What In-Context Learning "Learns" In-Context: Disentangling Task Recognition and Task Learning

Jane Pan **Tianyu Gao** **Howard Chen** **Danqi Chen**
Department of Computer Science, Princeton University
{jp7224,tianyug,howardchen,danqic}@cs.princeton.edu

Abstract

Large language models (LLMs) exploit in-context learning (ICL) to solve tasks with only a few demonstrations, but its mechanisms are not yet well-understood. Some works suggest that LLMs only recall already learned concepts from pre-training, while others hint that ICL performs implicit learning over demonstrations. We characterize two ways through which ICL leverages demonstrations. *Task recognition* (TR) captures the extent to which LLMs can recognize a task through demonstrations – even without ground-truth labels – and apply their pre-trained priors, whereas *task learning* (TL) is the ability to capture new input-label mappings unseen in pre-training. Using a wide range of classification datasets and three LLM families (GPT-3, LLaMA and OPT), we design controlled experiments to disentangle the roles of TR and TL in ICL. We show that (1) models can achieve non-trivial performance with only TR, and TR does not further improve with larger models or more demonstrations; (2) LLMs acquire TL as the model scales, and TL's performance consistently improves with more demonstrations in context. Our findings unravel two different forces behind ICL and we advocate for discriminating them in future ICL research due to their distinct nature.[1]

1 Introduction

Large language models (LLMs) have demonstrated the ability to perform in-context learning (ICL), i.e., "learning" to perform a task purely from examples in the context without any parameter updates (Brown et al., 2020). This powerful and flexible phenomenon enables LLMs to be used as general-purpose models that can perform any task with a small set of labeled examples.

However, there is still no consensus on how in-context learning works. Some previous work hypothesizes that during pre-training, LLMs implicitly learn tasks required for downstream applications, and the in-context demonstrations merely provide information that allow the model to recognize which task is required (Xie et al., 2022). Min et al. (2022) show empirical evidence of this hypothesis by demonstrating that ICL performance is insensitive to the usage of ground-truth labels.

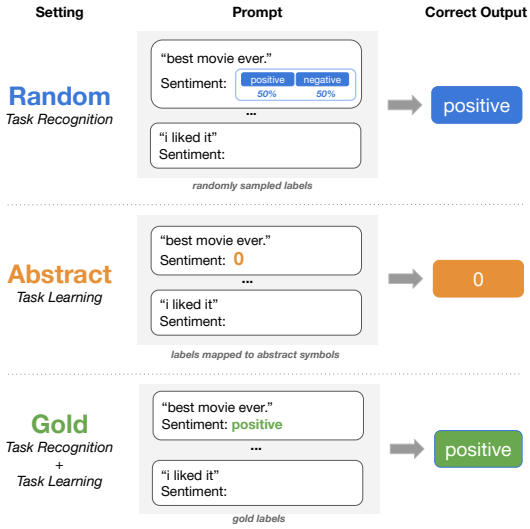

Figure 1: We perform experiments in three settings: RANDOM (top), ABSTRACT (middle), and GOLD (bottom). Our experiments demonstrate that *task recognition* (TR; shown by RANDOM) does not scale with model sizes and number of demonstrations, while *task learning* (TL; shown by ABSTRACT) does.

On the other hand, Akyürek et al. (2023); von Oswald et al. (2022) construct theories that Transformer-based models may perform implicit gradient descent to update an "inner-model", and Dai et al. (2023) demonstrate similarities between in-context learning and explicit fine-tuning through a series of metrics on real-world datasets. Such hypotheses assume the correct input-output mappings are important and ICL actually performs implicit learning over demonstrations.

In this paper, we disentangle ICL into **task**

[1]Our code is publicly available at https://github.com/princeton-nlp/WhatICLLearns.

Findings of the Association for Computational Linguistics: ACL 2023, pages 8298–8319
July 9-14, 2023

recognition (TR), which recognizes the task from demonstrations and applies LLMs' pre-trained priors, and **task learning** (TL), which learns a new input-label mapping from demonstrations. In common ICL scenarios where ground-truth labels are provided, TR and TL take effect simultaneously. We propose two settings to tease them apart: 1) RANDOM, where the labels are uniformly sampled from the label space (Min et al., 2022), in order to restrict LLMs to only apply TR; 2) ABSTRACT, where the labels are replaced with abstract symbols (e.g., numbers or letters) that never co-occurred with the inputs in pre-training. We focus on how the two abilities in ICL evolve with two factors – *model sizes* and *numbers of demonstrations*, which have been neglected in related literature.

Through extensive experiments with a series of classification datasets on GPT-3 (Brown et al., 2020), LLaMA (Touvron et al., 2023), and OPT (Zhang et al., 2022), we find:

• The gap between GOLD and RANDOM is small with smaller models, corroborating with Min et al. (2022). However, with larger models and more examples, the gap becomes larger. This suggests TR plays a significant role in ICL, but it does not scale with increasing parameters or examples.

• LLMs also perform TL, which emerges with larger models and more demonstrations. With the largest model and more than 16 examples, ABSTRACT outperforms RANDOM, pointing to a paradigm shift in in-context learning at scale.

Together, our findings provide a better way to understand ICL behaviors.[2]

2 Task Recognition and Task Learning

An LLM (parameterized by θ) performs ICL by conditioning on the input-label pair demonstrations $\mathcal{D}_{\text{demo}} = (x_1, y_1, x_2, y_2, \ldots, x_K, y_K)$ and the test input x_{test} to predict the label $y_{\text{test}} \sim p_\theta(y \mid \mathcal{D}_{\text{demo}}, x_{\text{test}})$, where the demonstrations elicit a mapping $f : \mathcal{X} \to \mathcal{Y}, x \in \mathcal{X}, y \in \mathcal{Y}$. We delineate two ways an LLM can leverage in-context demonstrations: *task recognition* and *task learning*.

Task recognition (TR) represents models' ability to recognize the mapping f purely by observing the input distribution $\{x_i\}_{i=1}^K$ and the label distribution $\{y_i\}_{i=1}^K$, without the provided (x_i, y_i) pairs. The LLM then applies its pre-trained priors to the recognized f. Formally, when only TR is enabled,

$$\begin{aligned} &p_\theta(y \mid x_{\text{test}}, \{x_i, y_i\}_{i=1}^K) \\ =&p_\theta(y \mid x_{\text{test}}, \{x_i\}_{i=1}^K, \{y_i\}_{i=1}^K), \end{aligned}$$

which suggests TR does not rely on the pair information. For example, an input distribution of movie reviews and a label distribution of "The sentiment is positive/negative" can be easily recognized as a sentiment classification task due to their prevalence during pre-training, and LLMs can make reasonable predictions without explicitly "learning" the task via ground-truth demonstrations. This leads to observations that the model can still perform well even when we provide wrong input-label mappings, e.g., "The movie is great. The sentiment is *negative*" (Min et al., 2022). Task learning (TL), on the other hand, characterizes how the model learns a new mapping from the input-label pairs through demonstrations. Unlike TR, TL allows models to learn novel mappings and thus correct input-label pairs will be crucial.

We posit that the two mechanisms occur under separate conditions, as recognizing an already learned task is easier than learning a new mapping. Models are able to perform TR at a small scale, but this ability does not drastically improve with increasing model sizes and demonstrations; on the other hand, TL improves significantly when model sizes and numbers of demonstrations increase. To show the above phenomenon, we disentangle TR and TL through *label space manipulation*, including three different setups (examples in Figure 1):

• GOLD: the standard ICL setting where we use natural prompts and gold input-label pairs. This setup reflects both TR and TL abilities.

• RANDOM: similar to Min et al. (2022), we use the same natural prompts as GOLD and sample demonstration labels uniformly at random from the label space. This setup reflects TR only.

• ABSTRACT: we use minimal prompts (which provide no task information) and characters with no clear semantic meanings (e.g. numbers, letters, and random symbols) as the label for each class. We found that even abstract labels may have biases in pre-training, e.g., "0" is biased towards negative. Hence, for *each* prompt $x_1, y_1, \ldots, x_K, y_K$, we randomly sample a 1-1 mapping $\phi : \mathcal{Y} \to \mathcal{Y}^*$ to avoid any bias, and no task-specific information is leaked in either the prompt template or the label

[2] We discuss the differences between our work and Min et al. (2022); Yoo et al. (2022) in Section 5, detailing how our findings deviate and converge with existing results.

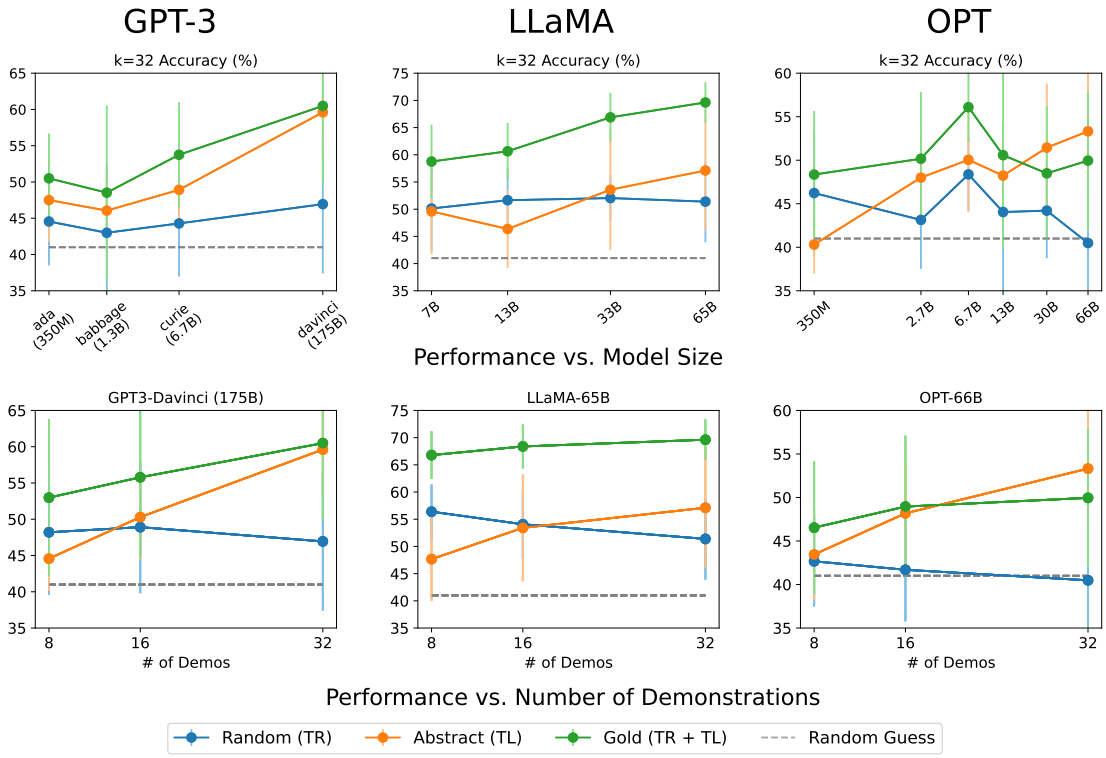

Figure 2: Averaged accuracy across 16 datasets for GPT-3 (left), LLaMA (middle), and OPT (right). Top graphs plot model sizes from small to large against performance, using 32 examples. Variance is calculated across three prompts. Bottom graphs plot #demonstrations against performance for `davinci`, LLaMA-65B, and OPT-66B.

space. To evaluate the model's ABSTRACT performance, we measure its accuracy using $\phi(y_{\text{test}})$ as target labels. Since these input-label mappings are never seen in pre-training, it reflects the TL ability.

