative christian parents and their estranged gay and lesbian children .`
		`Sentiment: Positive`

MRPC	`Sentence 1: He said the foodservice pie business doesn 't fit the company 's long-term growth strategy .`
	`Sentence 2: " The foodservice pie business does not fit our long-term growth strategy .`
	`Do the above sentences convey the same meaning? Yes or No.`
	`Answer:`
		`Sentence 1: He said the foodservice pie business doesn 't fit the company 's long-term growth strategy .`
		`Sentence 2: " The foodservice pie business does not fit our long-term growth strategy .`
		`Answer: Yes`

PAWS	`Sentence 1: Bradd Crellin represented BARLA Cumbria on a tour of Australia with 6 other players representing Britain , also on a tour of Australia .`
	`Sentence 2: Bradd Crellin also represented BARLA Great Britain on a tour through Australia on a tour through Australia with 6 other players representing Cumbria .`
	`Are these sentences paraphrases of each other? Yes or No.`
	`Answer:`
		`Sentence 1: Bradd Crellin represented BARLA Cumbria on a tour of Australia with 6 other players representing Britain , also on a tour of Australia .`
		`Sentence 2: Bradd Crellin also represented BARLA Great Britain on a tour through Australia on a tour through Australia with 6 other players representing Cumbria .`
		`Answer: No`

QQP	`Question 1: Why are African-Americans so beautiful?`
	`Question 2: Why are hispanics so beautiful?`
	`Are Questions 1 and 2 asking the same thing? Yes or No.`
	`Answer:`
		`Question 1: Why are African-Americans so beautiful?`
		`Question 2: Why are hispanics so beautiful?`
		`Answer: No`

PAWSX	`Sentence 1: El Consejo Shawnee Trail nació de la unión entre el Consejo Four Rivers y el Consejo Audubon.`
	`Sentence 2: El Consejo de caminos de los Shawnee se formó por la fusión del Consejo de Four Rivers y el Consejo de Audubon.`
	`Are the above sentences paraphrases of each other? Yes or No.`
	`Answer:`
		`Sentence 1: El Consejo Shawnee Trail nació de la unión entre el Consejo Four Rivers y el Consejo Audubon.`
		`Sentence 2: El Consejo de caminos de los Shawnee se formó por la fusión del Consejo de Four Rivers y el Consejo de Audubon.`
		`Answer: Yes`

Table 6: The example templates we use for example selection and in-context learning for the various datasets. See also Table 5, 7 and 8.

Dataset	Selector Example Template	ICL Example Template

CMSQA	`Select one of the choices that best answers the following question:`
	`Question: A revolving door is convenient for two direction travel, but it also serves as a security measure at a what?`
	`Option A: bank`
	`Option B: library`
	`Option C: department store`
	`Option D: mall`
	`Option E: new york`
	`Answer:`
		`Question: A revolving door is convenient for two direction travel, but it also serves as a security measure at a what?`
		`Option A: bank`
		`Option B: library`
		`Option C: department store`
		`Option D: mall`
		`Option E: new york`
		`Answer: A`

AGNews	`Classify the following news article into one of these categories: World, Sports, Business, Technology.`
	`Fears for T N pension after talks Unions representing workers at Turner Newall say they are ' disappointed' after talks with stricken parent firm Federal Mogul .`
	`Category:`
		`Article: Fears for T N pension after talks Unions representing workers at Turner Newall say they are ' disappointed' after talks with stricken parent firm Federal Mogul .`
		`Category: Business`

GSM8K	`Give the step-by-step reasoning process and then the final answer.Janet's ducks lay 16 eggs per day. She eats three for breakfast every morning and bakes muffins for her friends every day with four. She sells the remainder at the farmers' market daily for $2 per fresh duck egg. How much in dollars does she make every day at the farmers' market?`	`Question: Janet's ducks lay 16 eggs per day. She eats three for breakfast every morning and bakes muffins for her friends every day with four. She sells the remainder at the farmers' market daily for $2 per fresh duck egg. How much in dollars does she make every day at the farmers' market?`
		`Solution: Janet sells 16 - 3 - 4 = <<16-3-4=9>>9 duck eggs a day.`
		`She makes 9 * 2 = $<<9*2=18>>18 every day at the farmer's market.`
		`#### 18`

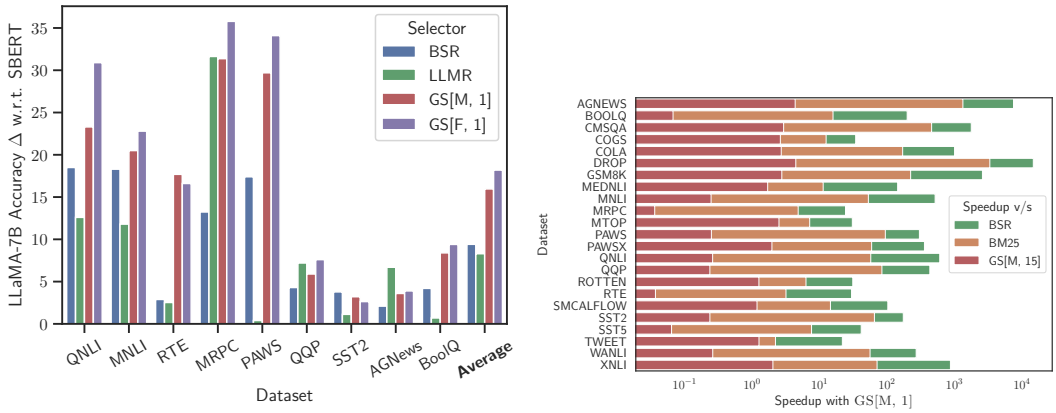
CoLA	`Is the following sentence grammatical ( Yes or No)?`
	`The sailors rode the breeze clear of the rocks.`
	`Answer:`
		`Sentence: The sailors rode the breeze clear of the rocks.`
		`Answer: Yes`

TweetEval	`Classify the emotion in the following tweet as one of anger, joy, optimism, or sadness..`
	`Tweet: @user @user Oh, hidden revenge and anger...I rememberthe time,she rebutted you.`
	`Answer:`
		`Tweet: @user @user Oh, hidden revenge and anger...I rememberthe time,she rebutted you.`
		`Answer: A`

Table 7: The example templates we use for example selection and in-context learning for the various datasets. See also Tables 5, 6 and 8.

