Can In-context Learners Learn a Reasoning Concept from Demonstrations?

Michal Štefánik and **Marek Kadlčík**
Faculty of Informatics, Masaryk University, Czech Republic
{stefanik.m,kadlcik}@mail.muni.cz

Abstract

Large language models show an emergent ability to learn a new task from a small number of input-output demonstrations. However, recent work shows that in-context learners largely rely on their pre-trained knowledge, such as the sentiment of the labels, instead of finding new associations in the input. However, the commonly-used few-shot evaluation settings using a *random* selection of in-context demonstrations can not disentangle models' ability to *learn* a new skill from demonstrations, as most of the randomly-selected demonstrations do not present relations *informative* for prediction beyond exposing the new task distribution.

To disentangle models' in-context learning ability independent of models' memory, we introduce a *Conceptual few-shot learning* method selecting the demonstrations sharing a possibly-informative *concept* with the predicted sample. We extract a set of such concepts from annotated explanations and measure how much can models benefit from presenting these concepts in few-shot demonstrations.

We find that smaller models are more sensitive to the presented concepts. While some of the models are able to benefit from concept-presenting demonstrations for each assessed concept, we find that *none* of the assessed in-context learners can benefit from all presented reasoning concepts consistently, leaving the in-context concept learning an open challenge.

1 Introduction

In-context learning (ICL) is the alternative to the conventional training of Large Language Models (LLMs) for specific task(s), where models are expected to learn a new task solely from the input text. In few-shot in-context learning that we focus on, the input text contains a set of *demonstrations*, i.e. the input-output examples of the task to be learned (Brown et al., 2020).

An ability to learn unseen tasks from natural instructions has practical and theoretical implications,

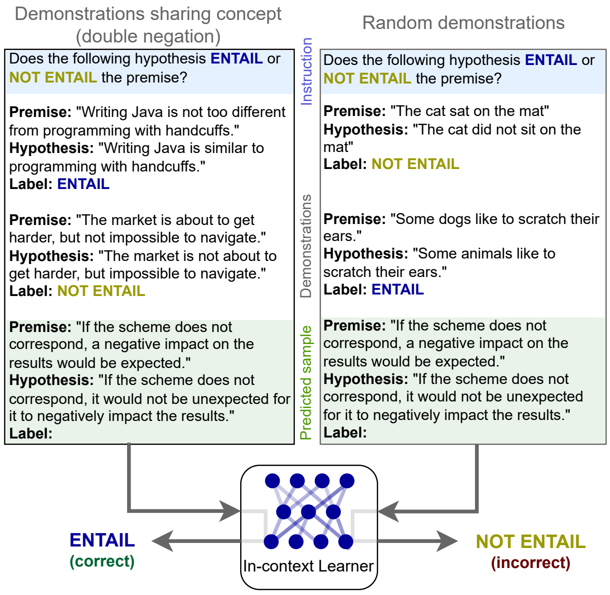

Figure 1: In this work, we assess In-context learners' ability to improve when presented with demonstrations using a reasoning concept applicable in the prediction (§2). We extract these concepts from human explanations (§3.2) and assess models' ability to learn to use these concepts, as reflected in improving their prediction quality.

both of which are of great significance; Understanding free-form user requests allow applying LLMs in applications of restricted, or limited data availability without over-specialization (Goodfellow et al., 2014). In-context learning can provide a *handle* of models' behaviour, enabling the model to avoid specific erroneous predictions. In theory, a training process resulting in accurate new-task learner defines the sufficient conditions for the emergence of a specific level of generalization.