In the following sections, we conduct comprehensive experiments with the above three different settings under two axes – model sizes and numbers of demonstrations – and show how TR and TL manifest under different conditions.

3 Experimental Setup

3.1 Datasets

We experiment on 16 classification datasets across 4 type of tasks: sentiment analysis, toxicity detection, natural language inference/paraphrase detection, and topic/stance classification. All datasets and references are in Appendix A. Our dataset selection largely follows Min et al. (2022), but we exclude multi-choice datasets since it is difficult to apply our ABSTRACT experiments on them.

3.2 Models

We use three state-of-the-art LLM families: GPT-3 (Brown et al., 2020), LLaMA (Touvron et al., 2023), and OPT (Zhang et al., 2022). We use GPT-3 ada (350M), babbage (1.3B), `curie` (6.7B), and `davinci` (175B) via the OpenAI API. For OPT, we use checkpoints from the Transformers library (Wolf et al., 2020), with model sizes of 350M, 2.7B, 6.7B, 13B, 30B, and 66B parameters. For LLaMA, we use model sizes of 7B, 13B, 33B, and 65B parameters.[3]

3.3 Task Setup

We adopt the sample-based evaluation protocol: for each test example, we sample a different set of demonstrations from the training set. We manually design 3 prompt templates for each type of classification tasks in a similar style to the prompts from Min et al. (2022). We report the mean by averaging across datasets and prompts, and standard variation across different prompts for each datapoint. For GPT-3, we sample 150 examples for each dataset. We use fewer examples due to budget constraints, and GPT-3 presents lower variance than other model families. For OPT and LLaMA, we sample 1,350 examples for all datasets.

[3]For GPT-3, we use the non-instruction legacy models for fair comparison to OPT and LLaMA models. We did not run experiments on the largest OPT-175B model due to computational constraints.

We design two kinds of prompts: *natural language prompts* (Table 1), which are similar to the manual prompts in Min et al. (2022), and *minimal prompts* (Table 3), which remove any natural language instructions for the task. For ABSTRACT, we tested three types of label choices: *numbers* ($0, \ldots, N-1$, where N is the number of classes), *letters* (N letters from A, B, C, ...), and *symbols* (first N symbols of "@", "#", "$", "%", "*", and "^"). For each test example, we randomly sample a new mapping between labels and abstract characters. We report the *number* abstract labels in all the main results and compare the three forms in §4.2.

4 Results

Figure 2 shows our main results with GPT-3, LLaMA, and OPT with our 3 settings: GOLD, RANDOM, and ABSTRACT. Below we summarize the trends of TR and TL across different conditions.

4.1 Main Results

Summary of overall trends. We first verify that GOLD consistently performs the best across model families and number of demonstrations, which is expected given that the GOLD setting provides the model with all information. Overall, the RANDOM curves do not increase with either model sizes or number of demonstrations, remaining largely flat; considering the scenario with *small* model sizes and *few* examples ($K = 8$), there is an insignificant gap between RANDOM and GOLD. Meanwhile, the ABSTRACT curves demonstrate an increasingly steep slope as the model sizes and the number of demonstrations grow; with small models or small K, ABSTRACT mostly underperforms RANDOM, whereas ABSTRACT with largest models and $K = 32$ performs well above RANDOM (and may even be competitive with GOLD). We note that the OPT curves demonstrate significant variance, which we hypothesize to be a result of the models potentially being under-trained. We elaborate the takeaways on TR and TL below.

Task recognition is a broader capability across scales. For all model families, the RANDOM setting shows similar performance at all sizes and numbers of demonstrations. Moreover, TR performance is significantly stronger than the random baseline, even with small models and few examples. For instance, even the smallest 350M parameter models are able to recognize the task using

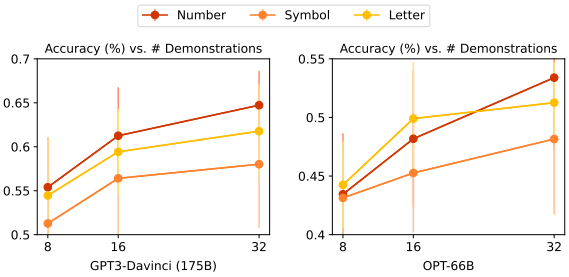

Figure 3: Performance of three types of ABSTRACT labels: numbers, letters, and symbols on `davinci` and OPT-66B.

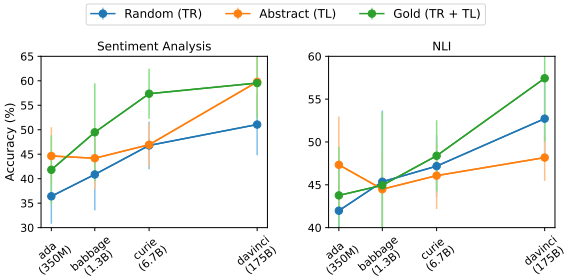

Figure 4: Average results of sentiment analysis datasets (left) vs. natural language inference datasets (right) on GPT-3 models, with $K = 32$.

just 8 examples, drawing around 10 points of average performance lead against the random baseline for GPT-3 `ada` and 5 points for OPT-350M. This shows that task recognition from in-context examples does not drastically scale with model sizes or numbers of examples.

Task learning is enabled with scale. We observe that TL is dependent on model sizes: smaller models perform roughly the same across all numbers of demonstrations (see Figure 6). On the other hand, larger models can utilize the provided mapping information and perform TL, as ABSTRACT (TL) performance increases drastically with larger sizes (first row of Figure 2). When using a larger model, the results also improve as the number of demonstration increases (second row of Figure 2). With only 16 examples, OPT-66B and `davinci` are able to match the performance of GOLD while using a new label mapping. While LLaMA-65B's ABSTRACT is not as competitive as its GOLD, the trend of improving ABSTRACT performance with larger size s or larger K is clear. This suggests that TL is only enabled by scales and further improves with more demonstrations.

4.2 Further Analysis

The trends for task learning generalize across different types of abstract labels. In Figure 3, we show ABSTRACT results with number, letter, and symbol labels respectively. We observe that all three versions show a similar trend and coincide with our main results. Numbers and letters perform consistently better than symbols. This may be because letters and numbers appear more frequently in the pre-training corpus, and therefore make for a more "natural" label space.

Task difficulty affects the trends. We notice that ABSTRACT scales better with sizes and examples when the task is simpler. In Figure 4 we compare two types of tasks: sentiment analysis and natural language inference (NLI). Since NLI is more difficult, we observe that it produces a flatter ABSTRACT curve, suggesting that the model relies more on the natural prompts and pre-training priors to solve those tasks. We demonstrate the full task-type breakdown results in §C.

5 Related Work

Many works have attempted to deepen empirical or theoretical understanding of ICL since its emergence in Brown et al. (2020). For instance, Xie et al. (2022) present a theoretical framework where latent "concepts" parameterize each document in pre-training. They posit that all concepts have been learned in pre-training; thus, ICL is the result of implicit Bayesian inference, where the LM uses in-context demonstrations as evidence to identify the correct concept. Min et al. (2022) present empirical evidence for this framework by showing that only limited information, rather than true input-label mappings, is needed to perform ICL.

Other works investigate the impact of the pre-training corpus on ICL. Chan et al. (2022) identify properties of the pre-training distribution that enable ICL behavior, including burstiness, label multiplicity, and a long-tailed class distribution – all of which are satisfied by natural language. Razeghi et al. (2022) show that the frequencies of terms in the pre-training corpora is positively correlated with model performance. Kirsch et al. (2022) show that both a rich training distribution and a sufficiently large model are critical to the development of in-context learning abilities.

More recently, several works have explored theoretical frameworks in which ICL can be seen as implicit gradient descent, treating a forward pass over the in-context demonstrations as an "update" to an implicit internal model. (Akyürek et al., 2023; von Oswald et al., 2022; Dai et al., 2023). For mechanistic perspectives on ICL, Olsson et al. (2022) and Bansal et al. (2022) identify induction heads (subnetworks that perform in-context pattern recognition) in small and large models, respectively.

While our conclusions align with aspects of previous studies, our work contributes novel insights on multiple axes. Min et al. (2022) also show that even small models can perform TR and argue that the performance gap between GOLD and RANDOM is consistently small, but most of their experiments are on $\leq$13B models with 16 demonstrations; we show that as model sizes scale, GOLD tends to improve while RANDOM does not. Thus, the performance deficit of RANDOM grows as models become larger. Yoo et al. (2022) also perform similar experiments to RANDOM and ABSTRACT, but they do not deeply investigate the effect of model sizes or numbers of demonstrations. Contemporary work Wei et al. (2023) obtain similar results; additionally, they show that instruction-tuning strengthens the model's semantic priors more than it improves TL. However, they primarily focus on closed-source models, whereas we also conduct experiments on public models such as LLaMA and OPT. Collectively, our findings offer a comprehensive understanding of how ICL works across scales.

6 Conclusion

While previous work often studies ICL as an umbrella term, regardless of model sizes and numbers of examples, we argue that there are two distinct characterizations of ICL – task recognition and task learning – and demonstrate that they emerge under different conditions. Even small models are capable of performing TR, but this ability does not scale. On the other hand, TL is an emergent ability of large models; small models are unable to perform TL even when provided with more demonstrations, whereas large models can leverage more demonstations to improve their TL performance. We suggest that future work on ICL should distinguish the two phenomena and clearly state the conditions under which the experiments are conducted.

Limitations

Though LLMs with in-context learning are capable of all kinds of NLP tasks, this work is limited to

classification tasks because they are easier to be adapted to our RANDOM and ABSTRACT setup. We leave other types of NLP tasks as future work.