Dataset	Selector Example Template	ICL Example Template
DROP	Hoping to rebound from their loss to the Patriots, the Raiders stayed at home for a Week 16 duel with the Houston Texans. Oakland would get the early lead in the first quarter as quarterback JaMarcus Russell completed a 20-yard touchdown pass to rookie wide receiver Chaz Schilens. The Texans would respond with fullback Vonta Leach getting a 1-yard touchdown run, yet the Raiders would answer with kicker Sebastian Janikowski getting a 33-yard and a 30-yard field goal. Houston would tie the game in the second quarter with kicker Kris Brown getting a 53-yard and a 24-yard field goal. Oakland would take the lead in the third quarter with wide receiver Johnnie Lee Higgins catching a 29-yard touchdown pass from Russell, followed up by an 80-yard punt return for a touchdown. The Texans tried to rally in the fourth quarter as Brown nailed a 40-yard field goal, yet the Raiders' defense would shut down any possible attempt. How many field goals did both teams kick in the first half? Answer:	Passage: Hoping to rebound from their loss to the Patriots, the Raiders stayed at home for a Week 16 duel with the Houston Texans. Oakland would get the early lead in the first quarter as quarterback JaMarcus Russell completed a 20-yard touchdown pass to rookie wide receiver Chaz Schilens. The Texans would respond with fullback Vonta Leach getting a 1-yard touchdown run, yet the Raiders would answer with kicker Sebastian Janikowski getting a 33-yard and a 30-yard field goal. Houston would tie the game in the second quarter with kicker Kris Brown getting a 53-yard and a 24-yard field goal. Oakland would take the lead in the third quarter with wide receiver Johnnie Lee Higgins catching a 29-yard touchdown pass from Russell, followed up by an 80-yard punt return for a touchdown. The Texans tried to rally in the fourth quarter as Brown nailed a 40-yard field goal, yet the Raiders' defense would shut down any possible attempt. Question: How many field goals did both teams kick in the first half? Answer: 2
BoolQ	Ethanol fuel -- All biomass goes through at least some of these steps: it needs to be grown, collected, dried, fermented, distilled, and burned. All of these steps require resources and an infrastructure. The total amount of energy input into the process compared to the energy released by burning the resulting ethanol fuel is known as the energy balance (or ``energy returned on energy invested''). Figures compiled in a 2007 report by National Geographic Magazine point to modest results for corn ethanol produced in the US: one unit of fossil-fuel energy is required to create 1.3 energy units from the resulting ethanol. The energy balance for sugarcane ethanol produced in Brazil is more favorable, with one unit of fossil-fuel energy required to create 8 from the ethanol. Energy balance estimates are not easily produced, thus numerous such reports have been generated that are contradictory. For instance, a separate survey reports that production of ethanol from sugarcane, which requires a tropical climate to grow productively, returns from 8 to 9 units of energy for each unit expended, as compared to corn, which only returns about 1.34 units of fuel energy for each unit of energy expended. A 2006 University of California Berkeley study, after analyzing six separate studies, concluded that producing ethanol from corn uses much less petroleum than producing gasoline. does ethanol take more energy make that produces (yes or no) Answer:	Passage: Ethanol fuel -- All biomass goes through at least some of these steps: it needs to be grown, collected, dried, fermented, distilled, and burned. All of these steps require resources and an infrastructure. The total amount of energy input into the process compared to the energy released by burning the resulting ethanol fuel is known as the energy balance (or ``energy returned on energy invested''). Figures compiled in a 2007 report by National Geographic Magazine point to modest results for corn ethanol produced in the US: one unit of fossil-fuel energy is required to create 1.3 energy units from the resulting ethanol. The energy balance for sugarcane ethanol produced in Brazil is more favorable, with one unit of fossil-fuel energy required to create 8 from the ethanol. Energy balance estimates are not easily produced, thus numerous such reports have been generated that are contradictory. For instance, a separate survey reports that production of ethanol from sugarcane, which requires a tropical climate to grow productively, returns from 8 to 9 units of energy for each unit expended, as compared to corn, which only returns about 1.34 units of fuel energy for each unit of energy expended. A 2006 University of California Berkeley study, after analyzing six separate studies, concluded that producing ethanol from corn uses much less petroleum than producing gasoline. Question: does ethanol take more energy make that produces Answer: no

Table 8: The example templates we use for example selection and in-context learning for the various datasets. See also Tables 5, 6 and 7.



(a) (b)

Figure 6: **(a)** Single-token GistScore v/s BSR and LLM-R with LLaMA-7B. All numbers are absolute gain in 8-shot ICL performance over SBERT. Both GS[F, 1] and GS[M, 1] consistently outperform LLM-R. **(b)** Example selection using GistScore (GS[M, 1]) is up to three orders faster than BSR, two orders faster than the Python implementation of BM25, and scales well with the number of gist tokens.

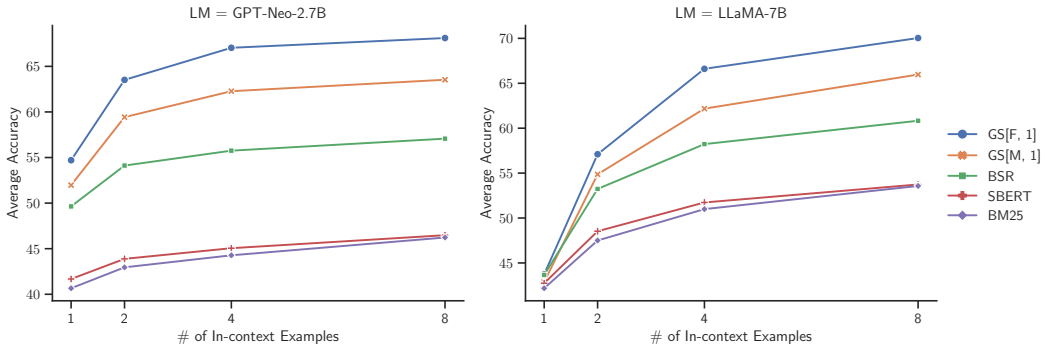


Figure 7: Average ICL performance with GPT-Neo-2.7B (top) and LLaMA-7B (bottom) for varying number of in-context examples. Both GS[F] and GS[M] are consistently better, and both surpass 8-shot ICL using SBERT and BM25 with just 2 examples!

D Additional Results

Results for GistScore-variations and all baselines Tables 10, 11, 12, 13, 14, 15 and 16 show 8-shot ICL results for all the datasets with GPT-Neo-2.7B, LLaMA-7B, LLaMA-13B, Mistral, Zephyr, Babbage, and Davinci, respectively.

Set-selection using Set-GS Figure 8 and Table 9 compare performance for different number of gist tokens and set-selection for different LLMs.

Varying number of shots The gains from GistScore persist across varying number of in-context examples as shown in Figure 7. In fact, with just 2 examples, GS[F, 1] outperforms 8-shots retrieved using general-purpose retrievers.

Impact of gist model size Table 13 shows results for GistScore-based selection using a larger multi-task gist model based on `flan-t5-xl`showing that a stronger gist model can further improve ICL performance.

Selection Speed Despite sharing its functional form and hence quadratic time-complexity in number of tokens, GistScore can be faster than BSR as it compares only a few gist tokens. Figure 6b shows that this yields thousands of times faster selection with single-token GistScore compared to BSR,

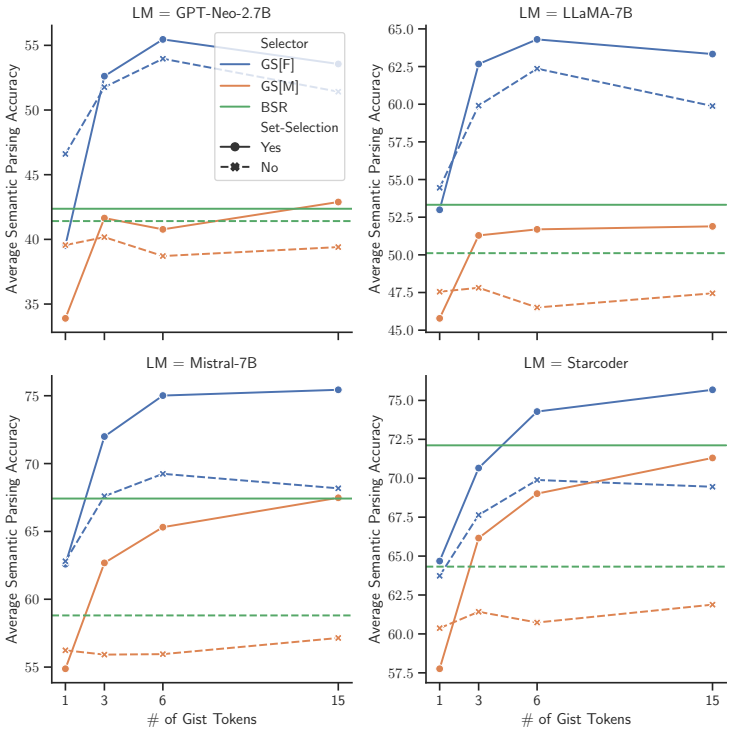


Figure 8: 8-shot ICL with different LLMs on semantic parsing datasets using BSR, GS[M], and GS[F] with varying number of gist tokens and set extension. Due to their complex compositional nature, Semantic Parsing datasets benefit from additional gist tokens and set-selection. With 15 tokens, SET-GS[M] matches the average 8-shot semantic parsing ICL performance of SET-BSR, while SET-GS[F] vastly outperforms it. See Table 2 for trained baselines and Table 9 for complete results.

which took over 20 seconds per test input for some datasets (see Table 17). Further, due to GPU acceleration, we found GistScore to be significantly faster than even BM25.