Recent LLMs trained on vast mixtures of tasks (Sanh et al., 2022a; Wang et al., 2022b; Chung et al., 2022) show a certain level of new-task ICL and gradually bring more attention and expectations in this direction. However, counter-intuitively to the overall evaluations, in-context

Proceedings of the 1st Workshop on Natural Language Reasoning and Structured Explanations (@ACL 2023), pages 107–115
June 13, 2023 ©2023 Association for Computational Linguistics

learners (ICLs) also expose surprising behavioural artefacts; Liu et al. (2022) show ICLs' sensitivity to the ordering of in-context demonstrations. Similarly, Lu et al. (2022) find surprising sensitivity of ICLs to the specific wording of the prompts. Min et al. (2022b) show that most of the model performance is persisted even when the contents of the demonstrations are randomly swapped. Contrary to the ability to learn from input, Wei et al. (2023) propose to attribute this to the over-reliance of in-context learners on *semantics* of the label tokens, especially in smaller models.

We find that the discrepancy in the perceived abilities of ICLs might be attributed to their limited evaluation, commonly performed with a *random* set of task demonstrations. However, for many open-ended tasks, such as Question Answering, or Translation, randomly-chosen demonstrations rarely present a reasoning pattern which can help with the prediction of new input (Figure 1; Right). We argue that the evaluation with mostly non-informative contexts also can not reflect on the ability of *learning*, as observed in humans[1], as the gain of extrapolating associations presented in non-informative demonstrations can only bring little benefit to the practice.

We note that in the absolute numbers, the random-demonstrations evaluation also favours very large LLMs with a capacity to remember a wider variety of input distributions from pre-training; Conditioned by the capacity, very large LLMs can better modulate the behaviour based on demonstrations' distribution, instead of learning new association(s) from the context.

Hence, in Section 2, we propose to evaluate models' in-context learning ability primed with the demonstrations that exhibit a reasoning *analogical* to the one required for a robust prediction of the predicted sample (Fig. 1). We measure how well can the recent few-shot learners *utilize* identified concepts for more accurate predictions (§3) and find large discrepancies among the models and concepts.

Our main contributions are following: (i) We introduce a task of Conceptual Few-shot Learning, disentangling models' ability to learn a new reasoning concept from other aspects of prediction quality. We show how such reasoning concepts can be extracted from human explanations. (ii) For a wide variety of recent in-context learners, we measure the ability to benefit from presented reasoning concepts. We show that while some models are better at learning concepts on average, this ability can not be attributed to the models' size or training strategy.

[1] We restrain from discussing a concept of *learning* in the psychological scope, but we note that Concept learning fits well into a definition of Associative learning (Plotnik, 2012).

Problem Definition Given a dataset $\mathcal{D} : \{(x_1 \rightarrow Y_1), .., (x_i \rightarrow Y_i)\} \in \mathcal{D}$ containing pairs of *input* x_j with associated *label* Y_j, an *in-context few-shot learner* $\Theta(x) \rightarrow y$ aims to predict a correct label $y_{k+1} = Y_{k+1}$ given a sequence of k input-output *demonstrations*, and the *predicted input* x_{k+1}:

$$\Theta([x_1 \rightarrow Y_1, .., x_k \rightarrow Y_k], x_{k+1}) \rightarrow y_{k+1} \quad (1)$$

We expect *in-context few-shot learner* Θ to model the relation of x_i and y_i by (i) *identifying* and (ii) *extrapolating* the relations of input and output presented in demonstrations. Each such relation is modelled by one or more *latent concepts* $\mathcal{C}$:

$$\forall\, (x_i, Y_i) \in \mathcal{D} : \exists\, \mathcal{C} : \mathcal{C}(x_i, Y_i) = 1 \quad (2)$$

We broadly define a *concept* $\mathcal{C}$ as any function $\mathcal{C}(x, y) \rightarrow \{0, 1\}$, constraining a space of valid outputs y to the ones where $\mathcal{C}(x, y) = 1$. Thus, if Θ *learns* a concept $\mathcal{C}$, it will never predict for x such y that $\mathcal{C}(x, y) = 0$. In a composition $\{\mathcal{C}\} = \{\mathcal{C}_1, .., \mathcal{C}_j\}$, all $\mathcal{C}_i \in \{C\}$ must evaluate to 1.