Another limitation of our work lies in the definition and discussion of task learning. Though we empirically show that large models are capable of acquiring a novel mapping to abstract labels like numbers or letters, how models "learn" mechanistically is still elusive. As suggested in previous work, LLMs may conduct implicit gradient descent over demonstrations, or they may alternatively map the patterns shown in the demonstrations back to concepts learned in pre-training. To some extent, these mechanisms could be considered an advanced form of "task recognition". This work only designs experiments to better observe and disentangle TR and TL, and we look forward to further studies that reveal more insights about the mechanistic inner-workings of these phenomena in ICL.

Acknowledgements

We thank the members of the Princeton NLP group for their valuable advice, thoughts, and discussions. We also appreciate the helpful feedback given by the anonymous reviewers and the area chairs. This project was partially supported by the National Science Foundation under Award IIS-2211779, and a Sloan Fellowship.

References

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. 2023. What learning algorithm is in-context learning? investigations with linear models. In International Conference on Learning Representations (ICLR).

Hritik Bansal, Karthik Gopalakrishnan, Saket Dingliwal, Sravan Bodapati, Katrin Kirchhoff, and Dan Roth. 2022. Rethinking the role of scale for in-context learning: An interpretability-based case study at 66 billion scale. arXiv preprint arXiv:2212.09095.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In Proceedings of the 13th International Workshop on Semantic Evaluation, pages 54–63. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In Empirical Methods in Natural Language Processing (EMNLP), pages 632–642.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In Advances in Neural Information Processing Systems (NeurIPS), volume 33, pages 1877–1901.

Stephanie Chan, Adam Santoro, Andrew Lampinen, Jane Wang, Aaditya Singh, Pierre Richemond, James McClelland, and Felix Hill. 2022. Data distributional properties drive emergent in-context learning in transformers. In Advances in Neural Information Processing Systems (NeurIPS), volume 35, pages 18878–18891.

Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Shuming Ma, Zhifang Sui, and Furu Wei. 2023. Why can GPT learn in-context? language models implicitly perform gradient descent as meta-optimizers. In ICLR 2023 Workshop on Mathematical and Empirical Understanding of Foundation Models.

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In Proceedings of the Third International Workshop on Paraphrasing (IWP2005).

Louis Kirsch, James Harrison, Jascha Sohl-Dickstein, and Luke Metz. 2022. General-purpose in-context learning by meta-learning transformers. arXiv preprint arXiv:2212.04458.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning.

P. Malo, A. Sinha, P. Korhonen, J. Wallenius, and P. Takala. 2014. Good debt or bad debt: Detecting semantic orientations in economic texts. Journal of the Association for Information Science and Technology.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In International Conference on Language Resources and Evaluation (LREC), pages 216–223.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In Empirical

Methods in Natural Language Processing (EMNLP), pages 11048–11064.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SemEval-2018 task 1: Affect in tweets. In Proceedings of the 12th International Workshop on Semantic Evaluation, pages 1–17. Association for Computational Linguistics.

Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas. 2020. Ethos: an online hate speech detection dataset. arXiv preprint arXiv:2006.08328.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. 2022. In-context learning and induction heads. arXiv preprint arXiv:2209.11895.

Yasaman Razeghi, Robert L Logan IV, Matt Gardner, and Sameer Singh. 2022. Impact of pretraining term frequencies on few-shot numerical reasoning. In Findings of Empirical Methods in Natural Language Processing (EMNLP), pages 840–854. Association for Computational Linguistics.

Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. 2018. CARER: Contextualized affect representations for emotion recognition. In Empirical Methods in Natural Language Processing (EMNLP), pages 3687–3697.

Emily Sheng and David Uthus. 2020. Investigating societal biases in a poetry composition system. In Proceedings of the Second Workshop on Gender Bias in Natural Language Processing, pages 93–106. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In Empirical Methods in Natural Language Processing (EMNLP), pages 1631–1642.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.

Johannes von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. 2022. Transformers learn in-context by gradient descent. arXiv preprint arXiv:2212.07677.

Ellen M Voorhees and Dawn M Tice. 2000. Building a question answering test collection. In Association for Computing Machinery Special Interest Group in Information Retrieval (ACM SIGIR), pages 200–207.

Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, et al. 2023. Larger language models do in-context learning differently. arXiv preprint arXiv:2303.03846.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In Empirical Methods in Natural Language Processing (EMNLP), pages 38–45.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2022. An explanation of in-context learning as implicit bayesian inference. In International Conference on Learning Representations (ICLR).

Kang Min Yoo, Junyeob Kim, Hyuhng Joon Kim, Hyunsoo Cho, Hwiyeol Jo, Sang-Woo Lee, Sang-goo Lee, and Taeuk Kim. 2022. Ground-truth labels matter: A deeper look into input-label demonstrations. In Empirical Methods in Natural Language Processing (EMNLP), pages 2422–2437.

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

A Datasets

We use a total of 16 datasets. **Sentiment analysis** includes SST-2 (Socher et al., 2013), financial_phrasebank (Malo et al., 2014), emotion (Saravia et al., 2018), and poem_sentiment (Sheng and Uthus, 2020) **Topic/stance classification** includes TREC (Voorhees and Tice, 2000), tweet_eval_atheist, and tweet_eval_feminist (Mohammad et al., 2018; Basile et al., 2019). **Toxicity detection** includes tweet_eval_hate, ethos_race, ethos_gender, ethos_national_origin, and ethos_religion (Mollas et al., 2020) **Natural language inference/paraphrase detection** includes SICK (Marelli et al., 2014), SNLI (Bowman et al., 2015), WNLI (Levesque et al., 2012), and MRPC (Dolan and Brockett, 2005).

We sample from the training set to construct the prompts; following Min et al. (2022), we use the development set for evaluation, using sampled $\max(1350, \text{dataset\_size})$ examples.

B Prompt Templates

For each task category (e.g. sentiment classification, topic detection), we manually design three natural language templates. Depending on exact specifications for the dataset, templates may be adjusted to better reflect the task (e.g. "Is this atheist?" for tweet_eval_atheist). We apply these templates to the natural language label sets (GOLD and RANDOM). All prompts are presented in Table 1.

We also design two task-agnostic variations on three minimal templates for ABSTRACT: one for single-sentence tasks and one for multi-sequence tasks (e.g. NLI tasks). We use these minimal templates on the abstract language label sets in order to prevent the model from being exposed to any information regarding the task from the prompt design. All minimal templates are presented in Table 3

All prompts are designed to be answered with single-token responses (e.g. "Yes/No", "True/False", "positive/negative/neutral", "0/1/2", "A/B/C") so that we can directly check models' last token prediction results instead of applying decoding methods.

C More Results

We demonstrate average model performance with respect to number of parameters in Figure 5. It is clear that small models struggle to perform ABSTRACT, regardless of how many examples, whereas the largest models (especially GPT-3 Davinci and OPT-66B) are able to perform ABSTRACT. Additionally, their performance improves even more when more demonstrations are provided.

We demonstrate average model performance with respect to numbers of demonstrations in Figure 6. We can see a clear trend that RANDOM (TR) does not change much but ABSTRACT improves drastically with more examples, especially for GPT-3 Davinci and OPT-66B.

Figure 7 shows all the ABSTRACT results and demonstrates a similar trend to what §4.2 describes.

Figure 8, Figure 9, Figure 10, and Figure 11 show task-type breakdown results. Though individual task-type results are more noisy, we can make a similar observation compared to the main result – ABSTRACT (TL) scales better with sizes and numbers of examples compared to RANDOM (TR).

Type	Template #	Example
Sentiment Analysis	1	<s> The sentiment is <positive/negative>
	2	<s> Sentiment: <positive/negative>
	3	<s> The sentiment of the text is <positive/negative>
Hate Speech	1	<s> Is this hate speech? <Yes/No>
	2	<s> Is the sentence hateful? <Yes/No>
	3	<s> The sentence contains hate speech. True or False? The answer is <True/False>
Stance Detection	1	<s> The stance is feminist. True or False? The answer is <True/False>
	2	<s> Does the sentence express a feminist view? <Yes/No>
	3	<s> Is the stance feminist? <Yes/No>
Topic Detection	1	<s> The topic is <label>
	2	<s> The sentence is about <label>
	3	<s> Sentence topic: <label>

Table 1: Natural prompts used as input in GOLD and RANDOM settings for single-sentence datasets. <s> denotes the input sequence; labels are illustrated in red.

Type	Temp. #	Example
Entailment	1	<*s1*> The question is: <*s2*>? True or False? The answer is <*True/False*>
	2	Hypothesis: <*s1*> Premise: <*s2*>? Do the sentences show entailment? <*Yes/No*>
	3	The hypothesis is: <*s1*> The premise is: <*s2*>? Is this entailment? <*Yes/No*>
NLI	1	<*s1*> The question is: <*s2*> True, False, or Unknown? The answer is <*True/False/Unknown*>
	2	Hypothesis: <*s1*> Premise: <*s2*>? Given the premise, is the hypothesis true? Yes, No, or Unknown? The answer is: <*Yes/No/Unknown*>
	3	The hypothesis is: <*s1*> The premise is: <*s2*>? According to the premise, the hypothesis is true. True, False, or Unknown? The answer is: <*True/False/Unknown*>
Paraphrase Detection	1	<*s1*> The question is: <*s2*> True or False? The answer is: <*True/False/*>
	2	Sentence 1: <*s1*> Sentence 2: <*s2*> These sentences are paraphrases. True or False? The answer is: <*True/False/*>
	3	Text: <*s1*> Consider this sentence: <*s2*> Does it paraphrase the text? <*Yes/No*>

Table 2: Natural prompts used as input in GOLD and RANDOM settings for multi-sentence datasets. <*s1*> and <*s2*> denote the input sequences; labels are illustrated in red.

Type	Template #	Example
Minimal (single sentence)	1	<*sentence*> <*label*>
	2	<*sentence*> Label: <*label*>
	3	Sentence: *sentence*> Label: <*label*>
Minimal (multiple sentences)	1	<*sentence1*> [SEP] <*sentence2*> <*label*>
	2	<*sentence1*> [SEP] <*sentence2*> Label: <*label*>
	3	Sentence 1: <*sentence1*> Sentence 2: <*sentence2*> Label: <*label*>

Table 3: Minimal prompts used for ABSTRACT.