E ANALYSIS

Effect of Selection Precision Figure 9 compares ICL accuracy with selection precision, *i.e.*, the fraction of labels with the test input's label, for classification tasks with fixed label sets and different LLMs. While the ICL accuracy of all LLMs improves with more accurate selection, larger LLMs are less reliant on it.

Gist Embeddings encode salient aspects Figure 10 shows t-SNE visualizations of salient information in gist embeddings for additional datasets. Figure 11 shows that the salient aspects seen in t-SNE visualizations in Figures 5 and 10 can also be observed in PCA visualizations.

Gist tokens are different from standard tokens Figure 12 qualitatively compares gist token embeddings with ordinary token embeddings through 3 types of pairwise distance distributions: NLP x NLP, Gist x Gist, and NLP x Gist. Clearly, gist tokens are embedded into a different geometry when compared to ordinary language tokens.

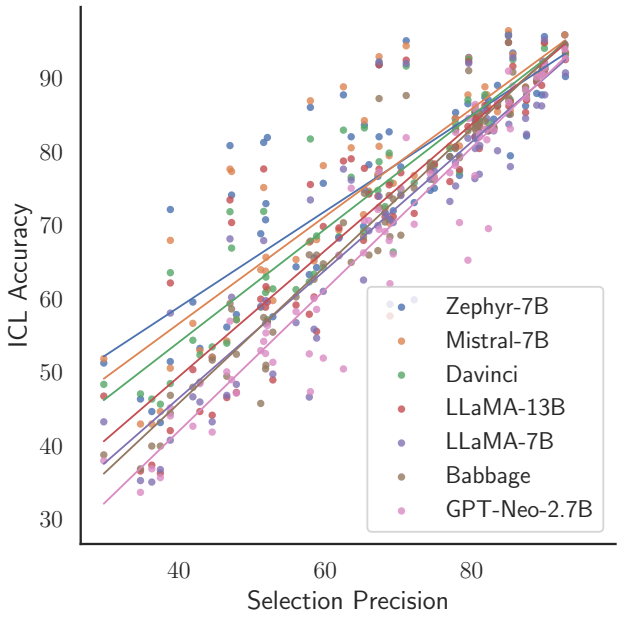


Figure 9: ICL accuracy v/s selection precision *i.e.* the fraction of in-context examples with the test label for the various classification datasets with fixed label sets, selectors, and LLMs. While the ICL accuracy of all LLMs improves with more accurate selection, larger LLMs are less reliant on it.

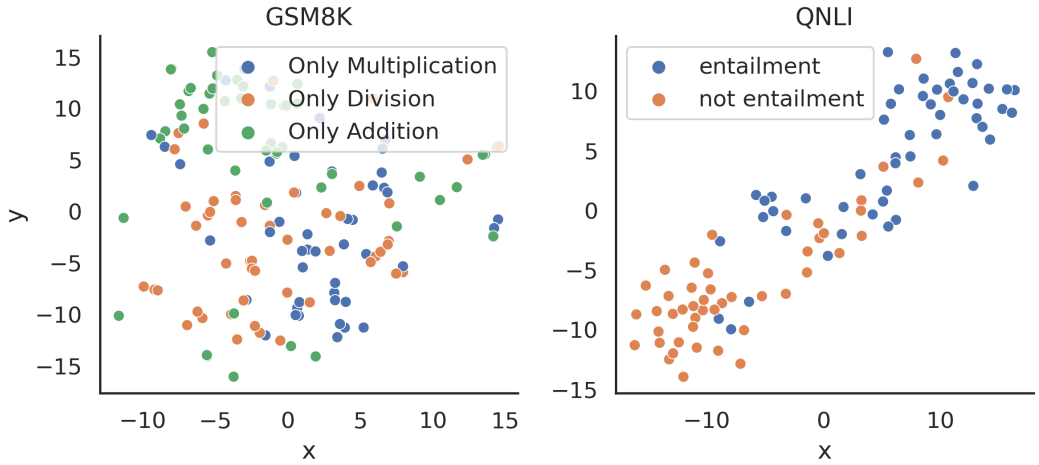


Figure 10: t-SNE Visualizations of gist embeddings for additional datasets. For QNLI, gist embeddings encode class labels. For GSM8K, they encode whether the solution can be obtained by a chain-of-thought reasoning comprising only addition, only multiplication, or only division.

