# GistScore: Learning Better Representations for In-Context Example Selection with Gist Bottlenecks

**Shivanshu Gupta** [1 †] **Clemens Rosenbaum** [†] **Ethan R. Elenberg** [2 †]


## 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 through supervised finetuning with an attention bottleneck between the inputs and outputs. These *gist models* form the basis for *GistScore*, a novel metric for scoring and selecting informative examples. Further, we experiment with two variations: (1) finetuning gist models for each dataset and (2) multi-task training a single model on a large collection of datasets. The latter can be used for new tasks out-of-the-box, enabling a training-free ICL pipeline. Evaluations with 21 datasets spanning 9 tasks and 8 diverse LLMs show that our fine-tuned models get 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 well to new tasks, datasets, and prompt templates. Selection using this model matches or outperforms prior methods while being three orders of magnitude faster than the strongest training-free baseline.[1]


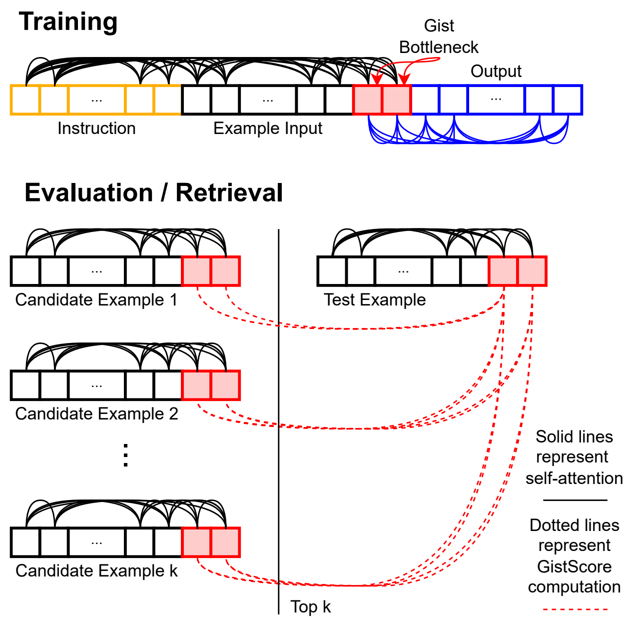


*Figure 1.* **Top** Example Gisting involves supervised training with an attention masking bottleneck. Here, gist tokens (red) may attend to example inputs (black) and the task instruction (yellow, optional), however, the output (blue) may only attend to the gist tokens. Training with such a bottleneck encourages concise, task-specifc encodings of salient aspects of inputs. Further, with multi-task training the model can be applied to new tasks out-of-the-box. **Bottom** Retrieval of the candidate examples with the highest GistScore with the test input (Task instruction omitted for brevity).

## 1. Introduction

In-Context Learning (ICL) (Brown et al., 2020) is a few-shot inference paradigm that leverages increasingly powerful large language models (LLMs) for new tasks by conditioning them on a prompt comprising a few task demonstrations. In contrast to traditional supervised finetuning, the training-free approach allows a single model to instantly switch between an arbitrary number of tasks with improved generalization (Anil et al., 2022; Qiu et al., 2022; Drozdov et al., 2023; Wei et al., 2022b) and reasoning skills (Wei et al., 2022b). Unfortunately, its performance is highly sensitive to the choice of examples placed in the prompt (Zhao et al., 2021; Liu et al., 2022b; Lu et al., 2022; Rubin et al., 2022; Schick & Schütze, 2021).

Despite extensive prior work on better example selection

[†]Work done at ASAPP [1]University of California, Irvine, CA, USA [2]Permanence AI, New York, NY, USA. Correspondence to: Shivanshu Gupta <shivag5@uci.edu>.

*Proceedings of the 41st International Conference on Machine Learning*, Vienna, Austria. PMLR 235, 2024. 

[1]Code and models available at https://github.com/Shivanshu-Gupta/gist-icl.

methods (Rubin et al., 2022; Ye et al., 2023a; Mualem et al., 2024; Gupta et al., 2023), the predominant approach in practice 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 training on the target task and/or training with feedback from a much larger Inference LLM (Rubin et al., 2022; Ye et al., 2023a; 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's quadratic complexity in input length makes it computationally expensive for long-text tasks. Further, its reliance on general-purpose encoders may lead to sub-optimal selection for many tasks.

To address these limitations, we seek to train computationally efficient retrievers that select informative in-context examples and can be used out-of-the-box for new tasks and datasets. We propose *Example Gisting*, a novel approach for training encoders for in-context example retrieval 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 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, while finetuning on each dataset can yield optimal performance, to enable a training-free ICL pipeline we also experiment with multi-task training a single gist model on a large collection of datasets. By gisting task instructions along with the example input, such a model can be used to select in-context examples for new tasks and datasets out-of-the-box.

Evaluating on 21 diverse datasets spanning 9 task categories and 8 diverse LLMs, we find that example selection using GistScore dramatically improves ICL. With finetuning, it consistently outperforms all prior selection methods, including ones that leverage task or LLM-specific training. In particular, it beats 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 can match prior trained methods and beat prior training-free methods even on held-out datasets and tasks while being thousands of times faster than BERTScore. Finally, congruent to Gupta et al. (2023), we find that the set-extension of GistScore is highly effective for the task of Semantic Parsing and compositional generalization. Our analysis shows that gist token embeddings capture abstract, task-specific salient aspects and can be effective for selection even for tasks that the gist model itself fails. Overall, our multi-task trained gist model presents the best tradeoff of performance, ease of use, and selection speed and can potentially replace the standard approach of using off-the-shelf retrievers.

## 2. Related Work

**In-Context Learning** Given the appeal of In-Context Learning as a training-free way to leverage LLMs, various example selection strategies to alleviate its sensitivity to choice of examples have been proposed: (1) selecting diverse examples to reduce redundancy among them (Su et al., 2023; Levy et al., 2023; Agrawal et al., 2023; Ye et al., 2023b), (2) selecting examples that minimize the entropy of the LLM's output distribution for the test input (Lu et al., 2022; Wu et al., 2023), (3) Bayesian inference (Wang et al., 2023), (4) influence functions (Nguyen & Wong, 2023), and (5) selecting examples as a set (Gupta et al., 2023; Ye et al., 2023a; Mualem et al., 2024).

Perhaps the most relevant to our work are Rubin et al. (2022) and Wang et al. (2024) which propose different ways to train example retrievers using feedback from a much larger LLM. However, task-specific finetuning requires abundant training data while also sacrificing the training-free nature of ICL limiting ease of use. Further, methods trained with an LLM-in-the-loop can lose effectiveness with larger Inference LLMs (Gupta et al., 2023). Orthogonally, Gupta et al. (2023) showed that simply using BERTScore-Recall (BSR) (Zhang et al., 2020) to score examples yields a training-free method that selects informative examples that demonstrate the *salient aspects* of the test input. However, BSR requires matching every pair of token embeddings in the candidate and the test input making it computationally expensive for long-text tasks. Moreover, the general-purpose encoders it uses may not capture informativeness for every task.

**Attention and Memory** Example Gisting is inspired from Mu et al. (2023)'s Gisting[2] which was used to compress prompts. Both 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. (2024) model

[2]We will refer to this method as Instruction Gisting to distinguish it from our proposed methods.

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; they model long dependencies with either overlapping windows or specific tokens with overlapping attention.

## 3. Preliminaries

**In-Context Learning (ICL)** LLMs have the ability to solve test inputs from new tasks when prompted with a few examples of that task. Formally, 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 to conditionally generate the following test output:

$$\mathbf{y}_{\text{test}} \sim \mathcal{P}_{LM}\left(\cdot \mid \mathcal{T}\left(\mathbf{x}_1, \mathbf{y}_1, \ldots, \mathbf{x}_k, \mathbf{y}_k, \mathbf{x}_{\text{test}}\right)\right) \quad (1)$$

**Example Selection** This work focuses on the problem of selecting the $k$ in-context examples from a pool of $N \gg k$ labeled candidates. This is often necessary due to the limited context windows of LLMs. Moreover, even if it were possible to fit the entire pool in the prompt, LLMs have been shown to be highly sensitive to both the order (Liu et al., 2022b) and the position of in-context examples (Liu et al., 2024). Thus, we seek to select the most relevant subset of candidates to improve both the computational efficiency and performance of ICL. Formally, the goal is to select a subset $\mathcal{S} \subset \{(x_i, y_i)\}_{i=1}^N$ of size $k$ that maximizes the probability of generating the desired $\mathbf{y}_{\text{test}}$ when the Inference LLM is conditioned on $\mathbf{x}_{\text{test}}$ and $\mathcal{S}$.

Beyond naïve random selection, the standard approach for this problem is to retrieve the top-$k$ examples from the candidate pool using either the BM25 algorithm (Robertson et al., 1993; Jones et al., 2000) or dense retrieval using an off-the-shelf encoder. However, such general-purpose retrievers are not trained for selecting examples for in-context learning and can yield sub-optimal performance. Moreover, standard approaches for training retrievers (Karpukhin et al., 2020) are not applicable as the gold retrieval is unknown. As described in § 2, prior approaches mitigate this by training with feedback from an Inference LLM (Rubin et al., 2022; Wang et al., 2024) or by using a more suitable general-purpose metric (Gupta et al., 2023).