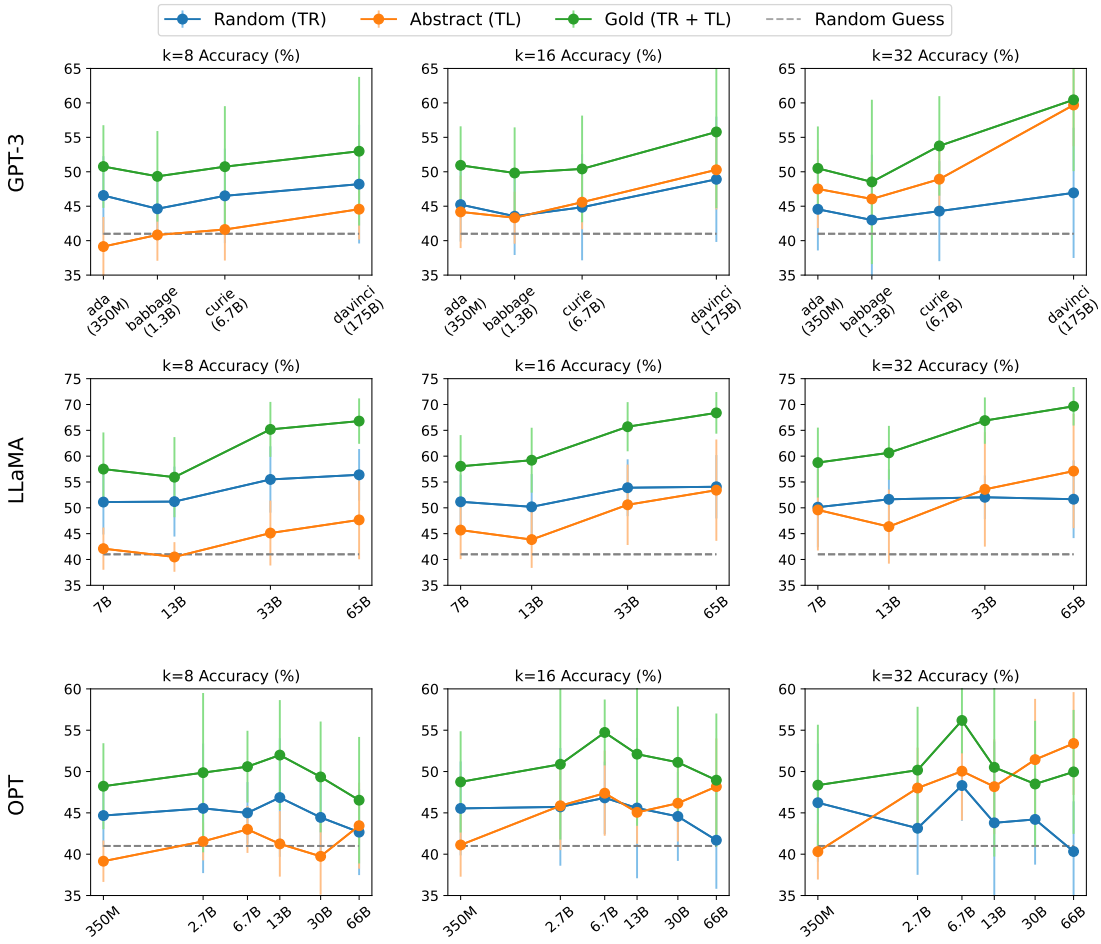

Figure 5: Averaged accuracy across 16 datasets for GPT-3 (top), LLaMA (middle), and OPT (bottom). X-axis shows model sizes from small to large. We run experiments with 8 (left), 16 (middle), and 32 (right) demonstrations respectively. Variance is calculated across three prompts.

	tweet_eval_hate			**tweet_eval_atheism**			**tweet_eval_feminist**			**sick**		
	Random	Abstract	Gold	Random	Abstract	Gold	Random	Abstract	Gold	Random	Abstract	Gold
ada	0.52	0.51	0.54	0.45	0.23	0.4	0.4	0.38	0.41	0.44	0.34	0.43
babbage	0.51	0.52	0.54	0.38	0.37	0.43	0.46	0.29	0.49	0.53	0.34	0.57
curie	0.55	0.54	0.6	0.28	0.33	0.32	0.39	0.32	0.4	0.56	0.36	0.56
davinci	0.56	0.55	0.59	0.34	0.33	0.33	0.4	0.4	0.38	0.4	0.39	0.44
	financial_phrasebank			**ethos_race**			**ethos_gender**			**ethos_religion**		
	Random	Abstract	Gold	Random	Abstract	Gold	Random	Abstract	Gold	Random	Abstract	Gold
ada	0.23	0.4	0.4	0.38	0.41	0.44	0.34	0.43	0.56	0.39	0.64	0.62
babbage	0.37	0.43	0.46	0.29	0.49	0.53	0.34	0.57	0.45	0.39	0.55	0.54
curie	0.33	0.32	0.39	0.32	0.4	0.56	0.36	0.56	0.54	0.42	0.63	0.53
davinci	0.33	0.33	0.4	0.4	0.38	0.4	0.39	0.44	0.4	0.56	0.44	0.52
	ethos_national_origin			**snli**			**sst2**			**trec**		
	Random	Abstract	Gold	Random	Abstract	Gold	Random	Abstract	Gold	Random	Abstract	Gold
ada	0.41	0.44	0.34	0.43	0.56	0.39	0.64	0.62	0.52	0.71	0.62	0.57
babbage	0.49	0.53	0.34	0.57	0.45	0.39	0.55	0.54	0.51	0.52	0.58	0.56
curie	0.4	0.56	0.36	0.56	0.54	0.42	0.63	0.53	0.52	0.6	0.48	0.54
davinci	0.38	0.4	0.39	0.44	0.4	0.56	0.44	0.52	0.51	0.54	0.47	0.52
	rte			**wnli**			**mrpc**			**poem**		
	Random	Abstract	Gold	Random	Abstract	Gold	Random	Abstract	Gold	Random	Abstract	Gold
ada	0.56	0.39	0.64	0.62	0.52	0.71	0.62	0.57	0.76	0.68	0.54	0.74
babbage	0.45	0.39	0.55	0.54	0.51	0.52	0.58	0.56	0.62	0.63	0.51	0.61
curie	0.54	0.42	0.63	0.53	0.52	0.6	0.48	0.54	0.55	0.56	0.6	0.54
davinci	0.4	0.56	0.44	0.52	0.51	0.54	0.47	0.52	0.5	0.48	0.53	0.62

Table 4: Single dataset accuracies across the GPT-3 model family, using 8 examples.

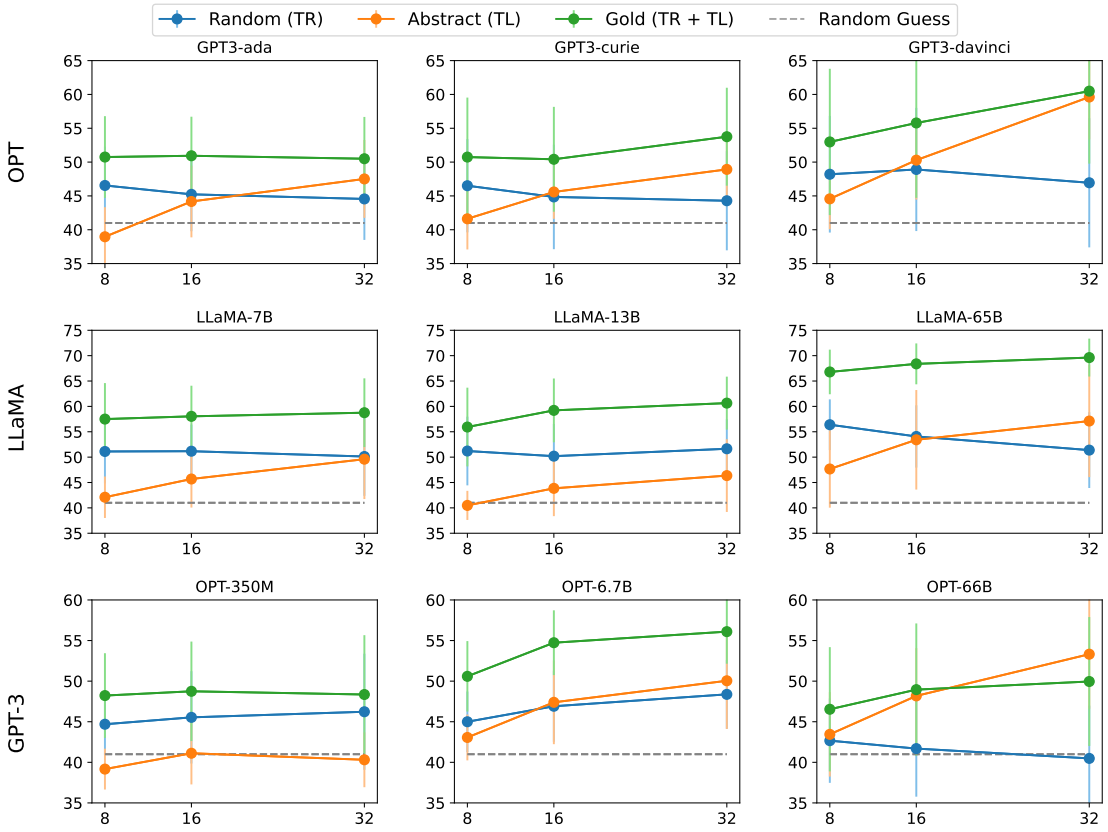

Figure 6: Averaged accuracy across 16 datasets for GPT-3 (top), LLaMA (middle), and OPT (bottom). x-axis shows number of demonstrations in the prompt. For each model, we run experiments for RANDOM (left), ABSTRACT(middle), and GOLD (right) demonstrations. Variance is calculated across three templates.

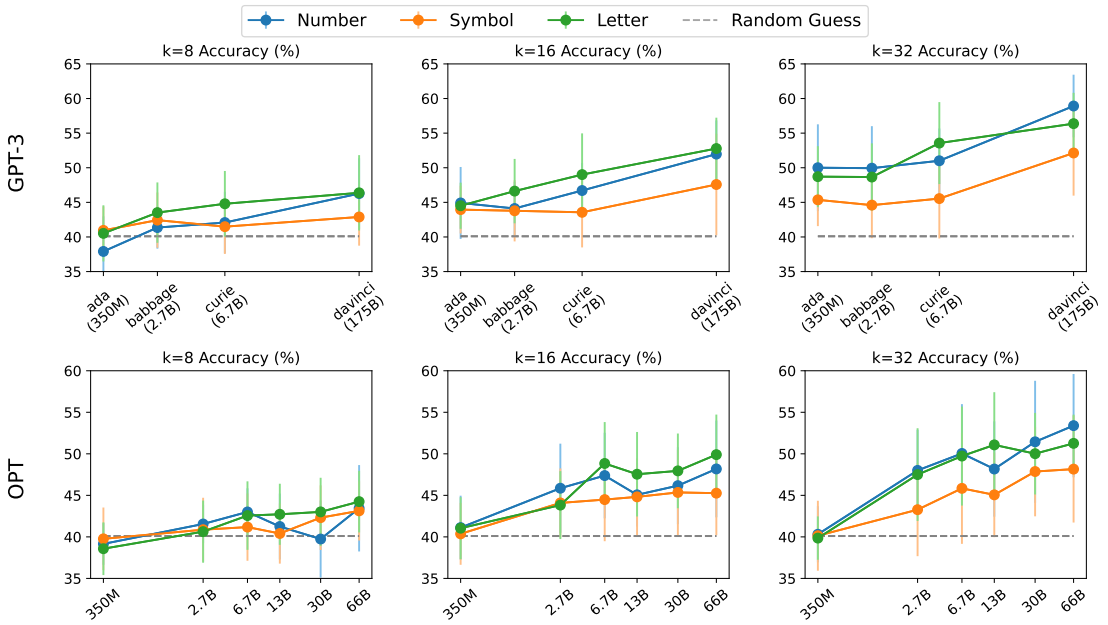

Figure 7: Performance of three types of ABSTRACT labels: numbers, letters, and symbols on OPT and GPT-3 models.