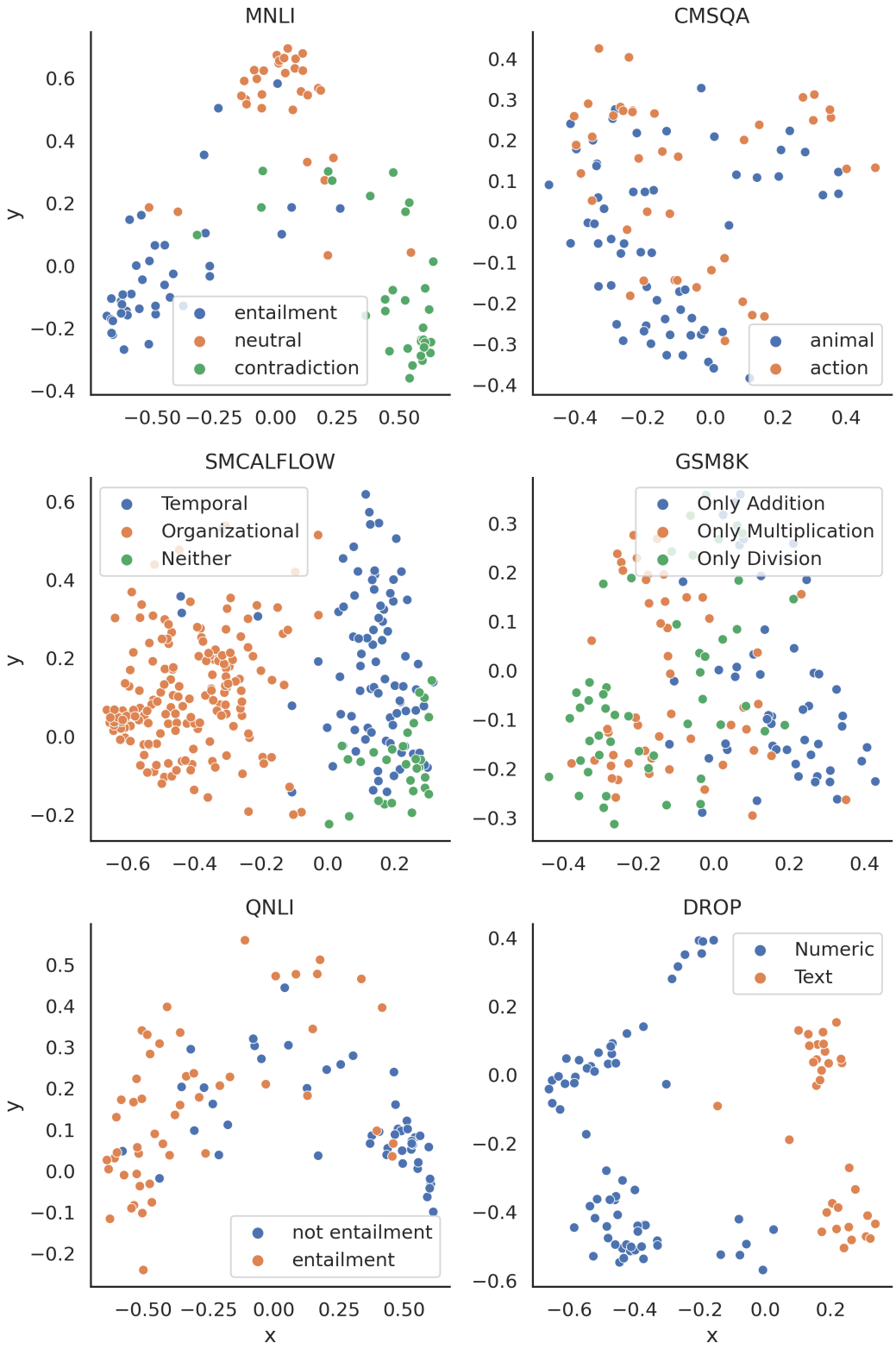


Figure 11: PCA visualizations of gist embeddings show similar results as t-SNE visualization in Figure 5 and 10. Gist embeddings encode task-specific salient information such as class labels (MNLI, QNLI) or more abstract information aspects (CMSQA, SMCalFlow, DROP, GSM8K) that help retrieve better in-context examples.

LM	Dataset	BSR	SET-BSR	GS[M]				SET-GS[M]				GS[F]				SET-GS[F]			
				l=1	l=3	l=6	l=15	l=1	l=3	l=6	l=15	l=1	l=3	l=6	l=15	l=1	l=3	l=6	l=15
GPT-Neo-2.7B	SMC CG	2.7	4.5	2.4	2	2.7	3.2	1.8	5	4.8	5.1	4.2	8	8.4	11.3	6.3	10.6	13	13
	SMC IID	36.6	37	33.4	35.5	34.1	35.3	30.4	36.3	34.7	37	50	54.8	53.8	53.2	38.5	53.9	53.9	54.5
	COGS CG	52.3	53	53.3	51.6	47.9	50	42.9	53.3	48.5	52	56.3	64.8	68.1	64.6	50.8	65	69.3	66.6
	COGS IID	61.4	64.2	55.9	58.9	55.4	55.3	48.5	60.5	62.8	65.7	62.4	70.3	76.7	67.9	52.9	72.3	79.6	72.9
	MTOP	54.1	53.1	52.8	52.9	53.4	53.2	45.9	53.2	53	54.6	60.1	60.9	62.8	60.1	49.1	61.3	61.5	60.8
	AVG	41.4	42.4	39.6	40.2	38.7	39.4	33.9	41.7	40.8	42.9	46.6	51.8	54.0	51.4	39.5	52.6	55.5	53.6
LLaMA-7B	SMC CG	8.9	17.8	5.7	6.6	8	7.7	9	16.7	16.9	16.4	8.7	15.8	16.1	17.6	11.6	24	24.6	24.9
	SMC IID	51.7	53	46.8	47.7	47.7	48.9	46.4	50.6	52.9	52.4	59.8	65.4	66.8	64.4	55.7	65.6	68.1	65.9
	COGS CG	59.3	59.6	57.1	57	52.1	54.7	53.6	59.2	55.9	56.5	63.9	70	72.6	70.5	64.5	73.2	73.7	73.1
	COGS IID	70.7	76	69.7	68.2	66.8	64.6	65.6	70	72.8	74	75.1	81	87.8	80.1	75.6	83.6	87.3	84.5
	MTOP	60	60.2	58.4	59.5	57.9	61.3	54.3	59.9	60	60.1	64.7	67.3	68.5	66.8	57.5	67	67.8	68.3
	AVG	50.1	53.3	47.5	47.8	46.5	47.4	45.8	51.3	51.7	51.9	54.4	59.9	62.4	59.9	53.0	62.7	64.3	63.3
Mistral	SMC CG	17.6	49.3	13.4	13.7	15.2	17.5	23.5	35.1	45.4	46.8	17.6	27.3	27.3	32.4	28.7	45.6	49.8	56.7
	SMC IID	62.4	69.8	57.6	59.7	61.9	61.3	57.9	63.7	65.3	68.7	71.5	74.8	74.6	71.5	65	73.7	77.2	75.1
	COGS CG	65.9	66.8	64.3	62.7	61.6	63.3	59.2	65.8	64.7	68	71.7	79.1	80.7	77.6	71.6	80	81.8	81.4
	COGS IID	80.4	82	79	76.9	74.3	75.3	70.7	78.8	83	84.8	81.8	86.5	90.7	86.5	82.1	88.2	92.5	90.5
	MTOP	67.7	69.2	66.9	66.6	66.7	68.3	63.1	69.9	68.2	69.1	71.4	70.3	72.9	72.9	65.6	72.5	73.8	73.5
	AVG	58.8	67.4	56.2	55.9	55.9	57.1	54.9	62.7	65.3	67.5	62.8	67.6	69.2	68.2	62.6	72.0	75.0	75.4
StarCoder	SMC CG	18.6	51.4	16	16.1	17.8	18.9	22.6	35.4	44.6	52.3	14.6	24.9	23.4	30.2	27.3	39.1	43.7	53.1
	SMC IID	65.3	69.6	58.2	60.6	59.1	63.1	55.3	63.4	65.7	69.2	69	71.6	73.3	70.7	64.5	73.4	74.8	73.7
	COGS CG	78	77.1	70.8	72.4	71.9	70.8	64	73.2	73.4	71.6	75	78.4	83.1	80.4	75.8	77.4	82.8	81.4
	COGS IID	91.8	92.4	88.4	88.8	86.3	87.5	81.7	88.9	92.6	91.7	89	91.8	95.6	92.6	89.9	91.3	95.6	94.7
	MTOP	68	70	68.5	69.2	68.6	69.1	65.2	69.8	68.7	71.7	71	71.5	74.1	73.4	65.9	72.1	74.5	75.5
	AVG	64.3	72.1	60.4	61.4	60.7	61.9	57.8	66.1	69.0	71.3	63.7	67.6	69.9	69.5	64.7	70.7	74.3	75.7

Table 9: 8-shot ICL results for varying number of gist tokens (l) and set-selection for semantic parsing datasets with different LLMs.