**Instruction Gisting** Mu et al. (2023) proposed Instruction Gisting for compressing instruction-following prompts into shorter *gists* for efficient LLM inference. To perform this mapping, they train a gisting model, $GM$, 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 initial LM and 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 ensures that the model must predict based on the information of $t$ encoded in the activations above $G$. Denoting this gist of $t$ as $G(t)$, this approach of instruction tuning with a gist bottleneck can also be seen as distillation between a standard instruction-tuned LM and the gisting model $GM$:

$$\begin{aligned}&\mathcal{L}_G\left(p_G, \mathcal{D}_I\right) = \\ &\mathop{\mathbb{E}}_{t,x,y \sim \mathcal{D}_I}\left[\mathrm{KL}\left(p_{\mathrm{LM}}(y \mid t, x) \parallel p_{GM}(y \mid G(t), x)\right)\right].\end{aligned} \quad (2)$$

The trained gisting model can be used for new instructions by feeding it the sequence $[t, G]$, precomputing the activations above $G$, and then prompting it with those activations instead of $t$.

## 4. Method

### 4.1. Intuition

The intuition behind our proposed approach follows from the work of Gupta et al. (2023) who showed that ICL requires examples that share the *salient aspects* of the test input i.e. task-specific features of the input such as reasoning patterns, rules, or similar properties that dictate its mapping to the output. Such examples are likely to be informative about how to solve the task. Selecting such examples requires a relevance metric that can measure the candidate examples' coverage of the test input's salient aspects. However, BSR (Zhang et al., 2020), the metric that Gupta et al. (2023) used, could only capture salient aspects explicitly expressed in text. In this work, we seek encoders that can extract all the task-specific salient information from the input into an encoding that can be used to select informative examples. Inspired by the Instruction Gisting approach of Mu et al. (2023), we posit that training a model to perform a task with a attention-masking bottleneck between inputs and outputs would enable the bottleneck to encode all the task-relevant information of the inputs.

### 4.2. Example Gisting

We now describe *Example Gisting*, our approach to training example encoders for ICL example selection. Consider an initial LM and a labeled dataset for target task $t$: $\mathcal{D}_t = \{(x_i, y_i)\}$. Analogous to Instruction Gisting, we finetune a model $GM$ to predict $y_i$ given the inputs $[x_i, G]$, where $G$ is the attention bottleneck comprising $l$ gist tokens. As in Eq. (2), this is akin to minimizing the following distillation objective:

$$\mathcal{L}_G\left(p_G, \mathcal{D}_t\right) = \underset{x,y\sim\mathcal{D}_t}{\mathbb{E}}\left[\mathrm{KL}\left(p_{\mathrm{LM}}(y \mid x) \| p_{GM}(y \mid G(x))\right)\right]. \tag{3}$$

As motivated in § 4.1, Example Gisting trains the model to encode task-specific salient information of the inputs in the activations of the gist tokens. Next section will describe how gist activations can be used to select in-context examples. However, note that, 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. Thus, Example Gisting is agnostic to the choice of Inference LLM and also does not suffer from the failure cases of Instruction Gisting, such as difficulty copying verbatim from the instruction (Mu et al., 2023).

### 4.3. Example Selection

A trained example gisting model can be used to select examples by mapping the candidates and the test input to sequences of *gist embeddings* that can then be used to score the candidates. Specifically, 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 as gist embeddings, *i.e.* $\mathbf{z} = \mathbf{z}_1, \dots \mathbf{z}_l = G(z)[-1]$ and $\mathbf{x} = \mathbf{x}_1, \dots \mathbf{x}_l = G(x_{\text{test}})[-1]$. Then we use the following metric, which we call *GistScore*, to measure the relevance of each candidate with respect to the test input:

$$\mathrm{GS}(x, z) = \frac{1}{l}\sum_{i=1}^{l} \max_{j=1,\dots l} \frac{\mathbf{x}_i^T \mathbf{z}_j}{\|\mathbf{x}_i\| \|\mathbf{z}_j\|} \tag{4}$$

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, as described in App A, GistScore admits Gupta et al. (2023)'s extension to a submodular set-level metric that can be greedily optimized to select examples together as a set. This is particularly useful in settings that require Compositional Generalization

### 4.4. Multi-Task Training

While task-specific finetuning with the approach described in § 4.2 can yield greater performance, it shares the ease-of-use limitations prior to trained methods described in § 3. To address we propose a multi-task training approach that enables the gisting model to be used out-of-the-box on new tasks without any additional training. This preserves the key advantage of ICL: the entire pipeline may be used with new tasks, domains, and prompt templates without any training. 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 an initial LM and 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 $G$ is the attention bottleneck as before. This is equivalent to minimizing the following multi-task distillation objective:

$$\mathcal{L}_G\left(p_G, \mathcal{D}_M\right) = \underset{t,x,y\sim\mathcal{D}_M}{\mathbb{E}}\left[\mathrm{KL}\left(p_{\mathrm{LM}}(y \mid t, x) \| p_{GM}(y \mid G(t, x))\right)\right]. \tag{5}$$

## 5. Experimental Setup

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

*Table 1.* Datasets used in this work. Red highlights datasets held-out from our multi-task collection. We use the German and Russian splits of XNLI, Spanish and French of PAWSX, and IID and Compositional Generalization (CG) splits of SMCalFlow and COGS.

### 5.1. Datasets

**Multi-task Corpus** For multi-task training gist models as described in § 4.4, we use a subset of the FLAN 2022 collection (Longpre et al., 2023) which comprises 15M zero and few-shot prompts from over 473 datasets and 146 task categories. Specifically, we subsample up to 10,000 zero-shot prompts at most 256 tokens long for every task category, yielding roughly 5M prompts.

**ICL Evaluation** We evaluate on 21 datasets spanning 9 diverse task categories and multiple languages as listed in Table 1. 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 Tweet-Eval (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.

We also evaluate on Semantic Parsing, a task that requires set-selection (Gupta et al., 2023) and that is completely absent in our multi-task collection, making it a hard test of generalization for our multi-task models. Further, in addition to IID splits as for other datasets, for SMCalFlow (Andreas et al., 2020) and COGS (Kim & Linzen, 2020), we also evaluate on compositional generalization (CG) splits. We include additional details about all the datasets, including splits, sample instances, selection and ICL templates, and metrics in App. B.

## 5.2. 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.

## 5.3. Methods

### 5.3.1. GistScore

As described in § 4, GistScore-based example selection involves training a gist model to produce example gists used within GistScore (GS) or its set-extension (Set-GS) for selection. While decoder-only LMs can also be used for gisting (Mu et al., 2023), we use encoder-decoder LMs. After example gisting training, we drop the decoder and retain only the encoder for gisting examples. As described in § 4.2 and § 4.4, we experiment with both finetuning gist models for each dataset as well as multi-task training a single model on the collection described in § 5.1 and then directly using it to gist and select in-context examples for downstream datasets. We refer to these models as GistScore-based selection using these as GS[Finetune] or **GS[F]** and GS[Multitask] or **GS[M]** and the set-extension as **Set-GS[F]** and **Set-GS[M]**, respectively. For GS[F], we use `flan-t5-base` (Chung et al., 2022) as the base LM and for GS[M], we use `flan-t5-large`. Each model is trained to produce gists of a fixed length $l$ denoted as GS[F, $l$] and GS[M, $l$]. We report results with $l = 1$ unless specified otherwise. Additional training details are provided in App. C.

### 5.3.2. Baselines

In addition to randomly selecting in-context examples (**Rand**), we compare with the following training-free ranking-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`[7] 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., 2023a) 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., 2024) 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. (2023a), 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. (2024)'s held-out tasks are included in our multi-task collection, we only compare with it on its held-in datasets.

## 5.4. 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.

[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.

[4] https://hf.co/mistralai/Mistral-7B-v0.1

[5] https://hf.co/HuggingFaceH4/zephyr-7b-alpha

[6] https://hf.co/bigcode/starcoder

[7] https://github.com/dorianbrown/rank_bm25

| 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 2.* Average 8-shot ICL performance across all datasets with single-token GistScore and training-free baselines for different LLMs. See App. D for complete results for each dataset and LLM. While finetuning (GS[F]) yields the best performance, GS[M] also outperforms the baselines and recovers much of GS[F]'s performance despite requiring no finetuning.

| Selector | SMC | | COGS | | MTOP | AVG |
|---|---|---|---|---|---|---|
| | IID | CG | IID | CG | | |
| 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 3.* 8-shot ICL using StarCoder for Semantic Parsing datasets with independent ranking (top) and set-selection (bottom) methods. 15-token SET-GS[F] outperforms all baselines while SET-GS[M] matches them despite never being trained for Semantic Parsing.