Given that modelling of each C valid for the task of $\mathcal{D}$ restrain a set of possible predictions of Θ *exclusively* from incorrect predictions, extending a set of concepts learned in-context with complementary one(s) should *never* decrease the performance of the model Θ on $\mathcal{D}$.

2 Conceptual Few-shot Learning

We reformulate in-context few-shot learning (1) to a *conceptual few-shot learning*, evaluating the ability of a few-shot learner Θ to *identify* and *apply* a user-chosen reasoning concept $\mathcal{C}$ shown in demonstrations. First, we *classify* evaluation samples such that the samples of the same category X_i require the concept $\mathcal{C}_i$ to map x to Y. Subsequently, in conceptual few-shot learning, we let the learner to infer a prediction for input x_{k+1} by presenting it with demonstrations $(x_j \rightarrow Y_j)_{1..k} \in X_i$, thus *sharing* the reasoning concept $\mathcal{C}_i$ with the predicted input x_{k+1}:

$$\begin{gathered}\Theta([x_1 \rightarrow Y_1, .., x_k \rightarrow Y_k], x_{k+1}) \\ \text{where } \forall (x_{1..k}, Y_{1..k}) \in X^i \text{ and } x_{k+1} \in X^i\end{gathered} \quad (3)$$

We note that Θ can rely on other features than $\mathcal{C}_i$, and such reliance is not easy to disentangle. Therefore, we propose to contextualize the results of Conceptual few-shot learning on a concept C_i a *difference* to the performance obtained in a *random* selection of demonstrations.

Additionally, to make the predictions based on two different sets of demonstrations mutually comparable without systematic bias (e.g. in samples' complexity), we perform both random and concept-sharing evaluations with the same predicted samples x_{k+1}, and only change the demonstrations[2].

Informative Concepts Extraction Constructing a scaled evaluation with annotated reasoning concepts $\mathcal{C}$ is challenging since the annotations of such concepts in associated with the datasets are rare.

However, we find such reasoning inherently captured in human explanations of some datasets, where annotators are asked to collect answers to a question "why is [input] assigned [output]?" (Wiegreffe and Marasović, 2021).

The form of these explanations ranges from free-text explanations, including annotator-specific slang and stylistics, to semi-structured and structured explanations, cast to a pre-defined format, often consisting of a set of relations in a form "[subject1] [relation] [subject2]" that transitively maps [input] to [output] (Jansen et al., 2018). We focus on extracting the concepts from the subset of the semi-structured and structured explanations where the format consistency and non-ambiguity of the operands are reassured.

3 Evaluations

This section introduces few-shot learners that we evaluate for Conceptual few-shot learning and the datasets allowing us to extract reasoning concepts.

3.1 Few-shot Learners

T0 (Sanh et al., 2022b) introduce a set of in-context learning models fine-tuned from a T5 model (Raffel et al., 2020) on a variety of tasks in zero-shot settings, aiming to perform well on a task of previously-unseen categories. T0 is trained for seq2seq generation over a large set of diverse tasks cast to a unified input-output format provided by task-specific templates of Promptsource project (Bach et al., 2022).

[2]The implementation of Conceptual few-shot learning is available on `https://github.com/MIR-MU/CoAT`.

TK-INSTRUCT (Wang et al., 2022a) is a set of models trained for comprehension of annotator-like instructions, consisting of a free-text task description and a set of input-output pairs, collected for more than 1,400 tasks of NATURALINSTRUCTIONS collection (Mishra et al., 2022). Note that, in contrary to T0, TK-INSTRUCT models can advance from being trained in the few-shot learning format, where the model was exposed to the format of a few input-output examples already in the fine-tuning.

FLAN (Chung et al., 2022) scales the approach of fine-tuning in a few-shot learning format to over 1,800 tasks of 146 categories including all resources of T0 and TK-INSTRUCT. Contrary to the former models, the training data mixture includes several datasets with chain-of-thought labels, where the model is trained to follow the annotated reasoning chain explicitly. We evaluate all publicly available T5-based FLAN models.