Dataset	RAND	SBERT	BM25	BSR	GS[M]				GS[F]		EPR	CEIL
					l=1	l=3	l=6	l=15	l=1	l=3		
SMCalFlow (CG)	0	2.6	1.1	2.7	2.4	2	2.7	3.2	4.2	8	3.6	3.8
SMCalFlow (IID)	3.3	30.7	31.6	36.6	33.4	35.5	34.1	35.3	50	54.8	54.5	59.1
MTOP	1.3	48.4	46.4	54.1	52.8	52.9	53.4	53.2	60.1	60.9	62.2	60.5
COGS (CG)	3.8	25.3	26	52.3	53.3	51.6	47.9	50	56.3	64.8		
COGS (IID)	8.1	30.1	34.7	61.4	55.9	58.9	55.4	55.3	62.4	70.3		
QNLI	54.8	56.8	56.3	82.6	86.8	85.9	85.5	85.8	91.4	93	74.9	84.2
MNLI	41.9	42.2	44	76.7	78.1	76.6	78.5	74.6	82	81.4	66.1	71.7
RTE	53.4	50.9	54.2	67.9	83	77.6	81.2	73.3	81.6	81.2		
WANLI	38.8	44.4	42.6	60	58.2	53.8	53	54.8	66.2	65.4		
XNLI (de)	33.9	36.6	33.6	41.8	58.5	56.2	58	56	62	62.6		
XNLI (ru)	32.9	34	36.8	35.6	47.1	46.5	44.3	45.7	51.3	52.5		
MedNLI	41.4	54.2	56.9	70.6	69.4	71.1	69.5	70.4	82.9	83		
SST2	86.9	82.6	81.9	90.9	92.1	92.4	92.5	89.6	93.9	94.3		
SST5	13	38.9	37.9	45.1	48.4	49.3	45.9	45.3	50	52.6	42.8	47
Rotten Tomatoes	83.1	78.1	77.2	84.5	88.9	87	88.2	85.3	90.5	90.3		
MRPC	51	57.6	52.5	70.1	83.1	88	84.1	75	87.3	85.3	76	80.2
QQP	65.9	71.3	75	86.4	85.6	85.2	85.7	84.8	86.7	88.6		
PAWS	48	55.2	52.5	75	90.1	90.2	88.1	84.7	92.7	91.6		
PAWSX (es)	47.5	54.5	52.9	72.1	77.1	79.2	80.7	76	88.4	86.6		
PAWSX (fr)	48	51.5	55.3	70.6	82.4	86.1	83	81	90.4	90.2		
CMSQA	19	17.5	18.1	20.1	54.3	55.6	55	44.5	59.9	57.2	36.8	37.2
AGNews	76.6	89.4	89.3	89.9	91.4	90.4	90.5	90.7	92.1	92.5		
GSM8K	1.7	4	2	2.4	3.4	1.8	3.5	3.6	3.1	3.5		
DROP	7.7	12.5	12.6	10.7	18.5	18.8	19.7	18	25.4	28.7		
BoolQ	39.3	49.6	47.3	50.4	65.2	65	66.3	59.7	69.5	66.3		
CoLA	60.3	64.4	64.9	69.7	76.4	75.9	74.4	70.4	80	80.3		
TweetEval (emotion)	42.5	44.7	48.9	51.9	66	69.8	64.7	59.6	70.3	70.9		
TweetEval (offensive)	58.8	66.5	69.1	65.9	77	73.9	72.6	75.1	76.4	77.2		
AVG (Held-out)	31.67	42.97	43.79	54.29	58.74	58.89	57.68	57.21	65.1	66.96		
AVG (All)	37.96	46.23	46.49	57.07	63.53	63.47	62.8	60.75	68.11	69.07		

Table 10: 8-shot ICL with GPT-Neo-2.7B with independent ranking-based selection. l is the number of gist tokens. Red highlights datasets or tasks that are held-out from our multi-task training collection. AVG (All) and AVG (Held-out) are average performances on all and only held-out datasets, respectively.

Dataset	RAND	SBERT	BM25	BSR	GS[M]				GS[F]		LLM-R
					l=1	l=3	l=6	l=15	l=1	l=3	
SMCalFlow (CG)	0	10	6.5	8.9	5.7	6.6	8	7.7	8.7	15.8	
SMCalFlow (IID)	6.8	45.3	45.2	51.7	46.8	47.7	47.7	48.9	59.8	65.4	
MTOP	3.4	54.3	53	60	58.4	59.5	57.9	61.3	64.7	67.3	
COGS (CG)	13.3	29.3	32.9	59.3	57.1	57	52.1	54.7	63.9	70	
COGS (IID)	10.6	35.9	42.1	70.7	69.7	68.2	66.8	64.6	75.1	81	
QNLI	51.5	57.4	56.8	75.3	80.1	82.2	81.7	79.1	87.7	90.2	69.4
MNLI	54.3	56.1	58	76.3	78.5	76	77.4	76.2	80.8	80.1	69.8
RTE	70	68.2	67.9	70.8	85.6	80.1	81.6	78.7	84.5	84.8	70.4
WANLI	45.8	47.1	46.6	55.8	56.7	55.2	53.3	52.4	62.5	63.1	
XNLI (de)	40.6	37.9	35.2	43.2	54.6	54.7	53.8	52.2	59.2	61.5	
XNLI (ru)	36.5	39.7	35	36.7	48.3	43.1	44	45.3	49.7	52.2	
MedNLI	60.4	69.2	68.1	74.8	73.9	75	74.6	75.3	82.8	83.6	
SST2	94.2	93.2	92	95.8	95.2	94.6	94.6	94.2	94.6	94.7	93.1
SST5	38.4	45.2	43.2	40.7	45.9	44.8	45.1	45.6	46.8	51.2	
Rotten Tomatoes	93.1	91.3	92.2	92	92.8	91.3	91.5	92.2	92.3	91.8	
MRPC	33.8	48.3	46.6	59.8	77.9	80.6	78.2	67.9	82.4	77.5	78.2
QQP	66.2	73.2	76.1	80.4	82	80.1	79.7	80.2	83.7	84.1	83.3
PAWS	59.1	57.2	56.6	74	86.3	88.1	87.2	80.6	90.7	89.3	57
PAWSX (es)	57.8	59.4	58.9	69.9	73.2	76.2	75.6	72.3	84.5	81.4	
PAWSX (fr)	56.8	59.6	59.7	69.2	76.9	79.3	78.9	74.2	86.2	87.4	
CMSQA	39.9	26.2	29.9	30.3	60.1	63.4	62.1	49.2	63.7	60	
AGNews	85.7	88.2	86.8	88.9	90.4	90.4	90.1	88.2	90.7	92.4	93.5
GSM8K	11	12.4	12.3	14.3	15.6	14	14.2	13.3	12.6	14.1	
DROP	24.4	28.5	27.6	27.4	32.7	32.2	31.9	31.4	36.5	39.2	
BoolQ	71.2	75.5	73.4	77.6	81.8	80.4	81.1	77.5	82.8	82.4	74.1
CoLA	60.1	67	70.3	70.3	74.4	71.9	73.8	72.4	77.4	77.5	
TweetEval (emotion)	42.8	55.6	60.2	61	70.3	72.2	68.4	65.8	79.4	76.7	
TweetEval (offensive)	67.6	68.7	71.6	68.2	76.2	75	74.8	74.7	77.3	77	
AVG (Held-out)	38.25	50.24	50.51	58.67	61.47	61.5	60.53	60.11	67.58	69.59	
AVG (All)	46.26	53.57	53.74	60.83	65.97	65.71	65.22	63.43	70.04	71.13	

Table 11: 8-shot ICL with LLaMA-7B with independent ranking-based selection. l is the number of gist tokens. Red highlights datasets or tasks that are held-out from our multi-task training collection. AVG (All) and AVG (Held-out) are average performances on all and only held-out datasets, respectively.