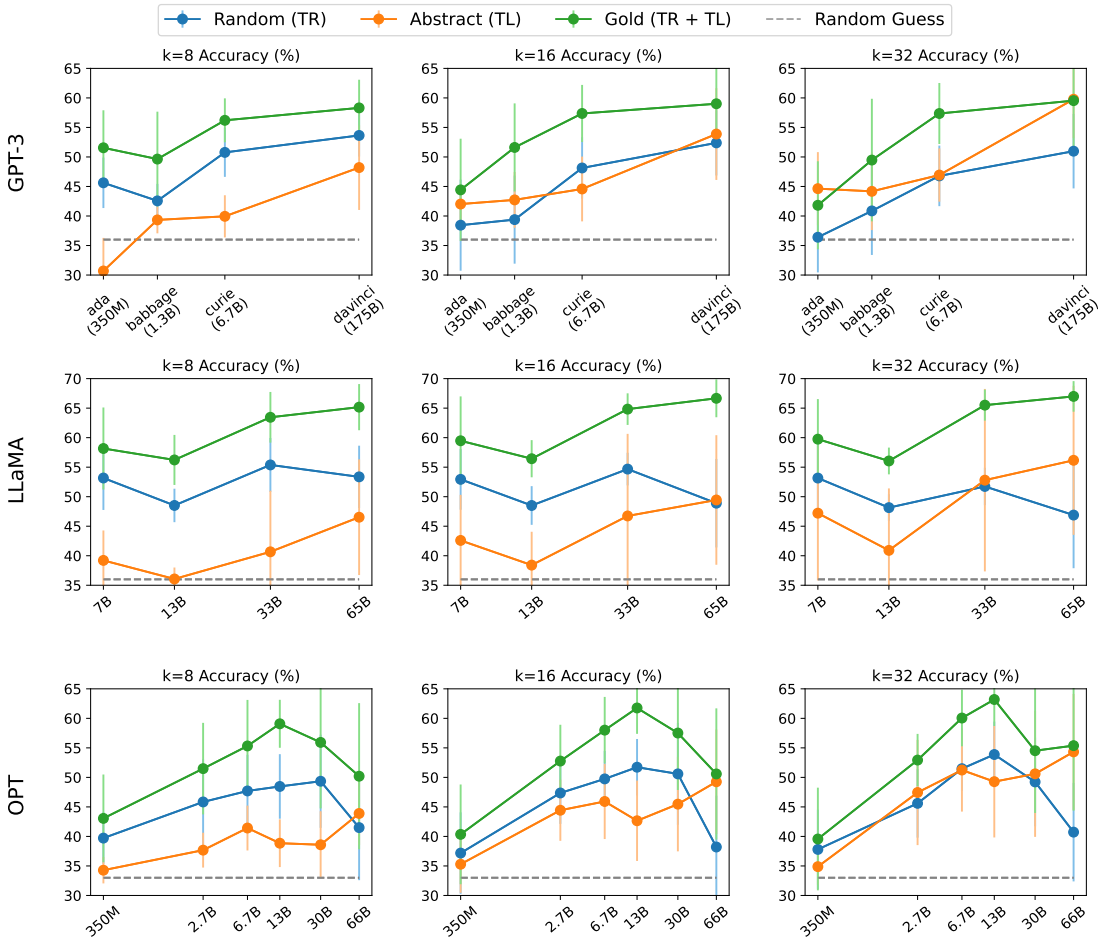

Figure 8: Average performance of **sentiment analysis** datasets.

	tweet_eval_hate			**tweet_eval_atheism**			**tweet_eval_feminist**			**sick**		
	Random	Abstract	Gold	Random	Abstract	Gold	Random	Abstract	Gold	Random	Abstract	Gold
ada	0.51	0.51	0.52	0.44	0.37	0.48	0.4	0.42	0.41	0.37	0.44	0.44
babbage	0.48	0.54	0.55	0.36	0.41	0.31	0.44	0.33	0.48	0.54	0.38	0.54
curie	0.54	0.58	0.62	0.28	0.48	0.3	0.33	0.38	0.32	0.56	0.41	0.56
davinci	0.56	0.6	0.64	0.34	0.42	0.39	0.29	0.44	0.38	0.46	0.49	0.49
	financial_phrasebank			**ethos_race**			**ethos_gender**			**ethos_religion**		
	Random	Abstract	Gold	Random	Abstract	Gold	Random	Abstract	Gold	Random	Abstract	Gold
ada	0.37	0.48	0.4	0.42	0.41	0.37	0.44	0.44	0.54	0.53	0.67	0.68
babbage	0.41	0.31	0.44	0.33	0.48	0.54	0.38	0.54	0.43	0.53	0.63	0.56
curie	0.48	0.3	0.33	0.38	0.32	0.56	0.41	0.56	0.5	0.55	0.71	0.54
davinci	0.42	0.39	0.29	0.44	0.38	0.46	0.49	0.49	0.38	0.63	0.49	0.51
	ethos_national_origin			**snli**			**sst2**			**trec**		
	Random	Abstract	Gold	Random	Abstract	Gold	Random	Abstract	Gold	Random	Abstract	Gold
ada	0.41	0.37	0.44	0.44	0.54	0.53	0.67	0.68	0.52	0.76	0.68	0.52
babbage	0.48	0.54	0.38	0.54	0.43	0.53	0.63	0.56	0.53	0.61	0.58	0.54
curie	0.32	0.56	0.41	0.56	0.5	0.55	0.71	0.54	0.56	0.55	0.49	0.55
davinci	0.38	0.46	0.49	0.49	0.38	0.63	0.49	0.51	0.6	0.56	0.5	0.57
	rte			**wnli**			**mrpc**			**poem**		
	Random	Abstract	Gold	Random	Abstract	Gold	Random	Abstract	Gold	Random	Abstract	Gold
ada	0.54	0.53	0.67	0.68	0.52	0.76	0.68	0.52	0.75	0.69	0.57	0.77
babbage	0.43	0.53	0.63	0.56	0.53	0.61	0.58	0.54	0.61	0.55	0.54	0.62
curie	0.5	0.55	0.71	0.54	0.56	0.55	0.49	0.55	0.59	0.49	0.55	0.59
davinci	0.38	0.63	0.49	0.51	0.6	0.56	0.5	0.57	0.59	0.54	0.67	0.63

Table 5: Single dataset accuracies across the GPT-3 model family, using 16 examples.

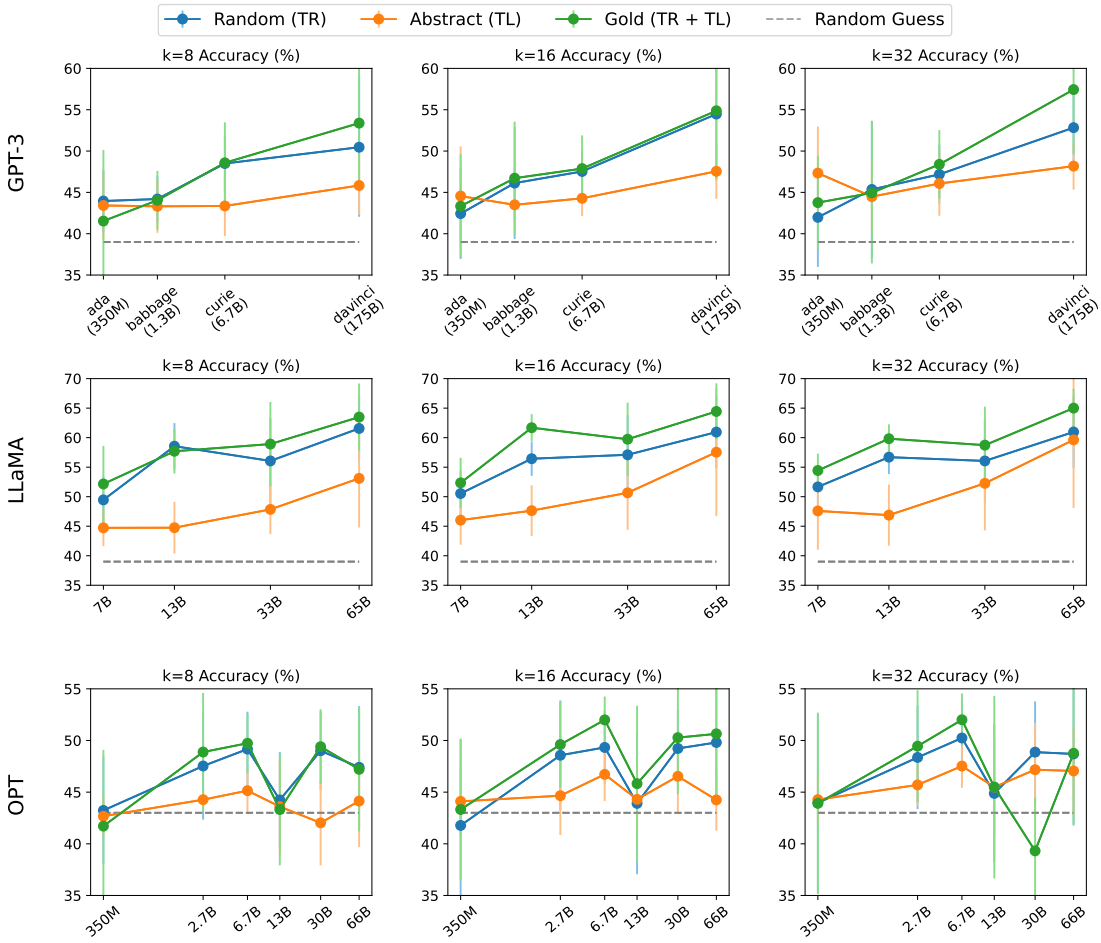

Figure 9: Average performance of **natural language inference/paraphrase detection** datasets.