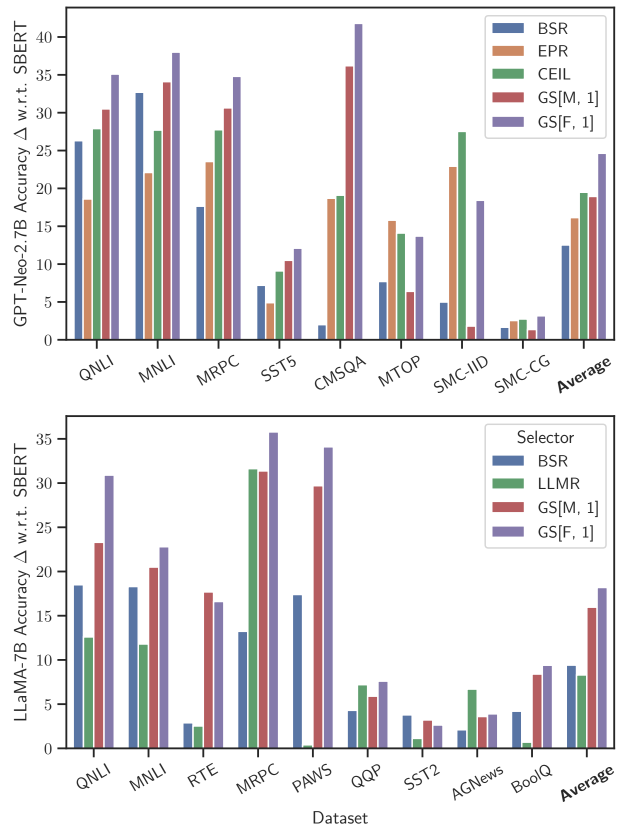


*Figure 2.* Single-token GistScore v/s BSR and trained baselines: EPR and CEIL with GPT-Neo-2.7B **(Top)** and LLM-R with LLaMA-7B **(Bottom)**. All numbers are absolute gain in 8-shot ICL performance over SBERT except EPR and CEIL on MNLI, SST5, MRPC, and CMSQA which are with 50 in-context examples. Both GS[F] and GS[M] consistently outperform all baselines, with GS[F] performing the best. Semantic parsing is an exception as it requires additional gist tokens and set-selection (see Table 3).

(2) ordering the examples by increasing relevance so that the more relevant examples are closer to the test input, (3) converting the ordered examples and the test input to text using the dataset's ICL example template in Tables 5, 6, and 7, and (4) concatenating the templated examples. For set-selection methods (SET-BSR and SET-GS), the examples are ordered by their corresponding instance-level score.

## 6. Results

**Finetuned GistScore is the superior method in-context example selection method.** Table 2 and Figure 2 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. Additional results for all datasets and LLMs are provided in App. D. 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. Finally, the gains from GistScore persist across varying number of in-context examples (Figure 4, Left) – with just 2 examples, it outperforms 8-shots retrieved using general-purpose retrievers.

**Semantic Parsing benefits from additional gist-tokens and set-selection.** While a single gist token works best for most datasets, it can be insufficient to capture all the salient information in complex compositional semantic parsing instances. Moreover, as shown in (Gupta et al., 2023), their compositional nature also necessitates set-selection as opposed to independent ranking-based selection, which can yield redundant examples while omitting information. Indeed, as shown in Figure 4 (Right), 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 3).

**Multi-task training yields strong performance out-of-the-box.** Table 2 and Figure 2 show that without task-specific

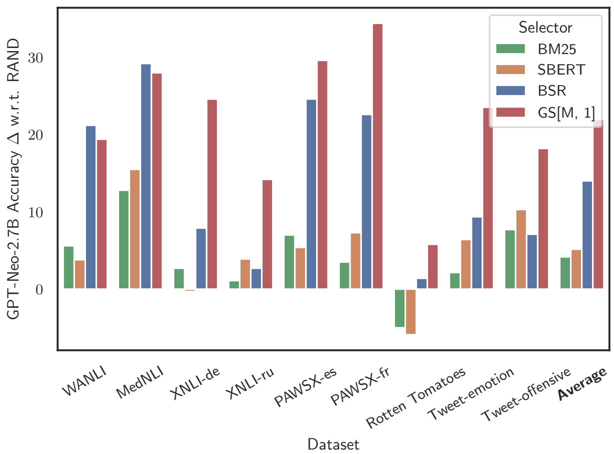


*Figure 3.* Comparison of training-free methods on held-out datasets. GS[M] is able to generalize out-of-the-box to held-out datasets, domains (e.g., tweet, medical), and languages, significantly outperforming both off-the-shelf retrievers as well as the stronger but slower BSR.

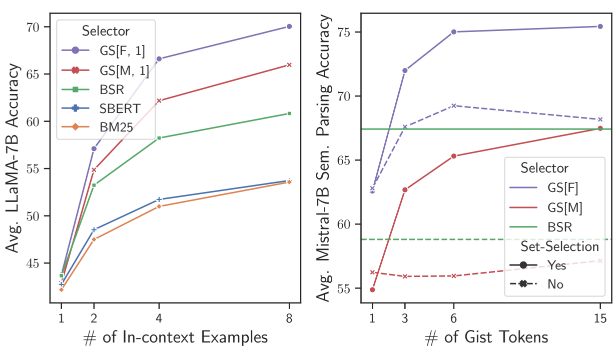


*Figure 4.* **Left** GS[F] and GS[M] consistently outperform baselines across varying number of in-context examples, requiring just 2 examples to surpass 8-shot ICL using SBERT and BM25. **Right** Due to their complex compositional nature, Semantic Parsing datasets benefit from additional gist tokens 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 3 for trained baselines and Table 11 for complete results.

finetuning, example selection using our multi-task trained gist model (GS[M]) is able to recover much of the performance of its finetuned counterparts (GS[F]) and matches or outperforms all baselines including trained ones like EPR, CEIL, and LLM-R. Further analyzing its held-out performance in Figure 3, we see that GS[M] is able to generalize out-of-the-box to held-out datasets, domains (e.g., tweet, medical), and languages, significantly outperforming off-the-shelf retrievers (BM25 and SBERT) as well as the

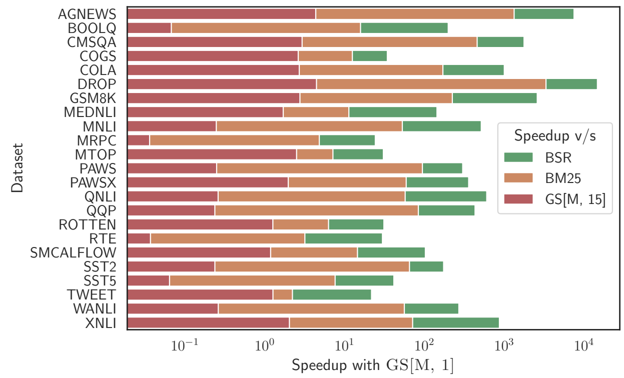


*Figure 5.* Example selection using GistScore (GS[M, 1]) is up to four (three) orders of magnitude faster than BSR (BM25), and scales well with the number of gist tokens.

stronger BSR. For semantic parsing datasets, unlike GS[F], just using more gist tokens without set-selection was only marginally effective (Figure 4, Right), likely because semantic parsing is not included in the multitask corpus, and so the gist model is unable to leverage the additional tokens well. However, when used for set-selection, additional tokens dramatically improve performance — with 15 tokens, despite not being trained for semantic parsing, SET-GS[M] is able to match even trained methods like EPR and CEIL (see Table 3). These results confirm that multi-task training for gisting both task instructions and example inputs enables generalization to new tasks making it a promising approach for an improved training-free ICL pipeline.

**Selection using GistScore is significantly faster than BSR.** 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 5 shows that this yields thousands of times faster selection with single-token GistScore compared to BSR, which took over 20 seconds per test input for some datasets (see Table 12). Further, due to GPU acceleration, we found GistScore to be significantly faster than even BM25.

## 7. Analysis

Having seen that GistScore-based example selection can effectively improve ICL performance, we now provide some analyses to better understand this improvement (see also App. E for additional analyses of gist embeddings, examples selected by different selectors, etc.).

**Gist embeddings encode task-specific salient aspects.** In § 4, we hypothesized that example-gisting training would enable gist models to encode task-specific salient information in gist embeddings. t-SNE visualizations of gist embed-

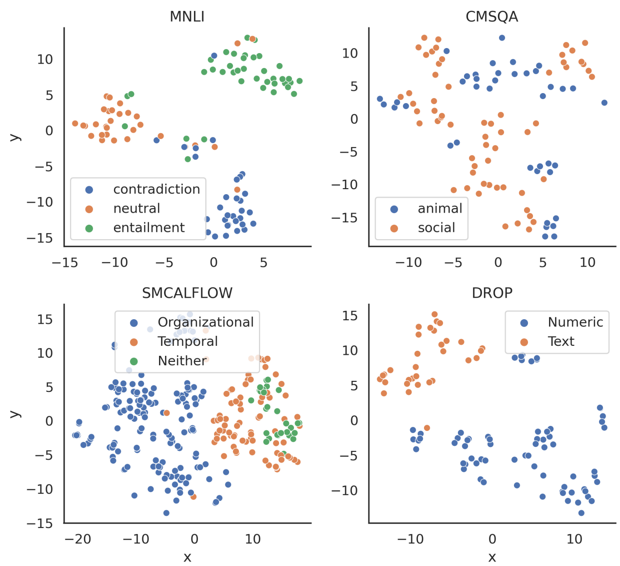


*Figure 6.* t-SNE visualizations of gist embeddings show that they encode task-specific salient information useful for retrieving informative in-context examples. For MNLI, a classification task, gist embeddings contain information about the class labels. For CMSQA, they encode relevant concepts in the question, *i.e.* whether it's about an animal or an action (*e.g.* "driving car," etc.). For SMCalFlow, they encode whether the input pertains to organizational hierarchy (*e.g.* `Who is Bill's manager?`), contains temporal information (*e.g* `Book me a dentist appointment before 3pm today`), or neither (*e.g.* `I need a meeting with Steve`). For DROP, they contain information whether the question expects a numeric or textual answer.