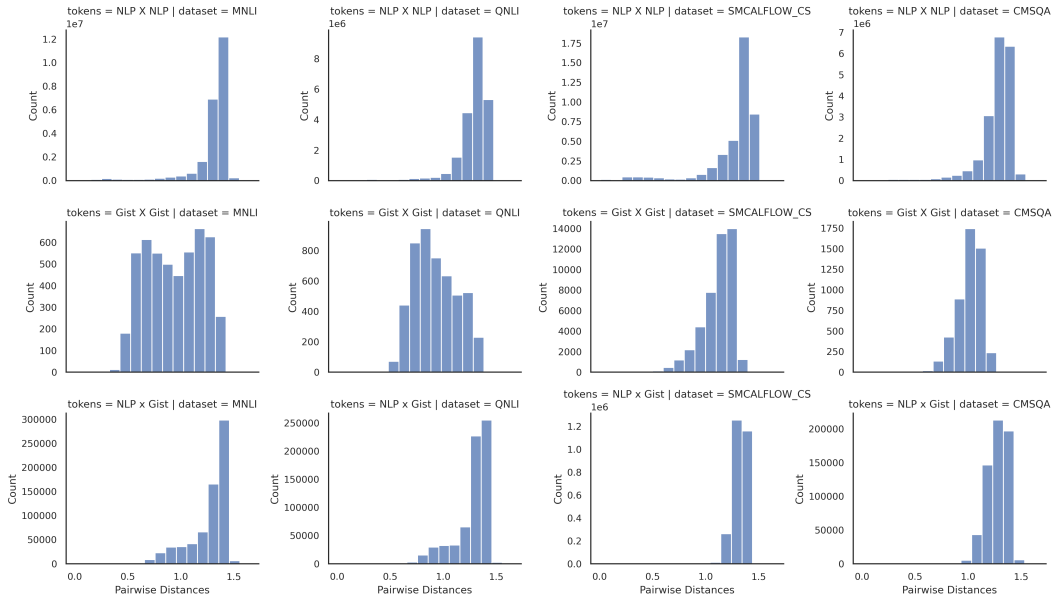


Figure 12: Pairwise Distances between Gist and NLP token activations.

Dataset	RAND	SBERT	BM25	BSR	GS[M]				GS[F]	
					l=1	l=3	l=6	l=15	l=1	l=3
SMCalFlow (CG)	0	12.4	9.5	12.7	10.1	8.7	8.7	11.5	10.6	19.9
SMCalFlow (IID)	15.3	48.8	49.4	57.4	50.3	55	52.4	53.5	60.7	62.8
MTOP	3.9	59.7	56.5	63.4	61.4	62.4	61.2	65	68.8	68.7
COGS (CG)	14.8	31.7	35.5	60.2	56.8	57.1	53	56.3	66.5	70.5
COGS (IID)	16.3	41.3	48.5	71.7	69.8	71.1	68.6	68.9	76.9	82.2
QNLI	56.7	59.7	59.5	80.6	86.2	86.1	85.4	85.8	91.2	92.6
MNLI	50.3	61.9	62.1	82	80.6	80.4	80.6	78.4	83.4	81.4
RTE	76.5	73.3	77.6	75.5	86.3	82.7	83	80.5	85.6	84.5
WANLI	44	50	50.3	60.2	59.1	58.1	56.4	59.5	67.9	67.1
XNLI (de)	36.1	40.6	36.5	44.1	55.5	57.8	56.6	53.6	57.6	59.9
XNLI (ru)	34.5	38	37.3	36.2	47.2	44.9	46.6	48.1	48.9	53.9
MedNLI	54.5	71.9	73.4	77.7	77.6	78	78.4	77.9	83.2	84.6
SST2	93.5	93	92.4	94.8	94.8	94.6	94.4	93.3	94.3	94.7
SST5	40	46.2	46.7	42	44.6	43.5	46.8	43.2	46.5	48
Rotten Tomatoes	87.1	91.6	91.8	92.2	91.6	92.1	91.8	92.9	91.7	91.5
MRPC	70.6	62.7	57.8	71.6	86.8	88	85.5	77.2	87	86
QQP	66.8	77.2	79	85.1	84.4	83.4	84.2	84.2	86.2	87.4
PAWS	59.7	58.5	58.8	77.1	89.4	90.2	89.3	85.3	92.5	91.7
PAWSX (es)	60.2	60.5	59.9	73.9	75.9	78.4	77.8	75.1	85.3	83.1
PAWSX (fr)	63.4	63.5	61.6	74.2	80.4	84.6	82.4	79.6	89	90
CMSQA	51.4	41	44	42.2	64.7	68.4	67.4	60.4	64.9	62.2
AGNews	83.9	91.6	91.2	91.3	92.9	92.8	92.7	91.2	93.4	93.9
GSM8K	15.4	16.4	16.7	19.4	16.8	18.2	18.1	18.6	18.9	17.3
DROP	31.1	33.5	32.9	33.2	37.3	36.7	38.4	36.7	42.7	42.9
BoolQ	63.4	77	75.5	78.7	83.4	82.7	82.6	80.3	83	82.7
CoLA	58.9	65.4	71	72.4	76	74.5	76.8	72.9	80.1	79.5
TweetEval (emotion)	55.3	67.9	70.3	69.8	71.1	73	74.6	74.1	77.5	78.6
TweetEval (offensive)	66.7	69.9	71.1	69.6	77.6	76	75.7	75.5	78.3	78.3
AVG (Held-out)	39.44	53.41	53.69	61.66	63.17	64.09	63.16	63.68	68.78	70.79
AVG (All)	48.94	57.33	57.74	64.61	68.16	68.55	68.19	67.13	71.88	72.71

Table 12: 8-shot ICL with LLaMA-13B with independent ranking-based selection. l is the number of gist tokens. Red highlights datasets or tasks that are held-out from our multi-task training collection. AVG (All) and AVG (Held-out) are average performances on all and only held-out datasets, respectively.