	tweet_eval_hate			**tweet_eval_atheism**			**tweet_eval_feminist**			**sick**		
	Random	Abstract	Gold	Random	Abstract	Gold	Random	Abstract	Gold	Random	Abstract	Gold
ada	0.48	0.52	0.53	0.4	0.37	0.42	0.41	0.38	0.42	0.24	0.45	0.27
babbage	0.53	0.58	0.52	0.32	0.38	0.35	0.42	0.35	0.38	0.44	0.4	0.5
curie	0.54	0.59	0.66	0.26	0.47	0.31	0.38	0.4	0.43	0.57	0.41	0.57
davinci	0.57	0.64	0.66	0.29	0.51	0.37	0.28	0.49	0.37	0.43	0.52	0.49
	financial_phrasebank			**ethos_race**			**ethos_gender**			**ethos_religion**		
	Random	Abstract	Gold	Random	Abstract	Gold	Random	Abstract	Gold	Random	Abstract	Gold
ada	0.37	0.42	0.41	0.38	0.42	0.24	0.45	0.27	0.55	0.56	0.69	0.66
babbage	0.38	0.35	0.42	0.35	0.38	0.44	0.4	0.5	0.51	0.58	0.65	0.51
curie	0.47	0.31	0.38	0.4	0.43	0.57	0.41	0.57	0.52	0.56	0.71	0.51
davinci	0.51	0.37	0.28	0.49	0.37	0.43	0.52	0.49	0.35	0.68	0.5	0.63
	ethos_national_origin			**snli**			**sst2**			**trec**		
	Random	Abstract	Gold	Random	Abstract	Gold	Random	Abstract	Gold	Random	Abstract	Gold
ada	0.42	0.24	0.45	0.27	0.55	0.56	0.69	0.66	0.55	0.73	0.69	0.61
babbage	0.38	0.44	0.4	0.5	0.51	0.58	0.65	0.51	0.57	0.63	0.6	0.59
curie	0.43	0.57	0.41	0.57	0.52	0.56	0.71	0.51	0.59	0.65	0.5	0.61
davinci	0.37	0.43	0.52	0.49	0.35	0.68	0.5	0.63	0.6	0.63	0.51	0.62
	rte			**wnli**			**mrpc**			**poem**		
	Random	Abstract	Gold	Random	Abstract	Gold	Random	Abstract	Gold	Random	Abstract	Gold
ada	0.55	0.56	0.69	0.66	0.55	0.73	0.69	0.61	0.73	0.65	0.63	0.77
babbage	0.51	0.58	0.65	0.51	0.57	0.63	0.6	0.59	0.64	0.57	0.56	0.65
curie	0.52	0.56	0.71	0.51	0.59	0.65	0.5	0.61	0.63	0.44	0.61	0.69
davinci	0.35	0.68	0.5	0.63	0.6	0.63	0.51	0.62	0.65	0.6	0.7	0.71

Table 6: Single dataset accuracies across the GPT-3 model family, using 32 examples.

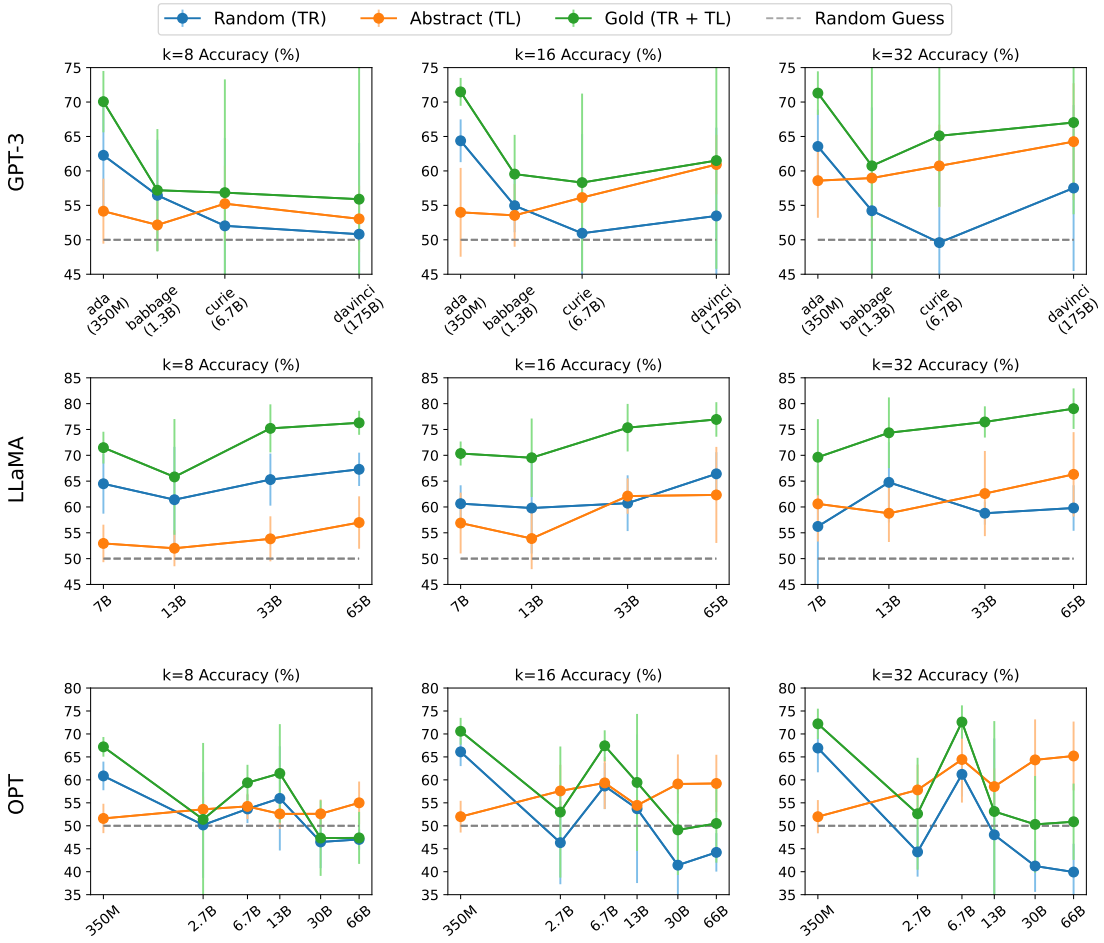

Figure 10: Average performance of **toxicity detection** datasets.

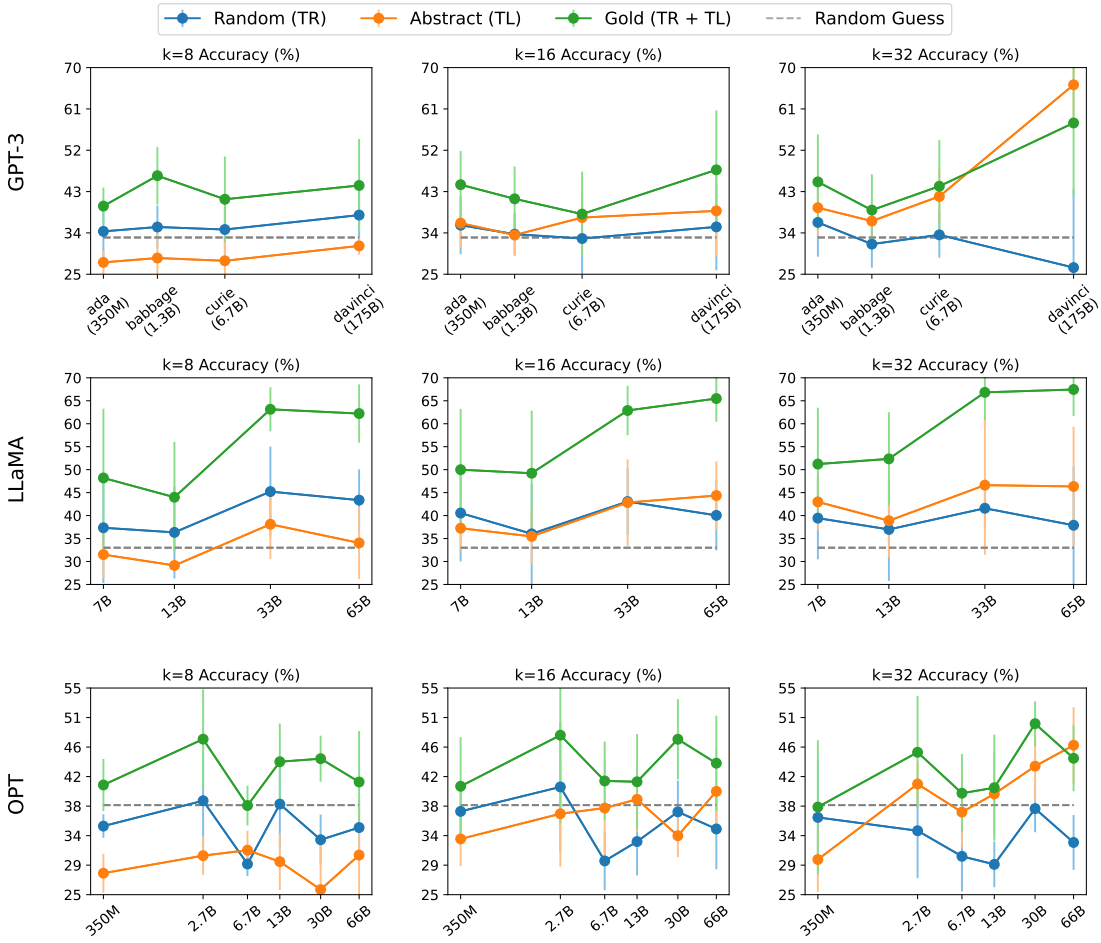

Figure 11: Average performance of **topic/stance classification** datasets.