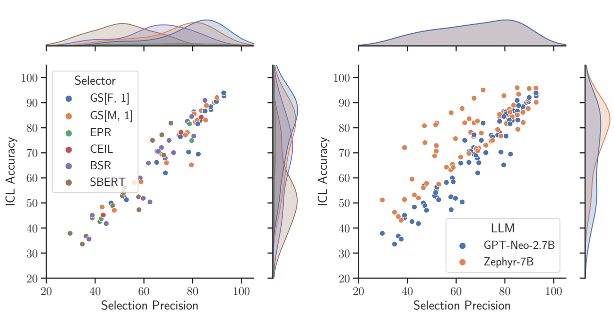


*Figure 7.* **Left** ICL accuracy (using GPT-Neo-2.7B) across all classification tasks is strongly correlated with the precision of the various selectors, *i.e.* per-dataset-average of the fraction of in-context examples with the same label as the test input. This suggests that retrieving such examples is the primary driver of ICL performance for these datasets. **Right** However, stronger LLMs like Zephyr can improve ICL performance beyond selection precision, especially when the selector is inaccurate.

| | Method | SST5 | QNLI | CMSQA | SMC CG | SMC IID | COGS CG | COGS IID | GSM | DROP |
|---|---|---|---|---|---|---|---|---|---|---|
| | 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 4.* Comparison of ICL performance with 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. In fact, on compositional splits and GSM8K which requires chain-of-thought reasoning (Wei et al., 2022b), GistScore improves ICL performance even though the gist model itself fails.

dings for various tasks (Figure 6) show that gist embeddings indeed encode features that would be considered salient for those tasks (*e.g.*, class label for MNLI, relevant concepts for CMSQA, type of question for DROP, etc.). This enables examples selected using GistScore to be more semantically relevant and informative about the test input, as shown qualitatively in Tables 9, 8, and 10.

**GistScore can improve ICL performance beyond that of the underlying gist model.** As gist models, along with the corresponding decoder, are trained to perform the task itself, their task performance could limit the downstream ICL performance. Table 4 shows that this is not necessarily the case—ICL using GistScore can yield performance exceeding that of the underlying gist model itself, especially when using stronger LLMs. This is best exemplified by tasks requiring compositional generalization and chain-of-thought (COT) reasoning, that are known to hard for smaller finetuned models (Qiu et al., 2022; Wei et al., 2022a). This is because, as shown in Figures 6 and 12 for SMCalFlow, DROP, and GSM8K, even in settigns where the gist models fails, its gists can encode abstract task-specific salient aspects useful for selecting informative examples.

**Classification performance is tied to the precision of the example selector.** For classification tasks, we found ICL accuracy using the different selection methods to be strongly correlated with the precision of selected examples' labels (Figure 7, Left). Similar trends were observed with majority-vote accuracies of the selectors as well (see App. E). This previously unshown phenomenon suggests that for classification, the various example selectors improve ICL performance not necessarily by selecting informative examples but solving the task and biasing the LLM's prediction towards the correct label. Note that this could be an artifact of the finite output space, where the most informative examples tend to have the same label as the test input. It also does not necessarily limit ICL performance for classification tasks—as shown in Figures 7 (Right) and 11, stronger LLMs are less reliant on accurate retrieval and can improve ICL performance beyond selection precision, especially when the selector is inaccurate. Nevertheless, whether it is possible to improve ICL performance without merely surfacing the correct label and biasing the LLM remains an important open question.

## 8. Conclusion

This work presents Example Gisting, a novel approach for training retrievers for in-context learning through supervised finetuning of encoder-decoder models with a bottleneck that forces encoding the salient information in inputs into a few tokens. We additionally propose GistScore, a novel metric to compare the gist encodings of candidates with the test input. Evaluation with a wide range of tasks and LLMs validates the efficacy of our approach by demonstrating the superior performance of our fine-tuned gist models. Finally, the out-of-the-box generalization of our multi-task trained models enables an improved yet training-free in-context learning pipeline. Future work could study the efficacy of gisting in other settings that require retrieval, such as retrieval augmented generation.

## Impact Statement

This paper presents a novel approach for retrieving examples for in-context learning with LLMs. While not a specific consequence of our approach, with LLMs, there is always a risk of generating biased, toxic, or non-factual outputs. Further, with in-context learning, the quality and factuality of the retrieved in-context examples also play a role. In turn, these would depend on the bias, toxicity, or factuality of the pool from which the examples are retrieved as well as the retrieval approach. In particular, for our approach, while we don't expect example gisting training to exacerbate these aspects, a gist model could retain any biases of the initial base model as well as the training data used.

## References


Agrawal, S., Zhou, C., Lewis, M., Zettlemoyer, L., and Ghazvininejad, M. In-context examples selection for machine translation. In Rogers, A., Boyd-Graber, J., and Okazaki, N. (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 8857–8873, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.564. URL https://aclanthology.org/2023.findings-acl.564.

Andreas, J., Bufe, J., Burkett, D., Chen, C., Clausman, J., Crawford, J., Crim, K., DeLoach, J., Dorner, L., Eisner, J., Fang, H., Guo, A., Hall, D., Hayes, K., Hill, K., Ho, D., Iwaszuk, W., Jha, S., Klein, D., Krishnamurthy, J., Lanman, T., Liang, P., Lin, C. H., Lintsbakh, I., McGovern, A., Nisnevich, A., Pauls, A., Petters, D., Read, B., Roth, D., Roy, S., Rusak, J., Short, B., Slomin, D., Snyder, B., Striplin, S., Su, Y., Tellman, Z., Thomson, S., Vorobev, A., Witoszko, I., Wolfe, J., Wray, A., Zhang, Y., and Zotov, A. 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.

Anil, C., Wu, Y., Andreassen, A. J., Lewkowycz, A., Misra, V., Ramasesh, V. V., Slone, A., Gur-Ari, G., Dyer, E., and Neyshabur, B. Exploring length generalization in large language models. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022.

Barbieri, F., Camacho-Collados, J., Espinosa Anke, L., and Neves, L. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In Cohn, T., He, Y., and Liu, Y. (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.

Beltagy, I., Peters, M. E., and Cohan, A. Longformer: The Long-Document transformer, 2020.

Bentivogli, L., Magnini, B., Dagan, I., Dang, H. T., and Giampiccolo, D. 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.

Black, S., Gao, L., Wang, P., Leahy, C., and Biderman, S. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow, March 2021. URL https://doi.org/10.5281/zenodo.5297715.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems*

*2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

Child, R., Gray, S., Radford, A., and Sutskever, I. Generating long sequences with sparse transformers, 2019.

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S. S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Castro-Ros, A., Pellat, M., Robinson, K., Valter, D., Narang, S., Mishra, G., Yu, A., Zhao, V., Huang, Y., Dai, A., Yu, H., Petrov, S., Chi, E. H., Dean, J., Devlin, J., Roberts, A., Zhou, D., Le, Q. V., and Wei, J. Scaling instruction-finetuned language models, 2022.

Clark, C., Lee, K., Chang, M.-W., Kwiatkowski, T., Collins, M., and Toutanova, K. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In Burstein, J., Doran, C., and Solorio, T. (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.

Conneau, A., Rinott, R., Lample, G., Williams, A., Bowman, S., Schwenk, H., and Stoyanov, V. XNLI: Evaluating cross-lingual sentence representations. In Riloff, E., Chiang, D., Hockenmaier, J., and Tsujii, J. (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.

Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q., and Salakhutdinov, R. 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. URL https://aclanthology.org/P19-1285.

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

Drozdov, A., Schärli, N., Akyürek, E., Scales, N., Song, X., Chen, X., Bousquet, O., and Zhou, D. Compositional semantic parsing with large language models. In *The Eleventh International Conference on Learning Representations*, 2023.

Dua, D., Wang, Y., Dasigi, P., Stanovsky, G., Singh, S., and Gardner, M. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In Burstein, J., Doran, C., and Solorio, T. (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.

Guo, M., Ainslie, J., Uthus, D., Ontanon, S., Ni, J., Sung, Y.-H., and Yang, Y. 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. URL https://aclanthology.org/2022.findings-naacl.55.

Gupta, S., Gardner, M., and Singh, S. Coverage-based example selection for in-context learning. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 13924–13950, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.930. URL https://aclanthology.org/2023.findings-emnlp.930.