GPT3 (Brown et al., 2020) is a well-known causal language model that has first shown that in-context few-shot learning ability can emerge solely from vast amounts of unsupervised training data and parametrization, without fine-tuning. Alternatively to other approaches, **INSTRUCTGPT** (Ouyang et al., 2022) fine-tunes GPT3 to follow human instructions using obtained user feedback. We evaluate both these models through OpenAI APIs[3].

3.2 Datasets

Following is a description of datasets that we use in Conceptual few-shot evaluation. Note that for each dataset, we highlight a single concept that we use in Conceptual few-shot evaluation as the $\mathcal{C}$ (§2). In the case of each model and dataset, we first evaluate all templates available in Promptsource and report the gain of utilising the chosen concept for the best-performing template.

WorldTree (Jansen et al., 2018; Xie et al., 2020) is a collection of 5,114 science exam questions with the explanations in the form of 9,216 shared facts supporting the assignment of the correct answer.

We use the shared facts as the concepts $\mathcal{C}$ and evaluate with the demonstrations of a maximal facts' intersection with the predicted sample. Contrary to the other datasets, in WorldTree evaluation,

[3]`https://beta.openai.com`

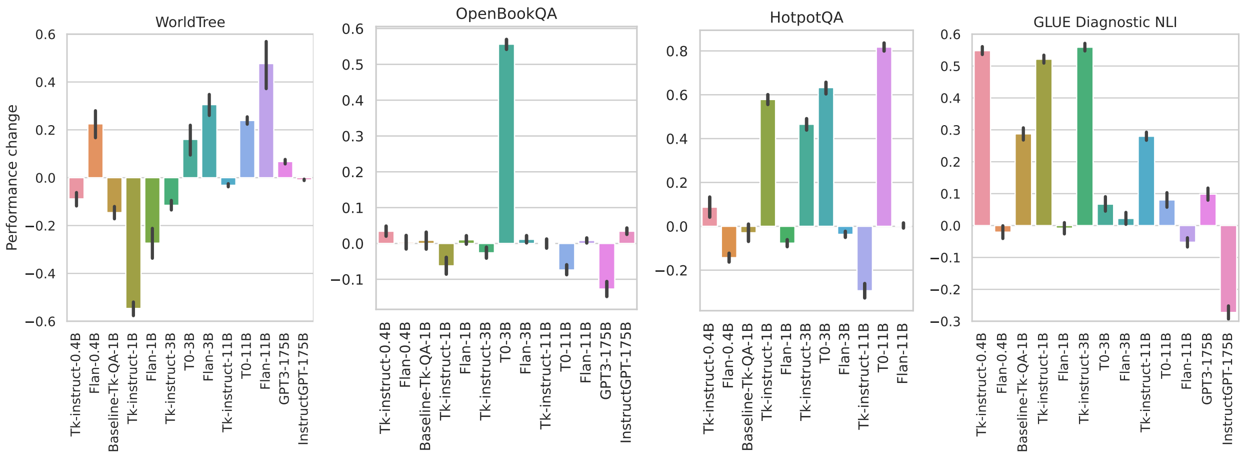

Figure 2: **Conceptual few-shot evaluation:** Relative performance change of the assessed in-context learners between using *random* demonstrations (k=3) and *concept-sharing* demonstrations (§2), with concepts of the datasets described in §3.2. Models are ordered by a number of parameters. Error bars show a 95% confidence interval of the bootstrapped results (100 samples, 200 repeats). Absolute results for both selection strategies are in Figure 4.