	tweet_eval_hate			tweet_eval_atheism			tweet_eval_feminist			sick		
	Random	Abstract	Gold	Random	Abstract	Gold	Random	Abstract	Gold	Random	Abstract	Gold
OPT-350M	0.49	0.51	0.53	0.43	0.34	0.48	0.41	0.31	0.45	0.33	0.34	0.29
OPT-2.7B	0.52	0.55	0.56	0.43	0.36	0.45	0.47	0.34	0.5	0.52	0.34	0.55
OPT-6.7B	0.53	0.53	0.57	0.26	0.33	0.27	0.33	0.39	0.36	0.46	0.36	0.48
OPT-13B	0.55	0.52	0.61	0.4	0.35	0.4	0.49	0.35	0.47	0.36	0.3	0.37
OPT-30B	0.52	0.54	0.55	0.28	0.24	0.35	0.4	0.34	0.46	0.53	0.31	0.55
OPT-66B	0.52	0.55	0.53	0.29	0.38	0.32	0.44	0.37	0.42	0.44	0.36	0.47
	financial_phrasebank			ethos_race			ethos_gender			ethos_religion		
	Random	Abstract	Gold	Random	Abstract	Gold	Random	Abstract	Gold	Random	Abstract	Gold
OPT-350M	0.34	0.48	0.41	0.31	0.45	0.33	0.34	0.29	0.48	0.36	0.48	0.6
OPT-2.7B	0.36	0.45	0.47	0.34	0.5	0.52	0.34	0.55	0.54	0.42	0.56	0.49
OPT-6.7B	0.33	0.27	0.33	0.39	0.36	0.46	0.36	0.48	0.63	0.44	0.74	0.55
OPT-13B	0.35	0.4	0.49	0.35	0.47	0.36	0.3	0.37	0.59	0.44	0.69	0.62
OPT-30B	0.24	0.35	0.4	0.34	0.46	0.53	0.31	0.55	0.56	0.43	0.61	0.44
OPT-66B	0.38	0.32	0.44	0.37	0.42	0.44	0.36	0.47	0.33	0.44	0.46	0.45
	ethos_national_origin			snli			sst2			trec		
	Random	Abstract	Gold	Random	Abstract	Gold	Random	Abstract	Gold	Random	Abstract	Gold
OPT-350M	0.45	0.33	0.34	0.29	0.48	0.36	0.48	0.6	0.49	0.66	0.65	0.51
OPT-2.7B	0.5	0.52	0.34	0.55	0.54	0.42	0.56	0.49	0.53	0.49	0.51	0.56
OPT-6.7B	0.36	0.46	0.36	0.48	0.63	0.44	0.74	0.55	0.54	0.59	0.53	0.57
OPT-13B	0.47	0.36	0.3	0.37	0.59	0.44	0.69	0.62	0.53	0.63	0.55	0.53
OPT-30B	0.46	0.53	0.31	0.55	0.56	0.43	0.61	0.44	0.49	0.42	0.46	0.57
OPT-66B	0.42	0.44	0.36	0.47	0.33	0.44	0.46	0.45	0.55	0.53	0.44	0.55
	rte			wnli			mrpc			poem		
	Random	Abstract	Gold	Random	Abstract	Gold	Random	Abstract	Gold	Random	Abstract	Gold
OPT-350M	0.48	0.36	0.48	0.6	0.49	0.66	0.65	0.51	0.71	0.66	0.53	0.73
OPT-2.7B	0.54	0.42	0.56	0.49	0.53	0.49	0.51	0.56	0.52	0.48	0.51	0.5
OPT-6.7B	0.63	0.44	0.74	0.55	0.54	0.59	0.53	0.57	0.61	0.53	0.52	0.62
OPT-13B	0.59	0.44	0.69	0.62	0.53	0.63	0.55	0.53	0.61	0.55	0.52	0.62
OPT-30B	0.56	0.43	0.61	0.44	0.49	0.42	0.46	0.57	0.47	0.46	0.51	0.46
OPT-66B	0.33	0.44	0.46	0.45	0.55	0.53	0.44	0.55	0.38	0.49	0.55	0.56

Table 7: Single dataset accuracies across the OPT model family, using 8 examples.

	tweet_eval_hate			tweet_eval_atheism			tweet_eval_feminist			sick		
	Random	Abstract	Gold	Random	Abstract	Gold	Random	Abstract	Gold	Random	Abstract	Gold
OPT-350M	0.52	0.53	0.55	0.47	0.37	0.49	0.42	0.42	0.44	0.33	0.36	0.35
OPT-2.7B	0.52	0.56	0.58	0.44	0.44	0.47	0.51	0.39	0.46	0.55	0.39	0.57
OPT-6.7B	0.52	0.57	0.57	0.22	0.39	0.28	0.39	0.43	0.41	0.48	0.42	0.54
OPT-13B	0.58	0.54	0.62	0.32	0.44	0.38	0.41	0.39	0.41	0.36	0.4	0.36
OPT-30B	0.51	0.57	0.57	0.34	0.4	0.35	0.41	0.32	0.5	0.55	0.45	0.56
OPT-66B	0.5	0.57	0.54	0.25	0.47	0.31	0.47	0.44	0.48	0.49	0.38	0.51
	financial_phrasebank			ethos_race			ethos_gender			ethos_religion		
	Random	Abstract	Gold	Random	Abstract	Gold	Random	Abstract	Gold	Random	Abstract	Gold
OPT-350M	0.37	0.49	0.42	0.42	0.44	0.33	0.36	0.35	0.45	0.4	0.47	0.65
OPT-2.7B	0.44	0.47	0.51	0.39	0.46	0.55	0.39	0.57	0.53	0.5	0.58	0.45
OPT-6.7B	0.39	0.28	0.39	0.43	0.41	0.48	0.42	0.54	0.66	0.53	0.8	0.59
OPT-13B	0.44	0.38	0.41	0.39	0.41	0.36	0.4	0.36	0.6	0.53	0.72	0.54
OPT-30B	0.4	0.35	0.41	0.32	0.5	0.55	0.45	0.56	0.56	0.52	0.64	0.35
OPT-66B	0.47	0.31	0.47	0.44	0.48	0.49	0.38	0.51	0.3	0.57	0.49	0.44
	ethos_national_origin			snli			sst2			trec		
	Random	Abstract	Gold	Random	Abstract	Gold	Random	Abstract	Gold	Random	Abstract	Gold
OPT-350M	0.44	0.33	0.36	0.35	0.45	0.4	0.47	0.65	0.52	0.71	0.71	0.52
OPT-2.7B	0.46	0.55	0.39	0.57	0.53	0.5	0.58	0.45	0.59	0.47	0.41	0.62
OPT-6.7B	0.41	0.48	0.42	0.54	0.66	0.53	0.8	0.59	0.56	0.71	0.62	0.61
OPT-13B	0.41	0.36	0.4	0.36	0.6	0.53	0.72	0.54	0.53	0.62	0.5	0.55
OPT-30B	0.5	0.55	0.45	0.56	0.56	0.52	0.64	0.35	0.57	0.43	0.38	0.63
OPT-66B	0.48	0.49	0.38	0.51	0.3	0.57	0.49	0.44	0.59	0.51	0.4	0.6
	rte			wnli			mrpc			poem		
	Random	Abstract	Gold	Random	Abstract	Gold	Random	Abstract	Gold	Random	Abstract	Gold
OPT-350M	0.45	0.4	0.47	0.65	0.52	0.71	0.71	0.52	0.76	0.73	0.51	0.76
OPT-2.7B	0.53	0.5	0.58	0.45	0.59	0.47	0.41	0.62	0.52	0.45	0.54	0.54
OPT-6.7B	0.66	0.53	0.8	0.59	0.56	0.71	0.62	0.61	0.69	0.64	0.61	0.74
OPT-13B	0.6	0.53	0.72	0.54	0.53	0.62	0.5	0.55	0.58	0.55	0.53	0.58
OPT-30B	0.56	0.52	0.64	0.35	0.57	0.43	0.38	0.63	0.5	0.41	0.59	0.51
OPT-66B	0.3	0.57	0.49	0.44	0.59	0.51	0.4	0.6	0.46	0.46	0.59	0.55

Table 8: Single dataset accuracies across the OPT model family, using 16 examples.

	tweet_eval_hate			tweet_eval_atheism			tweet_eval_feminist			sick		
	Random	Abstract	Gold	Random	Abstract	Gold	Random	Abstract	Gold	Random	Abstract	Gold
OPT-350M	0.53	0.53	0.55	0.42	0.35	0.42	0.43	0.33	0.4	0.36	0.34	0.35
OPT-2.7B	0.51	0.59	0.59	0.31	0.42	0.42	0.43	0.39	0.42	0.53	0.4	0.57
OPT-6.7B	0.55	0.59	0.6	0.26	0.29	0.24	0.4	0.39	0.42	0.49	0.44	0.53
OPT-13B	0.56	0.58	0.59	0.25	0.45	0.36	0.39	0.38	0.42	0.4	0.38	0.37
OPT-30B	0.52	0.59	0.57	0.32	0.47	0.42	0.47	0.42	0.47	0.54	0.45	0.6
OPT-66B	0.48	0.58	0.51	0.27	0.5	0.26	0.4	0.46	0.5	0.45	0.43	0.47
	financial_phrasebank			ethos_race			ethos_gender			ethos_religion		
	Random	Abstract	Gold	Random	Abstract	Gold	Random	Abstract	Gold	Random	Abstract	Gold
OPT-350M	0.35	0.42	0.43	0.33	0.4	0.36	0.34	0.35	0.44	0.38	0.44	0.67
OPT-2.7B	0.42	0.42	0.43	0.39	0.42	0.53	0.4	0.57	0.51	0.56	0.58	0.46
OPT-6.7B	0.29	0.24	0.4	0.39	0.42	0.49	0.44	0.53	0.68	0.61	0.82	0.63
OPT-13B	0.45	0.36	0.39	0.38	0.42	0.4	0.38	0.37	0.61	0.6	0.72	0.48
OPT-30B	0.47	0.42	0.47	0.42	0.47	0.54	0.45	0.6	0.57	0.57	0.7	0.4
OPT-66B	0.5	0.26	0.4	0.46	0.5	0.45	0.43	0.47	0.37	0.64	0.57	0.41
	ethos_national_origin			snli			sst2			trec		
	Random	Abstract	Gold	Random	Abstract	Gold	Random	Abstract	Gold	Random	Abstract	Gold
OPT-350M	0.4	0.36	0.34	0.35	0.44	0.38	0.44	0.67	0.51	0.73	0.71	0.51
OPT-2.7B	0.42	0.53	0.4	0.57	0.51	0.56	0.58	0.46	0.55	0.49	0.43	0.6
OPT-6.7B	0.42	0.49	0.44	0.53	0.68	0.61	0.82	0.63	0.65	0.74	0.62	0.65
OPT-13B	0.42	0.4	0.38	0.37	0.61	0.6	0.72	0.48	0.56	0.57	0.44	0.64
OPT-30B	0.47	0.54	0.45	0.6	0.57	0.57	0.7	0.4	0.55	0.42	0.36	0.66
OPT-66B	0.5	0.45	0.43	0.47	0.37	0.64	0.57	0.41	0.63	0.52	0.36	0.67
	rte			wnli			mrpc			poem		
	Random	Abstract	Gold	Random	Abstract	Gold	Random	Abstract	Gold	Random	Abstract	Gold
OPT-350M	0.44	0.38	0.44	0.67	0.51	0.73	0.71	0.51	0.77	0.74	0.52	0.79
OPT-2.7B	0.51	0.56	0.58	0.46	0.55	0.49	0.43	0.6	0.54	0.41	0.56	0.48
OPT-6.7B	0.68	0.61	0.82	0.63	0.65	0.74	0.62	0.65	0.78	0.65	0.64	0.77
OPT-13B	0.61	0.6	0.72	0.48	0.56	0.57	0.44	0.64	0.5	0.45	0.53	0.5
OPT-30B	0.57	0.57	0.7	0.4	0.55	0.42	0.36	0.66	0.46	0.4	0.71	0.54
OPT-66B	0.37	0.64	0.57	0.41	0.63	0.52	0.36	0.67	0.49	0.4	0.69	0.56