Herlihy, C. and Rudinger, R. MedNLI is not immune: Natural language inference artifacts in the clinical domain. In Zong, C., Xia, F., Li, W., and Navigli, R. (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.

Hu, Y., Lee, C.-H., Xie, T., Yu, T., Smith, N. A., and Ostendorf, M. In-context learning for few-shot dialogue state tracking. In Goldberg, Y., Kozareva, Z., and Zhang, Y. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 2627–2643, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.193. URL https://aclanthology.org/2022.findings-emnlp.193.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. Mistral 7b, 2023.

Jones, K. S., Walker, S., and Robertson, S. E. A probabilistic model of information retrieval: development and comparative experiments - part 2. *Inf. Process. Manag.*, 36:809–840, 2000.

Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., and Yih, W.-t. Dense passage retrieval for open-domain question answering. In Webber, B., Cohn, T., He, Y., and Liu, Y. (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6769–6781, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.550. URL https://aclanthology.org/2020.emnlp-main.550.

Kim, N. and Linzen, T. COGS: A compositional generalization challenge based on semantic interpretation. In Webber, B., Cohn, T., He, Y., and Liu, Y. (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.

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

Levy, I., Bogin, B., and Berant, J. Diverse demonstrations improve in-context compositional generalization. In Rogers, A., Boyd-Graber, J., and Okazaki, N. (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1401–1422, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.78. URL https://aclanthology.org/2023.acl-long.78.

Li, H., Arora, A., Chen, S., Gupta, A., Gupta, S., and Mehdad, Y. MTOP: A comprehensive multilingual task-oriented semantic parsing benchmark. In Merlo, P., Tiedemann, J., and Tsarfaty, R. (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.

Li, R., allal, L. B., Zi, Y., Muennighoff, N., Kocetkov, D., Mou, C., Marone, M., Akiki, C., LI, J., Chim, J., Liu, Q., Zheltonozhskii, E., Zhuo, T. Y., Wang, T., Dehaene, O., Lamy-Poirier, J., Monteiro, J., Gontier, N., Yee, M.-H., Umapathi, L. K., Zhu, J., Lipkin, B., Oblokulov, M., Wang, Z., Murthy, R., Stillerman, J. T., Patel, S. S., Abulkhanov, D., Zocca, M., Dey, M., Zhang, Z., Bhattacharyya, U., Yu, W., Luccioni, S., Villegas, P., Zhdanov, F., Lee, T., Timor, N., Ding, J., Schlesinger, C. S., Schoelkopf, H., Ebert, J., Dao, T., Mishra, M., Gu, A., Anderson, C. J., Dolan-Gavitt, B., Contractor, D., Reddy, S., Fried, D., Bahdanau, D., Jernite, Y., Ferrandis, C. M., Hughes, S., Wolf, T., Guha, A., Werra, L. V., and de Vries, H. Starcoder: may the source be with you! *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=KoFOg41haE. Reproducibility Certification.

Lin, C.-Y. 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.

Liu, A., Swayamdipta, S., Smith, N. A., and Choi, Y. WANLI: Worker and AI collaboration for natural language inference dataset creation. In Goldberg, Y., Kozareva, Z., and Zhang, Y. (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.

Liu, J., Shen, D., Zhang, Y., Dolan, B., Carin, L., and Chen, W. What makes good in-context examples for GPT-3? In Agirre, E., Apidianaki, M., and Vulić, I. (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.

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., and Liang, P. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 02 2024. ISSN 2307-387X. doi: 10.1162/tacl_a_00638. URL https://doi.org/10.1162/tacl_a_00638.

Longpre, S., Hou, L., Vu, T., Webson, A., Chung, H. W., Tay, Y., Zhou, D., Le, Q. V., Zoph, B., Wei, J., and Roberts, A. The flan collection: Designing data and methods for effective instruction tuning. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 22631–22648. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/longpre23a.html.

Lu, Y., Bartolo, M., Moore, A., Riedel, S., and Stenetorp, P. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In Muresan, S., Nakov, P., and Villavicencio, A. (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.

Meron, J. Simplifying semantic annotations of SMCalFlow. In Bunt, H. (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.

Mu, J., Li, X. L., and Goodman, N. Learning to compress prompts with gist tokens. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Mualem, L., Elenberg, E. R., Feldman, M., and Karbasi, A. Submodular minimax optimization: Finding effective sets. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, 2024.

Nguyen, T. and Wong, E. In-context example selection with influences, 2023.

Pang, B. and Lee, L. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In Knight, K., Ng, H. T., and Oflazer, K. (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.

Qiu, L., Shaw, P., Pasupat, P., Shi, T., Herzig, J., Pitler, E., Sha, F., and Toutanova, K. Evaluating the impact of model scale for compositional generalization in semantic parsing. In Goldberg, Y., Kozareva, Z., and Zhang, Y. (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 9157–9179, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.624. URL https://aclanthology.org/2022.emnlp-main.624.

Rae, J. W., Potapenko, A., Jayakumar, S. M., Hillier, C., and Lillicrap, T. P. 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. URL https://openreview.net/forum?id=SylKikSYDH.

Reimers, N. and Gurevych, I. 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. URL https://aclanthology.org/D19-1410.

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

Rubin, O., Herzig, J., and Berant, J. Learning to retrieve prompts for in-context learning. In Carpuat, M., de Marneffe, M.-C., and Meza Ruiz, I. V. (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.

Schick, T. and Schütze, H. Exploiting cloze-questions for few-shot text classification and natural language inference. In Merlo, P., Tiedemann, J., and Tsarfaty, R. (eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 255–269, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.20. URL https://aclanthology.org/2021.eacl-main.20.

Shazeer, N. and Stern, M. Adafactor: Adaptive learning rates with sublinear memory cost. In Dy, J. G. and Krause, A. (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. URL http://proceedings.mlr.press/v80/shazeer18a.html.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., and Potts, C. Recursive deep models for semantic compositionality over a sentiment treebank. In Yarowsky, D., Baldwin, T., Korhonen, A., Livescu, K., and Bethard, S. (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.

Su, H., Kasai, J., Wu, C. H., Shi, W., Wang, T., Xin, J., Zhang, R., Ostendorf, M., Zettlemoyer, L., Smith, N. A., and Yu, T. Selective annotation makes language models better few-shot learners. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/pdf?id=qY1hlv7gwg.

Talmor, A., Herzig, J., Lourie, N., and Berant, J. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In Burstein, J., Doran, C., and Solorio, T. (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.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. Llama: Open and efficient foundation language models, 2023.

Tunstall, L., Beeching, E., Lambert, N., Rajani, N., Rasul, K., Belkada, Y., Huang, S., von Werra, L., Fourrier, C., Habib, N., Sarrazin, N., Sanseviero, O., Rush, A. M., and Wolf, T. Zephyr: Direct distillation of lm alignment, 2023.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In Linzen, T., Chrupała, G., and Alishahi, A. (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.

Wang, L., Yang, N., and Wei, F. Learning to retrieve in-context examples for large language models. In Graham, Y. and Purver, M. (eds.), *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1752–1767, St. Julian's, Malta, March 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.eacl-long.105.

Wang, X., Zhu, W., Saxon, M., Steyvers, M., and Wang, W. Y. Large language models are latent variable models: Explaining and finding good demonstrations for in-context learning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=BGvkwZEGt7.

Warstadt, A., Singh, A., and Bowman, S. R. 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.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., and Fedus, W. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022a. ISSN 2835-8856. Survey Certification.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., brian ichter, Xia, F., Chi, E. H., Le, Q. V., and Zhou, D. Chain of thought prompting elicits reasoning in large language models. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022b. URL https://openreview.net/forum?id=_VjQlMeSB_J.

Williams, A., Nangia, N., and Bowman, S. A broad-coverage challenge corpus for sentence understanding through inference. In Walker, M., Ji, H., and Stent, A. (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.

Wu, Z., Wang, Y., Ye, J., and Kong, L. Self-adaptive in-context learning: An information compression perspective for in-context example selection and ordering. In Rogers, A., Boyd-Graber, J., and Okazaki, N. (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1423–1436, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.79. URL https://aclanthology.org/2023.acl-long.79.

Xiao, G., Tian, Y., Chen, B., Han, S., and Lewis, M. Efficient streaming language models with attention sinks. In

*The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=NG7sS51zVF.

Yang, Y., Zhang, Y., Tar, C., and Baldridge, J. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. In Inui, K., Jiang, J., Ng, V., and Wan, X. (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.

Ye, J., Wu, Z., Feng, J., Yu, T., and Kong, L. Compositional exemplars for in-context learning. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023a.

Ye, X., Iyer, S., Celikyilmaz, A., Stoyanov, V., Durrett, G., and Pasunuru, R. Complementary explanations for effective in-context learning. In Rogers, A., Boyd-Graber, J., and Okazaki, N. (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 4469–4484, Toronto, Canada, July 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.273. URL https://aclanthology.org/2023.findings-acl.273.

Yin, P., Fang, H., Neubig, G., Pauls, A., Platanios, E. A., Su, Y., Thomson, S., and Andreas, J. Compositional generalization for neural semantic parsing via span-level supervised attention. In Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., and Zhou, Y. (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.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. 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. URL https://openreview.net/forum?id=SkeHuCVFDr.