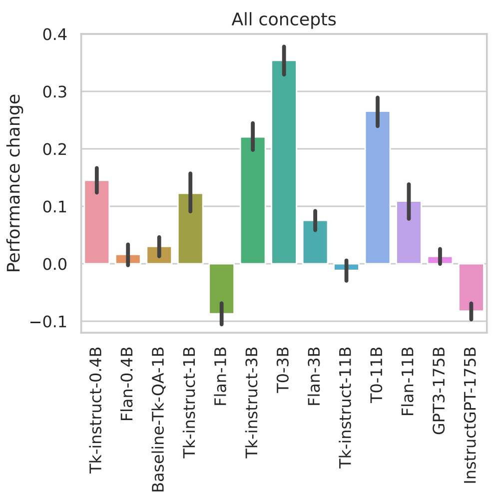

Figure 3: **Conceptual few-shot evaluation: all concepts:** Error change of the assessed in-context learners between random demonstrations and concept-sharing demonstrations (§2) aggregated over all assessed concepts. Experimental setup is consistent with Figure 2.

we prepend the facts for all the demonstrations in the context before the demonstrations.

OpenBookQA (Mihaylov et al., 2018) is a collection of elementary-grade single-choice questions requiring common sense knowledge about the world. A set of 4,957 explanations take the form of a triple of *(object, relation, object)*, such as "a stove generates heat" for a question "Which one of these can help a person cook their food? [four options]" and a correct option "a counter cooker appliance".

To extract informative concepts $\mathcal{C}$, we perform syntactic analysis of the explanation and extract the *relation*, identified as a *root* of the sentence's parse tree. Hence, in conceptual few-shot learning, we prime the aforementioned question with other question-options-answer pairs of the questions answerable by relating the input to output through the "generate" relation.

HotpotQA (Yang et al., 2018) is a QA dataset composed of questions requiring the QA model to jointly reason over multiple passages of multi-document contexts. Inoue et al. (2020) enrich the dataset with explanations from three human annotators. The explanations are structured in the form of triples (e_1, r, e_2), associating two entities (e_1 and e_2) through a relation r, such as ("Scott Derrickson", "is", "an American director").

We extract the shared concepts $\mathcal{C}$ as pairs of (r, e_2); Hence, Conceptual few-shot will prime the prediction with questions and contexts presenting the same entities in analogical relations to the ones the model should understand for correct prediction.

GLUE Diagnostic (Wang et al., 2018) contains approximately 1,100 diagnostic samples of Natural Language Inference intended to fool a simple statistical model. While the concepts are heuristically extracted in other cases, GLUE diagnostic directly annotates 30 distinct logical concepts needed in prediction, such as *double negation*, *conjunction*, or *existential quantification*. We directly use these logical concepts as the reasoning concepts $\mathcal{C}$.

3.3 Baseline model (BASELINE-TK-QA-1B)

To contextualize the results of existing In-context learners, we additionally evaluate a simple newly-created few-shot in-context learner trained on a single QA dataset. Similarly to TK-INSTRUCT, we construct the training examples of the meta-learning task in the explicit few-shot learning format, as initially proposed by Min et al. (2022a), where the model is updated to predict correct labels with a set of randomly-selected demonstrations included in the input (Eq. (1)). This way, we fine-tune a T5-LARGE model (Raffel et al., 2020) on AdversarialQA dataset (Bartolo et al., 2021) until convergence on a validation split. We assess the resulting model on Conceptual few-shot learning together with other in-context learners, denoting its results as BASELINE-TK-QA-1B.

4 Results and Discussion

Figure 2 shows the change of models' error between a random selection of demonstrations and Conceptual few-shot learning, i.e. with demonstrations sharing a selected concept (§2), ordered by models' size.

For each of the assessed concepts, we observe statistically significant improvement for at least one of the models, which confirms our initial assumption on the informativeness of the extracted concepts in prediction.

However, we can see that the selection of demonstrations makes a large difference in many cases, and the difference also largely depends on the inspected concept. Following the results of specific models, we see many cases where the model is able to utilise one concept but fails to utilise, or even worsen the prediction then exposed to the other. The variance is larger for instruction-tuned TK-INSTRUCT models, excelling in utilising shared reasoning logic of GLUE, but to the contrary, degrading when being exposed to demonstrations supported by the shared facts in WorldTree. Contrary to these results is the case of *InstructGPT* that is agnostic to concepts except for GLUE.