Table 9: Single dataset accuracies across the OPT model family, using 32 examples.

	tweet_eval_hate			tweet_eval_atheism			tweet_eval_feminist			sick		
	Random	Abstract	Gold	Random	Abstract	Gold	Random	Abstract	Gold	Random	Abstract	Gold
7B	0.59	0.53	0.64	0.33	0.31	0.37	0.41	0.43	0.45	0.32	0.36	0.38
13B	0.63	0.53	0.65	0.31	0.34	0.28	0.43	0.34	0.44	0.39	0.41	0.41
30B	0.64	0.58	0.72	0.38	0.47	0.52	0.57	0.49	0.65	0.37	0.43	0.41
65B	0.69	0.58	0.72	0.4	0.42	0.58	0.54	0.42	0.58	0.38	0.46	0.41
	financial_phrasebank			ethos_race			ethos_gender			ethos_religion		
	Random	Abstract	Gold	Random	Abstract	Gold	Random	Abstract	Gold	Random	Abstract	Gold
7B	0.31	0.37	0.41	0.43	0.45	0.32	0.36	0.38	0.64	0.4	0.7	0.65
13B	0.34	0.28	0.43	0.34	0.44	0.39	0.41	0.41	0.42	0.35	0.61	0.61
30B	0.47	0.52	0.57	0.49	0.65	0.37	0.43	0.41	0.65	0.38	0.79	0.69
65B	0.42	0.58	0.54	0.42	0.58	0.38	0.46	0.41	0.6	0.44	0.83	0.69
	ethos_national_origin			snli			sst2			trec		
	Random	Abstract	Gold	Random	Abstract	Gold	Random	Abstract	Gold	Random	Abstract	Gold
7B	0.45	0.32	0.36	0.38	0.64	0.4	0.7	0.65	0.56	0.73	0.61	0.53
13B	0.44	0.39	0.41	0.41	0.42	0.35	0.61	0.61	0.52	0.66	0.59	0.5
30B	0.65	0.37	0.43	0.41	0.65	0.38	0.79	0.69	0.52	0.76	0.65	0.52
65B	0.58	0.38	0.46	0.41	0.6	0.44	0.83	0.69	0.55	0.75	0.65	0.56
	rte			wnli			mrpc			poem		
	Random	Abstract	Gold	Random	Abstract	Gold	Random	Abstract	Gold	Random	Abstract	Gold
7B	0.64	0.4	0.7	0.65	0.56	0.73	0.61	0.53	0.7	0.71	0.52	0.78
13B	0.42	0.35	0.61	0.61	0.52	0.66	0.59	0.5	0.64	0.71	0.54	0.78
30B	0.65	0.38	0.79	0.69	0.52	0.76	0.65	0.52	0.77	0.67	0.56	0.86
65B	0.6	0.44	0.83	0.69	0.55	0.75	0.65	0.56	0.77	0.73	0.6	0.87

Table 10: Single dataset accuracies across the LLaMA model family, using 8 examples.

	tweet_eval_hate			tweet_eval_atheism			tweet_eval_feminist			sick		
	Random	Abstract	Gold	Random	Abstract	Gold	Random	Abstract	Gold	Random	Abstract	Gold
7B	0.61	0.58	0.66	0.33	0.49	0.37	0.41	0.35	0.45	0.31	0.43	0.36
13B	0.6	0.58	0.66	0.27	0.5	0.34	0.4	0.34	0.42	0.37	0.42	0.41
30B	0.67	0.67	0.74	0.37	0.54	0.53	0.47	0.5	0.62	0.36	0.51	0.42
65B	0.66	0.62	0.73	0.37	0.56	0.6	0.52	0.53	0.6	0.38	0.55	0.42
	financial_phrasebank			**ethos_race**			**ethos_gender**			**ethos_religion**		
	Random	Abstract	Gold	Random	Abstract	Gold	Random	Abstract	Gold	Random	Abstract	Gold
7B	0.49	0.37	0.41	0.35	0.45	0.31	0.43	0.36	0.65	0.46	0.72	0.6
13B	0.5	0.34	0.4	0.34	0.42	0.37	0.42	0.41	0.41	0.39	0.59	0.56
30B	0.54	0.53	0.47	0.5	0.62	0.36	0.51	0.42	0.64	0.49	0.84	0.6
65B	0.56	0.6	0.52	0.53	0.6	0.38	0.55	0.42	0.56	0.54	0.87	0.62
	ethos_national_origin			**snli**			**sst2**			**trec**		
	Random	Abstract	Gold	Random	Abstract	Gold	Random	Abstract	Gold	Random	Abstract	Gold
7B	0.45	0.31	0.43	0.36	0.65	0.46	0.72	0.6	0.53	0.72	0.57	0.59
13B	0.42	0.37	0.42	0.41	0.41	0.39	0.59	0.56	0.51	0.66	0.59	0.5
30B	0.62	0.36	0.51	0.42	0.64	0.49	0.84	0.6	0.58	0.74	0.6	0.65
65B	0.6	0.38	0.55	0.42	0.56	0.54	0.87	0.62	0.58	0.75	0.66	0.65
	rte			**wnli**			**mrpc**			**poem**		
	Random	Abstract	Gold	Random	Abstract	Gold	Random	Abstract	Gold	Random	Abstract	Gold
7B	0.65	0.46	0.72	0.6	0.53	0.72	0.57	0.59	0.67	0.65	0.59	0.78
13B	0.41	0.39	0.59	0.56	0.51	0.66	0.59	0.5	0.73	0.69	0.54	0.78
30B	0.64	0.49	0.84	0.6	0.58	0.74	0.6	0.65	0.74	0.65	0.64	0.85
65B	0.56	0.54	0.87	0.62	0.58	0.75	0.66	0.65	0.78	0.73	0.64	0.85

Table 11: Single dataset accuracies across the LLaMA model family, using 16 examples.

	tweet_eval_hate			tweet_eval_atheism			tweet_eval_feminist			sick		
	Random	Abstract	Gold	Random	Abstract	Gold	Random	Abstract	Gold	Random	Abstract	Gold
7B	0.58	0.58	0.64	0.33	0.51	0.35	0.4	0.38	0.47	0.36	0.46	0.4
13B	0.6	0.59	0.68	0.3	0.46	0.37	0.41	0.42	0.46	0.36	0.42	0.42
30B	0.65	0.64	0.73	0.32	0.53	0.6	0.48	0.51	0.63	0.35	0.55	0.42
65B	0.64	0.68	0.78	0.38	0.51	0.6	0.45	0.49	0.63	0.36	0.62	0.43
	financial_phrasebank			**ethos_race**			**ethos_gender**			**ethos_religion**		
	Random	Abstract	Gold	Random	Abstract	Gold	Random	Abstract	Gold	Random	Abstract	Gold
7B	0.51	0.35	0.4	0.38	0.47	0.36	0.46	0.4	0.64	0.5	0.74	0.61
13B	0.46	0.37	0.41	0.42	0.46	0.36	0.42	0.42	0.38	0.38	0.56	0.65
30B	0.53	0.6	0.48	0.51	0.63	0.35	0.55	0.42	0.61	0.61	0.88	0.66
65B	0.51	0.6	0.45	0.49	0.63	0.36	0.62	0.43	0.52	0.66	0.88	0.59
	ethos_national_origin			**snli**			**sst2**			**trec**		
	Random	Abstract	Gold	Random	Abstract	Gold	Random	Abstract	Gold	Random	Abstract	Gold
7B	0.47	0.36	0.46	0.4	0.64	0.5	0.74	0.61	0.59	0.67	0.47	0.62
13B	0.46	0.36	0.42	0.42	0.38	0.38	0.56	0.65	0.53	0.73	0.67	0.57
30B	0.63	0.35	0.55	0.42	0.61	0.61	0.88	0.66	0.6	0.74	0.55	0.6
65B	0.63	0.36	0.62	0.43	0.52	0.66	0.88	0.59	0.63	0.76	0.58	0.66
	rte			**wnli**			**mrpc**			**poem**		
	Random	Abstract	Gold	Random	Abstract	Gold	Random	Abstract	Gold	Random	Abstract	Gold
7B	0.64	0.5	0.74	0.61	0.59	0.67	0.47	0.62	0.69	0.65	0.64	0.79
13B	0.38	0.38	0.56	0.65	0.53	0.73	0.67	0.57	0.76	0.7	0.62	0.83
30B	0.61	0.61	0.88	0.66	0.6	0.74	0.55	0.6	0.8	0.57	0.65	0.82
65B	0.52	0.66	0.88	0.59	0.63	0.76	0.58	0.66	0.77	0.63	0.73	0.87

Table 12: Single dataset accuracies across the LLaMA model family, using 32 examples.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
5

☒ A2. Did you discuss any potential risks of your work?
Our investigation focuses on providing an empirical explanation of the behavior of ICL. We do not spot immediate concerns for risk but are happy to supplement as needed.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
1

☑ A4. Have you used AI writing assistants when working on this paper?
Some authors used Writefull for Overleaf, which is a grammar checker and phrase-suggestion tool. It was used to double-check spelling and phrasing for all sections of the paper.

B ☑ Did you use or create scientific artifacts?

3

☑ B1. Did you cite the creators of artifacts you used?
Appendix A

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
The datasets that we use are publicly available on Huggingface.

☒ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
We do not create any artifacts; we use publicly available datasets to study the behavior of LMs performing ICL. We do not spot immediate concerns for inappropriate use but are happy to supplement as needed.

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
We use publicly available datasets to study the behavior of LMs performing ICL and assume that the dataset creators have rigorously anonymized the sources of their data.

☒ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
We use publicly available English-based datasets across a wide spread of domains; dpcumentation is available at Huggingface and other public repositories.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
We describe how many demonstrations we use per prompt and how many examples we use to evaluate model performance. We sample at most 32 demonstrations per prompt and do not perform any fine-tuning; thus, we do not currently find details of train/test/dev splits to be critical, but are happy to supplement as needed.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

C ☑ Did you run computational experiments?

3

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
3

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
3

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
4

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
3

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Not applicable. Left blank.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Not applicable. Left blank.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.