Zhang, X., Zhao, J. J., and LeCun, Y. Character-level convolutional networks for text classification. In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R. (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. URL https://proceedings.neurips.cc/paper/2015/hash/250cf8b51c773f3f8dc8b4be867a9a02-Abstract.html.

Zhang, Y., Baldridge, J., and He, L. PAWS: Paraphrase adversaries from word scrambling. In Burstein, J., Doran, C., and Solorio, T. (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.

Zhao, Z., Wallace, E., Feng, S., Klein, D., and Singh, S. Calibrate before use: Improving few-shot performance of language models. In Meila, M. and Zhang, T. (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. URL http://proceedings.mlr.press/v139/zhao21c.html.

## A. 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 $x_{\text{test}}$ as a recall of salient aspects with the following form,

$$\text{cover}\,(x_{\text{test}}\,, z) = \sum_{s \in S_{x_{\text{test}}}} \text{c}(s, z) \tag{6}$$

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:

$$\text{setcov}\,(x_{\text{test}}\,, Z) = \sum_{s \in S_{x_{\text{test}}}} \max_{z \in Z} \text{c}(s, z) \tag{7}$$

For $l > 1$ GistScore, as defined in Eq. 4, has the form of Eq. 6 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:

$$\text{Set-GS}_{l>1}(x, Z) = \frac{1}{l} \sum_{i=1}^{l} \max_{z \in Z} \max_{j=1,\ldots l} \frac{\mathbf{x}_i^T \mathbf{z}_j}{\|\mathbf{x}_i\| \|\mathbf{z}_j\|} \tag{8}$$

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 $c(s, z) = \frac{\mathbf{x}_1[i] \mathbf{z}_1[i]}{\|\mathbf{x}_1\| \|\mathbf{z}_1\|}$:

$$\text{Set-GS}_{l=1}(x, Z) = \sum_{i=1}^{d} \max_{z \in Z} \frac{\mathbf{x}_1[i] \mathbf{z}_1[i]}{\|\mathbf{x}_1\| \|\mathbf{z}_1\|} \tag{9}$$

## B. ICL Evaluation

**Splits** We experiment with 21 datasets spanning 9 task categories. See Table 1 for a summary of the datasets used in this work. 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.

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. (2023), we use the 8-S split. Additionally, we use the programs with the simplified syntax provided by (Meron, 2022).

Following prior work (Gupta et al., 2023; Rubin et al., 2022; Ye et al., 2023a), for 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.

**Templates** Tables 5, 6, and 7 contain the textual templates we use to linearize the instances for example selection and ICL. The ICL prompt is constructed by concatenating the templatized demonstrations and the test instance using `\n\n` as the separator.

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

## C. Training Details

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 finetune Flan-T5-base Chung et al. (2022) models to produce gists of varying lengths on each individual dataset using the procedure described in § 4.2. 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. We use 95% of this sub-

| 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 .`<br>`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`<br>`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 .`<br>`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.`<br>`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?`<br>`Sentence: As of that day, the new constitution heralding the Second Republic came into force.`<br>`Answer: Yes` |
| MNLI | `Premise: The new rights are nice enough`<br>`Does the above premise entail the hypothesis that "Everyone really likes the newest benefits " (Yes, Maybe, or No)?`<br>`Answer:` | `Premise: The new rights are nice enough`<br>`Hypothesis: Everyone really likes the newest benefits`<br>`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.`<br>`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.`<br>`Hypothesis: Christopher Reeve had an accident.?`<br>`Answer: No` |
| MedNLI | `Premise: No history of blood clots or DVTs, has never had chest pain prior to one week ago.`<br>`Is the hypothesis that "Patient has angina." an entailment, contradiction or neutral with respect to the above premise?`<br>`Answer:` | `Premise: No history of blood clots or DVTs, has never had chest pain prior to one week ago.`<br>`Hypothesis: Patient has angina.`<br>`Answer: Yes` |
| WANLI | `Premise: In the past, I have found that there is no point in making a speech unless you have prepared it.`<br>`Is the hypothesis that "You should prepare a speech." an entailment, contradiction or neutral with respect to the above premise?`<br>`Answer:` | `Premise: In the past, I have found that there is no point in making a speech unless you have prepared it.`<br>`Hypothesis: You should prepare a speech.`<br>`Answer: Yes` |
| XNLI | `Premise: Et il a dit, maman, je suis à la maison .`<br>`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?`<br>`Answer:` | `Premise: Et il a dit, maman, je suis à la maison .`<br>`Hypothesis: Il a appelé sa mère dès que le bus scolaire l'a déposé.`<br>`Answer: No` |
| 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?`<br>`Solution: Janet sells 16 - 3 - 4 = <<16-3-4=9>>9 duck eggs a day.`<br>`She makes 9 * 2 = $<<9*2=18>>18 every day at the farmer's market.`<br>`#### 18` |
| AGNews | `Classify the following news article into one of these categories: World, Sports, Business, Technology.`<br>`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.`<br>`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.`<br>`Category: Business` |

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

| 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 .<br>Does the review above see the movie as terrible, bad, OK, good, or great?<br>Answer: | Review: in his first stab at the form , jacquot takes a slightly anarchic approach that works only sporadically .<br>Sentiment: OK |
| SST2 | Review: it 's a charming and often affecting journey .<br>Is the sentiment of the above review Negative or Positive?<br>Answer: | Review: it 's a charming and often affecting journey .<br>Sentiment: Positive |
| Rotten Tomatoes | Review: compassionately explores the seemingly irreconcilable situation between conservative christian parents and their estranged gay and lesbian children .<br>Is the sentiment of the above review Negative or Positive?<br>Answer: | Review: compassionately explores the seemingly irreconcilable situation between conservative christian parents and their estranged gay and lesbian children .<br>Sentiment: Positive |
| MRPC | Sentence 1: He said the foodservice pie business doesn 't fit the company 's long-term growth strategy .<br>Sentence 2: " The foodservice pie business does not fit our long-term growth strategy .<br>Do the above sentences convey the same meaning? Yes or No.<br>Answer: | Sentence 1: He said the foodservice pie business doesn 't fit the company 's long-term growth strategy .<br>Sentence 2: " The foodservice pie business does not fit our long-term growth strategy .<br>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 .<br>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 .<br>Are these sentences paraphrases of each other? Yes or No.<br>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 .<br>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 .<br>Answer: No |
| QQP | Question 1: Why are African-Americans so beautiful?<br>Question 2: Why are hispanics so beautiful?<br>Are Questions 1 and 2 asking the same thing? Yes or No.<br>Answer: | Question 1: Why are African-Americans so beautiful?<br>Question 2: Why are hispanics so beautiful?<br>Answer: No |
| PAWSX | Sentence 1: El Consejo Shawnee Trail nació de la unión entre el Consejo Four Rivers y el Consejo Audubon.<br>Sentence 2: El Consejo de caminos de los Shawnee se formó por la fusión del Consejo de Four Rivers y el Consejo de Audubon.<br>Are the above sentences paraphrases of each other? Yes or No.<br>Answer: | Sentence 1: El Consejo Shawnee Trail nació de la unión entre el Consejo Four Rivers y el Consejo Audubon.<br>Sentence 2: El Consejo de caminos de los Shawnee se formó por la fusión del Consejo de Four Rivers y el Consejo de Audubon.<br>Answer: Yes |
| CoLA | Is the following sentence grammatical (Yes or No )?<br>The sailors rode the breeze clear of the rocks.<br>Answer: | Sentence: The sailors rode the breeze clear of the rocks.<br>Answer: Yes |
| TweetEval | Classify the emotion in the following tweet as one of anger, joy, optimism, or sadness..<br>Tweet: @user @user Oh, hidden revenge and anger ...I rememberthe time,she rebutted you.<br>Answer: | Tweet: @user @user Oh, hidden revenge and anger ...I rememberthe time,she rebutted you.<br>Answer: A |
| CMSQA | Select one of the choices that best answers the following question:<br>Question: A revolving door is convenient for two direction travel, but it also serves as a security measure at a what?<br>Option A: bank<br>Option B: library<br>Option C: department store<br>Option D: mall<br>Option E: new york<br>Answer: | Question: A revolving door is convenient for two direction travel, but it also serves as a security measure at a what?<br>Option A: bank<br>Option B: library<br>Option C: department store<br>Option D: mall<br>Option E: new york<br>Answer: A |

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

| 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.<br>How many field goals did both teams kick in the first half?<br>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.<br>Question: How many field goals did both teams kick in the first half?<br>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.<br>does ethanol take more energy make that produces (yes or no)<br>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.<br>Question: does ethanol take more energy make that produces<br>Answer: no |

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

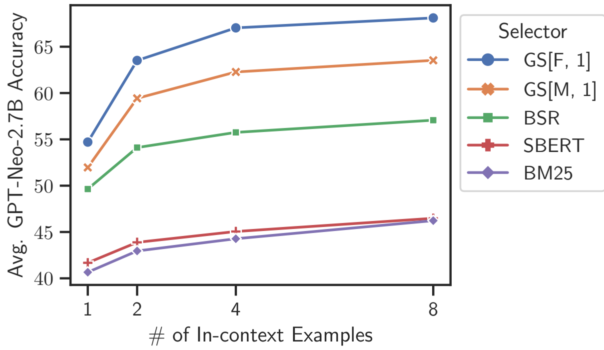


*Figure 8.* Average ICL performance with GPT-Neo-2.7B for varying number of in-context examples. Both GS[F] and GS[M] are consistently better, and both surpass 8-shot ICL using SBERT and BM25 with just 1 example!