Figure 3 shows the average of changes of Conceptual few-shot evaluation over the inspected four concepts. The aggregation uncovers that the gain from providing informative demonstrations largely varies among models, with T0 and smaller models ($\leq$3B) benefiting from the presented concepts slightly more often; This could be caused by larger models' increasing reliance on their memorized knowledge. However, within the model-type groups, we also note that this trend is disputed by T0 and FLAN models.

5 Conclusion

This work introduces a task of conceptual few-shot learning that reflects on in-context learners' ability to learn to apply a specific reasoning concept that can be informative for prediction. We assess a set of recent in-context learners for this ability over a set of concepts extracted from human explanations.

We find that none of the learners can benefit consistently from all concepts, even though at least one of the other models proves the concept to bear an informative value. Despite that, we still observe some interesting trends, such as the models of T0 are able to benefit from the concepts more often than others or that the concept-learning ability does not appear to relate to the model size.

We believe the future work can inspire in identifying possibly complex reasoning concepts in the explanations of human annotators and will scale the conceptual evaluation to a wider variety of concepts. We trust that an evaluation with a comprehensive selection of the concepts will allow us to more realistically assess the abilities of the newly-designed language models in the fast-progressing development of new in-context learners.

Limitations

Concepts In this work, we extract the concepts from semi-structured explanations whose format reassures consistency and non-ambiguity of the exploited concept(s). The selection of datasets and corresponding *concepts* is primarily conditioned by data availability, as the semi-structured explanations are available merely for a small set of datasets.

We acknowledge that our selection of concepts is not representative for a vast variance of concepts that users might expect models to learn from context in interaction. Some important concepts' features that we identify are following: **(i)** a number of premises or reasoning inference steps needed to map the input to output, **(ii)** the granularity of the reasoning steps, **(iii)** a type of the premises; For instance, whether the familiarity with a given concept requires a *memorization* of an entity property (such as "*sun emits light*"), or a *reasoning mechanics* such as analogical reasoning ("*if animals can run and cat is an animal, then a cat can run*").

We invite future work to identify or propose a taxonomy that would better reflect the wide variance of reasoning concepts that models are expected to comprehend in order to serve a wide scope of unseen tasks. Such taxonomy can motivate a more targeted collection of concepts from explanations, or annotation of new explanations demonstrating new concepts.

Models We acknowledge the limitation in a variance of evaluated models given by their availability and our computational possibilities. We evaluate only two models of the GPT family due to the usage limits of OpenAI API. Outside GPT models, we do not evaluate models over 20B parameters, given the infrastructure requirements of such settings. Nevertheless, we argue that the relevance of the models with constrained access, or resource requirements exceeding the limits of most organizations also remains a subject of open question.

Datasets One should note that the sizes of our evaluation datasets, for which we are able to extract concepts from explanations (Fig. 2), are too small to compare concept sensitivity between models. The sizes of our sensitivity evaluation datasets are the following: WorldTree: 2,204 samples, OpenBookQA: 792, GLUE Diagnostics: 282 samples, HotpotQA: 182 samples.

Ethical Considerations & Broader Impact

As outlined in Section 1, in-context learning recently presents a research direction of broad public interest, where the outstanding results on NLP benchmarks often do not meet the users' expectations. It is understandable that the focus of development in in-context learning LLMs goes to measurable improvements on existing benchmarks, as ecologically-valid evaluations (de Vries et al., 2020) on end use-cases are timely and challenging to compare to related work.

Nevertheless, in this highly-exposed and fast-paced direction, we identify the necessity for the emergence of fast proxy measures that can shed light on the decision-making of the LLMs as expected by their end users.

The presented evaluation of models' sensitivity to demonstrated reasoning concepts introduces a technical framework for quickly assessing models' compliance with our expected functioning; However, a selection of a comprehensive set of concepts that we can agree our models should be able to learn remains a subject of open discussion.