Dataset	RAND	BM25	SBERT	BSR	GS[M, LARGE]				GS[F]		GS[M, XL]
					$l=1$	$l=3$	$l=6$	$l=15$	$l=1$	$l=3$	$l=1$
SMCalFlow (CG)	0	21.6	15.7	17.6	13.4	13.7	15.2	17.5	17.6	27.3	18.1
SMCalFlow (IID)	13.7	55.7	57.1	62.4	57.6	59.7	61.9	61.3	71.5	74.8	64
MTOP	7	63.5	60.2	67.7	66.9	66.6	66.7	68.3	71.4	70.3	67.1
COGS (CG)	14.2	35.2	42.7	65.9	64.3	62.7	61.6	63.3	71.7	79.1	60.5
COGS (IID)	18.4	48	58.7	80.4	79	76.9	74.3	75.3	81.8	86.5	75.9
QNLI	56.4	62.8	61.2	83.3	85.4	86.4	86.9	85.8	90.6	92.3	87.9
MNLI	62	67.6	67.9	85.6	84.5	82.2	82.2	82.5	85	85.7	85.7
RTE	80.1	77.3	75.1	79.4	88.8	84.5	83.4	83.8	87.7	84.8	88.4
WANLI	54.5	56.3	56.6	65.1	65.3	60.1	63	61.8	71.4	71.3	65.7
XNLI (de)	35.1	46.3	42.9	52	68	66.9	68.1	63.8	70.2	70.9	71.1
XNLI (ru)	33.4	42.8	42.9	44.6	57.1	55.5	55.1	54.3	59.7	58.3	60.4
MedNLI	75.4	78.7	77.6	84.2	80.7	82	83.3	82.5	83.1	85	83.5
SST2	95.5	94.5	94.4	96.4	94.7	94.7	96	95	95.9	95.6	94.8
SST5	51.1	51.1	51.8	50.5	52.9	53.2	52.7	52.7	54.2	55.4	53.6
Rotten Tomatoes	93.3	91.9	92.9	92.7	93.2	92.5	92.5	93.5	90.7	91.8	92.6
MRPC	72.8	70.6	67.6	76.7	85.5	88	84.6	79.7	87	87	90.4
QQP	73.8	78.5	80.5	86.1	84.8	84.4	84.3	85.5	86.9	88.5	85.1
PAWS	71.2	60.8	63.7	74.1	90.5	91.3	90.4	88.1	93.5	92.5	92.5
PAWSX (es)	68.8	63.3	63.9	76.9	80.7	82.2	82.2	77.8	88.8	87.2	86.3
PAWSX (fr)	71.7	63.8	65.6	74.6	83.9	86.4	84.1	82.5	90.8	90.7	86.8
CMSQA	73.5	67.6	70.6	69	75.1	76.4	76.5	72.7	74.2	73.3	77.8
AGNews	88.3	93.4	93.2	93.1	94.6	94.4	93.7	92.9	93.8	94.4	94.5
GSM8K	34.8	37.3	37	40	37.9	37.6	39.4	38.7	38.5	40.3	42.2
DROP	41.1	48.3	48.2	48.4	56	54.8	53.9	54.8	58.5	59.2	56.2
BoolQ	86.4	87.3	86.9	88.8	87.7	88.9	87.2	87.9	86	86.5	89.1
CoLA	82.1	82.2	83.1	82.2	81.8	80.3	81.1	82	83	83.2	83
TweetEval (emotion)	59.1	75.4	77.3	78.1	75.7	78.3	78.6	77.5	80.7	82.9	78.9
TweetEval (offensive)	65.7	69.3	72.2	69.3	77.4	75.1	75.4	74.3	76.5	76.9	78.8
AVG (Held-out)	43.59	57.99	59.02	66.54	68.8	68.47	68.71	68.12	73.28	75.21	70.69
AVG (All)	56.41	63.97	64.55	70.9	73.69	73.42	73.37	72.71	76.45	77.56	75.39

Table 13: 8-shot ICL with Mistral with independent ranking-based selection. l is the number of gist tokens. Red highlights datasets or tasks that are held-out from our multi-task training collection. AVG (All) and AVG (Held-out) are average performances on all and only held-out datasets, respectively.

Dataset	RAND	SBERT	BM25	BSR	GS[M]				GS[F]	
					l=1	l=3	l=6	l=15	l=1	l=3
SMCalFlow (CG)	0	19	13.4	15.8	11.8	12.1	13.3	15.1	16.1	23.7
SMCalFlow (IID)	5.9	51.1	50.8	56.6	51.1	53.6	57.6	59.7	66.8	69.3
MTOP	4.7	59	54	61.3	61	61.1	59.8	62.3	67	65.9
COGS (CG)	15.4	33.8	39.7	63.3	61.4	59.6	59.3	61.8	68.5	76.1
COGS (IID)	17.7	46.6	55.4	77.4	74.7	72	72.4	70.9	78	83
QNLI	81.7	81.3	81.9	85.3	89	87.8	88.4	88.8	91.6	92.3
MNLI	73.4	72.5	72.1	84.3	84.5	83.3	83.7	83.7	85.2	84.5
RTE	80.5	81.6	81.2	82.7	87.4	83.4	85.2	85.6	86.3	85.2
WANLI	50.5	58.8	59.5	65.5	64.3	62.1	63.4	63.4	69.8	69.3
XNLI (de)	42.5	45.9	46.3	52	64.2	64.6	64.1	61.5	70.8	69.3
XNLI (ru)	42.8	44.7	44.6	43.1	57.8	55.4	53.1	53.5	57.5	58.9
MedNLI	76.3	80	80.8	83.6	82	83.9	83.8	82.3	84.4	85.3
SST2	95.6	94.8	95.1	96	95.6	96.1	96.1	96.1	95.9	96.1
SST5	52.3	51.6	51.2	51.4	53.2	52.8	52.7	53.9	56.1	55.2
Rotten Tomatoes	92.5	91.1	91.8	92.8	93.4	93.3	92.9	93.3	91.3	92.4
MRPC	74.3	67.9	63.2	73	79.4	83.3	80.6	74.3	82.1	82.4
QQP	80.2	80	82	82	81.7	82.3	81.5	83.5	85.1	84.6
PAWS	71.7	68.5	70.7	77.9	87.9	85.8	85.7	84.7	90.2	88.9
PAWSX (es)	73.5	69.1	68.8	76.6	79.3	81.4	81.7	77.7	86.2	86
PAWSX (fr)	72.9	69.9	72.9	78.2	82.6	82.9	81.8	80.9	87.7	87.4
CMSQA	72.5	67.7	71.6	68.9	71.8	74.1	72.9	71.5	73	72.2
AGNews	87.8	93.3	92.6	93.1	93.8	93.5	93.9	92.3	92.6	93.5
GSM8K	37.9	38.1	35.9	42	38.3	38.9	38.7	39.2	39	37.5
DROP	37	47	46.3	46.5	52.3	53.8	53.2	53.6	53.6	54.6
BoolQ	86.5	87	86	87.7	86.5	86.9	87.4	87.2	87	88
CoLA	80.2	79.4	81.6	80.8	80.1	80.5	80.4	80.7	83.7	83.1
TweetEval (emotion)	71.7	72.5	74.1	75.7	71.9	77.3	75.1	76.5	76.7	78.1
TweetEval (offensive)	68.2	70.5	71.7	68.3	74.7	73	72.2	73.1	75	76.3
AVG (Held-out)	45.33	58	58.84	65.01	66.44	66.59	66.46	66.57	71.13	72.93
AVG (All)	58.79	65.1	65.54	70.06	71.85	71.96	71.82	71.68	74.9	75.68

Table 14: 8-shot ICL with Zephyr with independent ranking-based selection. l is the number of gist tokens. Red highlights datasets or tasks that are held-out from our multi-task training collection. AVG (All) and AVG (Held-out) are average performances on all and only held-out datasets, respectively.

Dataset	RAND	BM25	SBERT	BSR	GS[M] $l=1$	GS[F] $l=1$
SMCalFlow (CG)	0	0.3	0.6	0.8	0.9	1.2
SMCalFlow (IID)	2.9	15.4	17.2	25.2	17.7	34.3
MTOP	2.3	45.9	46.2	52.3	50.6	58.6
COGS (CG)	2.1	10.3	13.2	29.6	29.9	31.4
COGS (IID)	2.2	13.7	17.4	35.4	31.1	32.7
QNLI	51.7	55.6	56.8	83	86.5	91.2
MNLI	35.4	43.5	46.8	83.2	80.4	85.1
RTE	59.9	56.3	57.4	74	84.1	83
WANLI	38.7	45	47.9	62.6	60.2	68.4
XNLI (de)	34	36.5	36.8	51.6	65.9	68.4
XNLI (ru)	32.9	38.6	39.9	39.9	52.4	55.4
MedNLI	36.7	53.4	59.3	74.3	72.4	83.2
SST2	90.7	88.2	87.6	94.8	92.1	94.6
SST5	31.4	36.8	38.7	44.4	48.6	49.4
Rotten Tomatoes	76.8	84.6	87.2	91	90.8	90.5
MRPC	68.4	68.9	65.9	75	85	87.7
QQP	56.6	56.4	64.9	83.8	82.8	87.3
PAWS	44.5	48.8	50.4	68.6	89.8	93.4
PAWSX (es)	51.9	47.2	45.7	66.5	79.3	88.7
PAWSX (fr)	50.6	50.6	50.9	65.9	83.3	91
CMSQA	20.9	20	19.6	20.4	55.5	63.4
AGNews	85.7	92.3	92.5	93.4	92.9	93.3
GSM8K	2.7	4.1	3.6	5	2.8	4.6
DROP	10.9	14.5	15.1	14	24.3	30.1
BoolQ	64.3	68	67.8	70.3	82.8	82.8
CoLA	68.6	64.3	67	69	76.6	79.3
TweetEval (emotion)	42.5	48.1	58.6	64.7	73.5	79.1
TweetEval (offensive)	52.5	64.7	70.4	65.8	78	76.1
AVG (Held-out)	30.44	39.59	42.24	51.83	56.14	61.36
AVG (All)	39.92	45.43	47.34	57.3	63.22	67.29