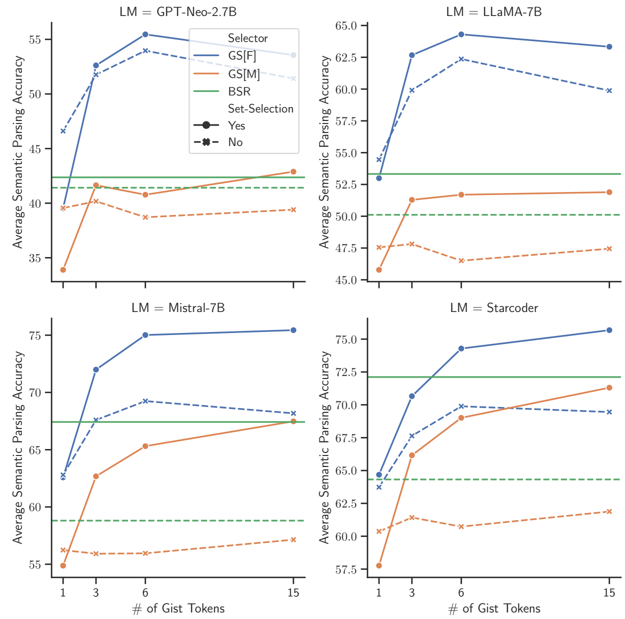


*Figure 9.* 8-shot ICL with different LLMs on semantic parsing datasets using multi-task trained and fine-tuned GistScore with varying number of gist tokens and its set extension.

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.

## D. Additional Results

**Results for GistScore-variations and all baselines** Tables 13, 14, 15, 16, 17, 18 and 19 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 9 and Table 11 compare performance for different number of gist tokens and set-selection for different LLMs.

**Varying number of shots** Figure 8 shows the average ICL performance with GPT-Neo-2.7B for varying number of in-context examples.

**Impact of gist model size** Table 16 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 Speeds** Table 12 provides the time taken to select 8 ICL examples using various selection methods.

## E. Additional Analyses

**Qualitative analysis of example selection** Tables 9, 8, and 10 compare examples selected by various selectors for instances from DROP, SST5 and SMCalFlow, respectively. In particular, Tables 9 and 8 show that examples selected using GS[F, 1] are semantically more relevant to test inputs than SBERT while Table 10 shows an example where set-selection using SET-GS[F, 1] is beneficial.

**Performance on classification tasks** Figure 10 shows that the majority-vote accuracy of the in-context examples selected by the various selectors is strongly correlated with ICL performance across all classification datasets. Figure 11 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 12 shows t-SNE visualizations of salient information in gist embeddings for additional datasets. Figure 13 shows that the salient aspects seen in t-SNE visualizations in Figures 6 and 12 can also be observed in PCA visualizations.

**Gist tokens are different from standard tokens** Figure 14 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

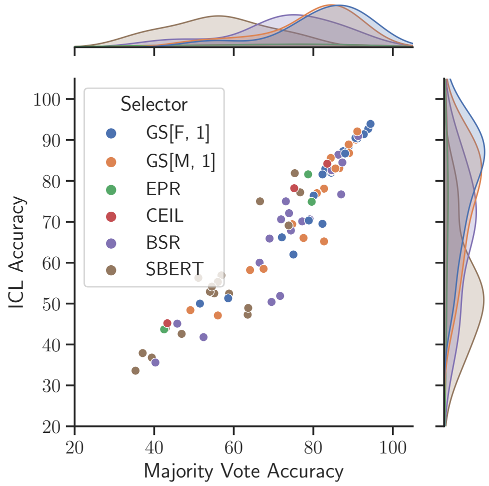


*Figure 10.* 8-shot GPT-Neo-2.7B ICL accuracy v/s the Majority Vote accuracy of the various selectors on all the classification datasets.

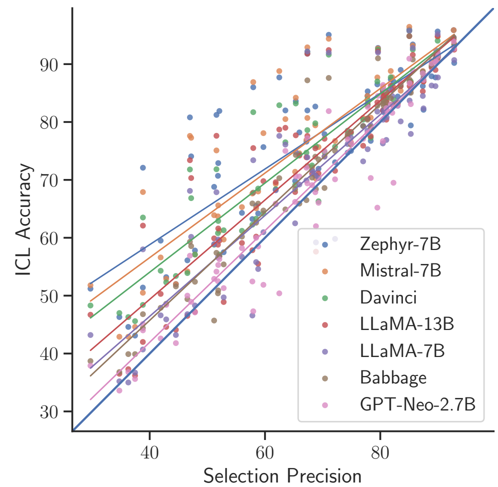


*Figure 11.* ICL accuracy v/s selection precision *i.e.* the fraction of in-context examples with the test label for the various classification datasets with fixed label sets, selectors, and LLMs. While the ICL accuracy of all LLMs improves with more accurate selection, larger LLMs are less reliant on it.

compared to ordinary language tokens.

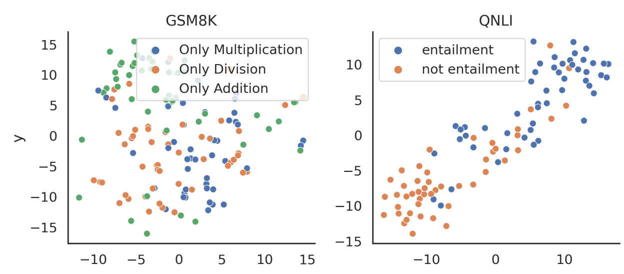


*Figure 12.* t-SNE Visualizations of gist embeddings for additional datasets. For QNLI, gist embeddings encode class labels. For GSM8K, they encode whether the solution can be obtained by a chain-of-thought reasoning comprising only addition, only multiplication, or only division.

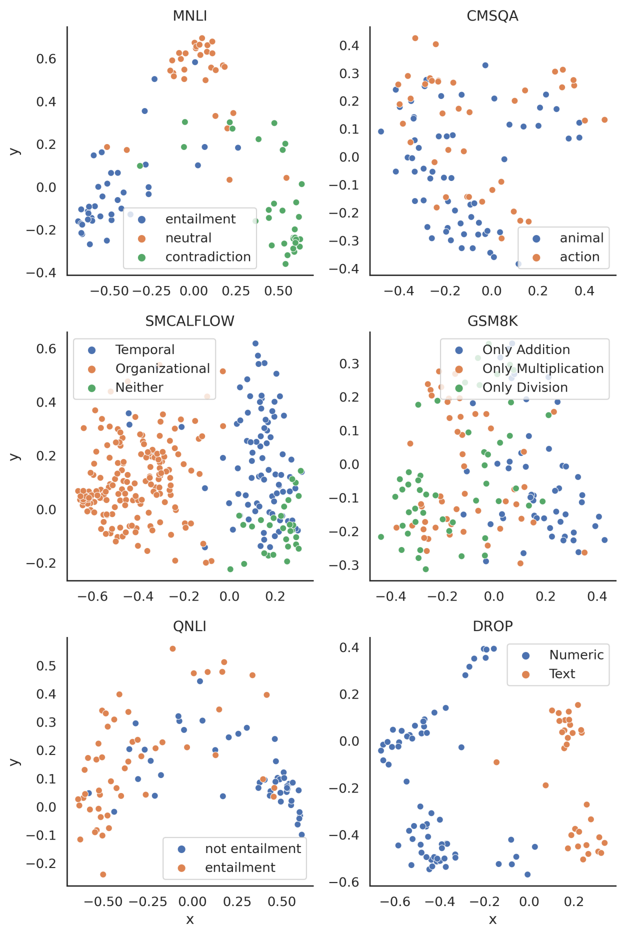


*Figure 13.* PCA visualizations of gist embeddings show similar results as t-SNE visualization in Figure 6 and 12. Gist embeddings encode task-specific salient information such as class labels (MNLI, QNLI) or more abstract information aspects (CMSQA, SMCalFlow, DROP, GSM8K) that help retrieve better in-context examples.

| Selector | Selected Examples |
| --- | --- |
| SBERT | `Review: at the one-hour mark, herzog simply runs out of ideas and the pace turns positively leaden as the movie sputters to its inevitable tragic conclusion.`<br>`Sentiment: terrible`<br><br>`Review: it is very difficult to care about the character, and that is the central flaw of the film.`<br>`Sentiment: terrible`<br><br>`Review: a film that loses sight of its own story.`<br>`Sentiment: bad`<br><br>`Review: one of those films that started with a great premise and then just fell apart.`<br>`Sentiment: terrible` |
| GS[F, 1] | `Review: the film starts promisingly, but the ending is all too predictable and far too cliched to really work.`<br>`Sentiment: bad`<br><br>`Review: all the pieces fall together without much surprise, but little moments give it a boost.`<br>`Sentiment: good`<br><br>`Review: despite an overwrought ending, the film works as well as it does because of the performances.`<br>`Sentiment: good`<br><br>`Review: the filmmakers keep pushing the jokes at the expense of character until things fall apart.`<br>`Sentiment: bad` |

*Table 8.* Examples selected by SBERT and GS[F, 1] for the test input *in the end, the movie collapses on its shaky foundation despite the best efforts of director joe carnahan.* from SST5 that has the correct sentiment label `terrible`. While GS[F, 1] selects other reviews that highlight some positives but have overall negative sentiment, SBERT also selects reviews of the opposite kind, i.e. noce which highlight some negative aspects but have overall positive sentiment.