Acknowledgements

We acknowledge the Centre for Biomedical Image Analysis at Masaryk University, supported by MEYS CR (LM2023050 and CZ.02.1.01/0.0/0.0/18_046/0016045 Czech-BioImaging) for their support in obtaining evaluations presented in this paper.

References

Stephen H. Bach, Victor Sanh, Zheng-Xin Yong, Albert Webson, Colin Raffel, Nihal V. Nayak, Abheesht Sharma, Taewoon Kim, M Saiful Bari, Thibault Fevry, Zaid Alyafeai, Manan Dey, Andrea Santilli, Zhiqing Sun, Srulik Ben-David, Canwen Xu, Gunjan Chhablani, Han Wang, Jason Alan Fries, Maged S. Al-shaibani, Shanya Sharma, Urmish Thakker, Khalid Almubarak, Xiangru Tang, Xiangru Tang, Mike Tian-Jian Jiang, and Alexander M. Rush. 2022. Promptsource: An integrated development environment and repository for natural language prompts.

Max Bartolo, Tristan Thrush, Robin Jia, Sebastian Riedel, Pontus Stenetorp, and Douwe Kiela. 2021. Improving question answering model robustness with synthetic adversarial data generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8830–8848, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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*, volume 33, pages 1877–1901. Curran Associates, Inc.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling Instruction-Finetuned Language Models. *arXiv e-prints*, page arXiv:2210.11416.

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

Harm de Vries, Dzmitry Bahdanau, and Christopher D. Manning. 2020. Towards ecologically valid research on language user interfaces. *ArXiv*, abs/2007.14435.

Ian J. Goodfellow, Mehdi Mirza, Xia Da, Aaron C. Courville, and Yoshua Bengio. 2014. An Empirical Investigation of Catastrophic Forgeting in Gradient-Based Neural Networks. *CoRR*, abs/1312.6211.

Naoya Inoue, Pontus Stenetorp, and Kentaro Inui. 2020. R4C: A benchmark for evaluating RC systems to get the right answer for the right reason. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6740–6750, Online. Association for Computational Linguistics.

Peter Jansen, Elizabeth Wainwright, Steven Marmorstein, and Clayton Morrison. 2018. WorldTree: A corpus of explanation graphs for elementary science questions supporting multi-hop inference. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. ACL.

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

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.

Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2022a. MetaICL: Learning to learn in context. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2791–2809, Seattle, United States. Association for Computational Linguistics.

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

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. Cross-task generalization via natural language crowdsourcing instructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3470–3487, Dublin, Ireland. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*.

Haig Plotnik, Rod; Kouyomdijan. 2012. *Discovery Series: Introduction to Psychology.*, 1 edition. Wadsworth Cengage Learning.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(146):1–67.

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

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine

Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M Rush. 2022b. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proc. of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. ACL.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, et al. 2022a. Super-naturalinstructions:generalization via declarative instructions on 1600+ tasks. In *EMNLP*.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Kuntal Kumar Pal, Maitreya Patel, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Savan Doshi, Shailaja Keyur Sampat, Siddhartha Mishra, Sujan Reddy A, Sumanta Patro, Tanay Dixit, and Xudong Shen. 2022b. Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, and Tengyu Ma. 2023. Larger language models do in-context learning differently.

Sarah Wiegreffe and Ana Marasović. 2021. Teach me to explain: A review of datasets for explainable nlp. In *Proceedings of NeurIPS*.