Table 15: 8-shot ICL with Babbage with independent ranking-based selection. l is the number of gist tokens. Red highlights datasets or tasks that are held-out from our multi-task training collection. AVG (All) and AVG (Held-out) are average performances on all and only held-out datasets, respectively.

Dataset	RAND	BM25	SBERT	BSR	GS[M] $l=1$	GS[F] $l=1$
SMCalFlow (CG)	0	0.8	1.6	2.4	3.2	1.2
SMCalFlow (IID)	0.8	12.4	17.2	29.6	22	22.8
MTOP	2.4	55.6	52.4	56.8	59.2	61.2
COGS (CG)	10	21.2	21.6	46.4	44.8	48.4
COGS (IID)	6.4	22.4	24.8	44	44.4	40
QNLI	45.2	57.2	52	82	84.4	92.4
MNLI	55.6	62.8	60	84.8	82.4	83.2
RTE	77.2	71.6	71.6	80	88.4	85.2
WANLI	49.2	50.8	52.8	65.2	62.4	71.6
XNLI (de)	42.8	46.8	44.4	52.4	73.2	69.6
XNLI (ru)	41.6	45.6	43.6	43.2	59.6	60.8
MedNLI	61.6	75.6	72.8	83.2	78.4	84
SST2	94.8	88.4	89.2	95.6	94	94
SST5	45.2	50.8	52	47.6	51.2	54.8
Rotten Tomatoes	93.2	91.2	94.8	94	94	94
MRPC	71.6	68.8	62.4	78	85.2	89.2
QQP	70.4	76.8	78.8	85.6	83.2	86
PAWS	67.6	55.6	60	80.8	90.4	94.4
PAWSX (es)	64.4	59.2	55.2	70.4	79.6	84.4
PAWSX (fr)	65.6	59.6	65.6	67.6	82.8	88.8
CMSQA	72.8	65.6	67.2	66.8	77.6	75.2
AGNews	86	94.8	93.6	92	93.6	92.8
GSM8K	32.8	30	33.6	37.2	36.8	35.2
DROP	36	38	42.8	37.6	49.6	49.6
BoolQ	82.8	84	88	88	91.6	88
CoLA	73.2	74.8	78.8	77.6	77.2	75.6
TweetEval (emotion)	58	62.8	69.2	64.8	66.8	79.6
TweetEval (offensive)	68.8	69.2	70.4	71.6	78.5	78.1
AVG (Held-out)	40.34	48.09	49.03	56.54	60.64	63.18
AVG (All)	52.71	56.87	57.73	65.19	69.09	70.72

Table 16: 8-shot ICL with Davinci with independent ranking-based selection. l is the number of gist tokens. Red highlights datasets or tasks that are held-out from our multi-task training collection. AVG (All) and AVG (Held-out) are average performances on all and only held-out datasets, respectively.

Dataset	SBERT	BM25	BSR	GS[M]				GS[F]	
				l=1	l=3	l=6	l=15	l=1	l=3
SMCalFlow (CG)	346.84	8.72	2411.3	22.86	52.32	48.71	49.82	10.53	11.22
SMCalFlow (IID)	341.16	8.52	2418.1	23.36	56.11	26.02	27.86	13.27	12.88
MTOP	169.04	5.99	723.89	23.49	55.38	54.92	59.26	10.17	10.81
COGS (CG)	305.8	8.62	818.34	30.41	58	58.08	61.93	10.71	10.25
COGS (IID)	297.53	8.92	816.26	23.61	56.33	51.97	62.43	11.04	11.06
QNLI	1416.9	18.21	10934	1.49	2.25	3.7	7.02	1.54	2.02
MNLI	1469.9	20.71	9565.4	1.47	2.54	3.34	6.87	1.58	2.04
RTE	68.81	1.01	696.73	0.79	0.8	0.69	0.92	0.57	0.83
WANLI	1351	19.45	6556.2	1.45	2.17	3.81	6.89	1.53	1.98
XNLI (de)	6271	51.67	28794	118.75	61.97	67.06	54.06	31.17	35.73
XNLI (ru)	5863.8	56.89	35382	125.5	56.74	56.95	62.26	35.84	30.65
MedNLI	285.6	4.78	3357.4	0.74	49	39.19	44.8	25.45	22
SST2	1119.4	22.36	3639.3	1.52	2.31	3.36	6.99	1.53	2.14
SST5	295.95	5.01	609.36	0.64	0.93	1.12	1.68	0.75	1.2
Rotten Tomatoes	755.88	11.7	963.7	63.07	35.3	35.01	32.83	11.5	11.32
MRPC	89.62	1.34	255.69	0.66	0.67	0.76	1.03	0.52	0.68
QQP	1336.2	20.13	8862.5	1.55	2.14	3.58	6.86	1.58	2.06
PAWS	1350.4	20.94	6712.2	1.72	2.24	3.46	6.92	1.6	2.02
PAWSX (es)	5266.5	52.59	11698	118.93	60.19	60.61	53.79	24.65	24.95
PAWSX (fr)	5367.5	52.03	11118	114.39	59.72	56.77	53.77	24.66	25.23
CMSQA	290.7	4.19	1124.5	0.63	0.92	1.15	1.86	0.68	0.91
AGNews	2098.1	20.78	11813	1.56	2.46	3.36	6.86	1.51	1.96
GSM8K	138.6	3.19	1605.9	0.61	0.84	1	1.71	0.63	0.9
DROP	5068.6	29.71	22340	1.51	2.23	3.24	6.75	1.48	2.01
BoolQ	413.46	3.75	4876	0.63	0.9	1.11	1.81	0.68	0.9
CoLA	109.66	3.82	644.41	0.64	0.87	1.03	1.74	0.72	0.88
TweetEval (emotion)	378.42	6.57	1032	104.86	40.61	45.11	38.31	11.4	11.82
TweetEval (irony)	146.37	6.22	1886	90.75	34.71	46.66	39.15	11.05	11.41
TweetEval (offensive)	1201.3	20.23	4154.5	98.87	44.79	49.41	41.29	11.75	11.42
TweetEval (sentiment)	4964.1	71.18	6870.6	124.91	61.1	60.22	62.68	24.54	25.12

Table 17: Time (in ms) to select 8-shots for the various datasets using the different training-free methods. The time for SBERT is higher than gisting-based retrieval because our implementation for it does not use FAISS indexing.