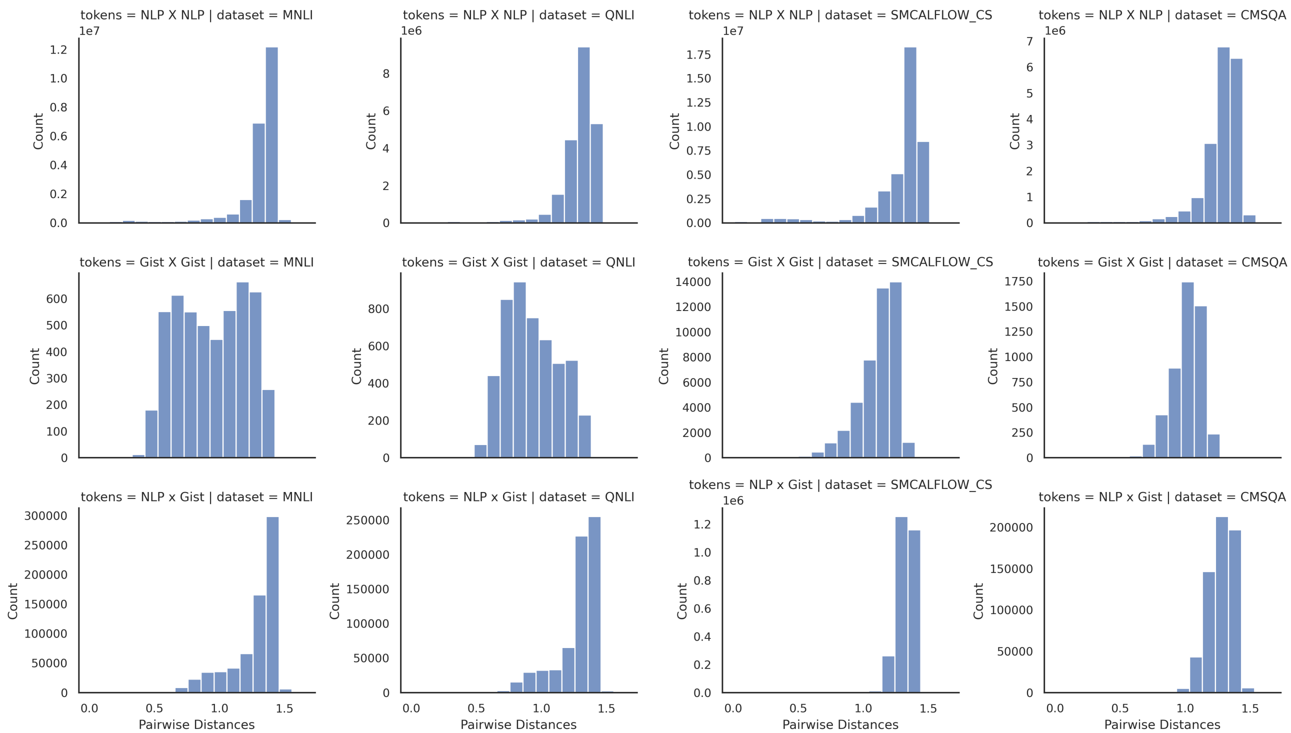


*Figure 14.* Pairwise Distances between Gist and NLP token activations.

| Selector | Selected Examples |
|---|---|
| SBERT | `Question: How many more TD passes did Romo throw than Palmer?`<br>`Answer: 1`<br>`Question: How many TD passes between 5 yards and 20 yards were thrown?`<br>`Answer: 3`<br>`Question: How many total yards were scored on passing touchdowns?`<br>`Answer: 104`<br>`Question: How many field goals did the Bengals score in the second quarter?`<br>`Answer: 2` |
| GS[F, 1] | `Question: How many field goals did Graham kick in the second quarter?`<br>`Answer: 2`<br>`Question: How many field goals did Lindell kick in the third quarter?`<br>`Answer: 2`<br>`Question: How many field goals did Bryant kick in the second half?`<br>`Answer: 3`<br>`Question: How many field goals were made in the second quarter?`<br>`Answer: 3` |

*Table 9.* Examples selected by SBERT and GS[F, 1] for the test input *How many second quarter field goals did the Bengals get?* from DROP dataset. Questions selected by GS[F, 1] are much more semantically relevant with similar reasoning patterns than SBERT.

| Selector | Selected Examples |
|---|---|
| SBERT | `Add meeting for work after drinks for two hours`<br>`CreateEvent(AND(has_subject(" meeting for work "),starts_at(:end(FindEvents(has_subject(" drinks ")))),has_duration(toHours(2))))`<br><br>`create work meeting with my boss`<br>`CreateEvent(AND(with_attendee(FindManager(CurrentUser())),has_subject(" work meeting ")))`<br><br>`create drinks with sarah after work`<br>`CreateEvent(AND(with_attendee(" sarah "),has_subject(" drinks "),starts_at(:end(FindEvents(has_subject(" work "))))))`<br><br>`create drinks after work`<br>`CreateEvent(AND(has_subject(" drinks "),starts_at(:end(FindEvents(has_subject(" work "))))))`<br><br>*`create work drinks after meeting with boss`*<br>`CreateEvent(AND(has_subject(" work drinks "),starts_at(:end(FindEvents(AND(has_subject(" meeting "),with_attendee(FindManager(CurrentUser()))))))))` |
| GS[F, 1] | `Let's put drinks with hannah after meeting with hannah`<br>`CreateEvent(AND(with_attendee(" hannah "),has_subject(" drinks "),starts_at(:end(FindEvents(with_attendee(" hannah "))))))`<br><br>`create drinks after work`<br>`CreateEvent(AND(has_subject(" drinks "),starts_at(:end(FindEvents(has_subject(" work "))))))`<br><br>`work meeting after school`<br>`CreateEvent(AND(has_subject(" work meeting "),starts_at(:end(FindEvents(has_subject(" school "))))))`<br><br>`Thank you. Can you schedule a coffee break immediately after my meeting with Susan`<br>`CreateEvent(AND(has_subject(" coffee break "),starts_at(:end(FindEvents(with_attendee(" Susan "))))))`<br><br>*`create work drinks after meeting with boss`*<br>`CreateEvent(AND(has_subject(" work drinks "),starts_at(:end(FindEvents(with_attendee(" boss "))))))` |
| SET-GS[F, 1] | `create work meeting with my boss`<br>`CreateEvent(AND(with_attendee(FindManager(CurrentUser())),has_subject(" work meeting ")))`<br><br>`Let's put drinks with hannah after meeting with hannah`<br>`CreateEvent(AND(with_attendee(" hannah "),has_subject(" drinks "),starts_at(:end(FindEvents(with_attendee(" hannah "))))))`<br><br>`work meeting after school`<br>`CreateEvent(AND(has_subject(" work meeting "),starts_at(:end(FindEvents(has_subject(" school "))))))`<br><br>`Thank you. Can you schedule a coffee break immediately after my meeting with Susan`<br>`CreateEvent(AND(has_subject(" coffee break "),starts_at(:end(FindEvents(with_attendee(" Susan "))))))`<br><br>*`create work drinks after meeting with boss`*<br>`CreateEvent(AND(has_subject(" work drinks "),starts_at(:end(FindEvents(with_attendee(FindManager(CurrentUser())))))))` |

*Table 10.* Examples selected by SBERT, GS[F, 1], and SET-GS[F, 1] for the test input *create work drinks after meeting with boss* from SMCalFlow's CG split. The correct model output is highlighted in green while incorrect ones are in red. Independent-ranking based selection using SBERT and GS[F, 1] fails to demonstrate all functionalities: SBERT doesn't select any instances that require finding the end time of a meeting with an attendee (`starts_at(:end(FindEvents(with_attendee())))`) whereas GS[F, 1] fails to select any examples demonstrating finding the user's manager (`FindManager(CurrentUser())`). Set-selection using SET-GS[F, 1] ensures that all the necessary functionalities are demonstrated leading to correct model output.

| 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 11.* 8-shot ICL results for varying number of gist tokens ($l$) and set-selection for semantic parsing datasets with different LLMs.

| 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 12.* 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.

| 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 13.* 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 14.* 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.

| 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 15.* 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 16.* 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 17.* 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 18.* 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 19.* 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.