Zhengnan Xie, Sebastian Thiem, Jaycie Martin, Elizabeth Wainwright, Steven Marmorstein, and Peter Jansen. 2020. WorldTree v2: A corpus of science-domain structured explanations and inference patterns supporting multi-hop inference. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5456–5473, Marseille, France. European Language Resources Association.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

A Details of Concept-aware Evaluations

Unless stated otherwise, we evaluate *all* models over *all* datasets and *both* demonstrations selection strategies consistently for ROUGE-L in default settings of Lin (2004), using a number of demonstrations $k = 3$ and contexts constructed in the following format:

"*Input:* x_1 *Prediction:* Y_1 *Input:* x_2 *Prediction:* Y_2 *Input:* x_3 *Prediction:* Y_3 *Input:* x_{pred}"

Among both random and concept-sharing evaluations, we share the same x_{pred} and only permute the demonstrations; We find cases where the filtering of predicted samples (x_{pred}) to the ones sharing a concept with sufficient amount of (3) different samples needed for demonstrations makes the task systematically easier.

We diverge from the stated configuration only in the following cases:

- TK-INSTRUCT-11B and HotpotQA: we limit the evaluation contexts to at most 3.500 unique words, as we can not fit longer contexts into the memory. This might make the absolute results in this configuration overly optimistic, but still comparable within the Conceptual few-shot evaluation.

- GPT and HotpotQA: We completely exclude these evaluations given the fixed context window size of these models will exclude the x_{pred} from prediction input in too many cases.

We choose evaluated GPT APIs based on OpenAI documentation[4], picking for GPT and INSTRUCTGPT models marked as DAVINCI and TEXT-DAVINCI-003. Note that these identifiers might change in time, thus disallowing us to guarantee the reproducibility of their evaluations.

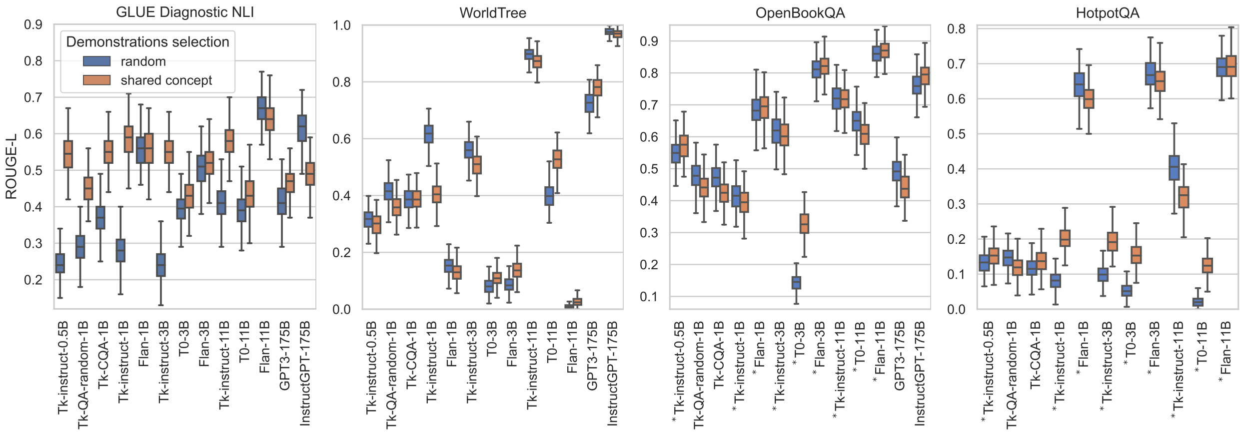

Figure 4: **Conceptual few-shot evaluation:** ROUGE-L of models using *random* demonstrations (left) and demonstrations exploiting a concept of prediction (§3.2; right). Boxes and confidence intervals cover 50% and 95% of the bootstrapped results, respectively (100 samples, 200 repeats). Models marked with * were exposed to the evaluation task (but not samples) in training. Training datasets of GPT* models are unknown.

B Computational Requirements

We run both training and evaluation experiments using single NVIDIA A100-SXM-80GB. The time and computational requirements of evaluation depend largely on the size of the evaluated model; We can evaluate the models up to 11B parameters on a single NVIDIA A100-SXM-80GB. The evaluation of Concept Few-shot learning on all our datasets, together with the Random reference evaluation takes approximately 2 hours for a 1B model.

[4] https://beta.openai.com/docs/model-index-for-researchers