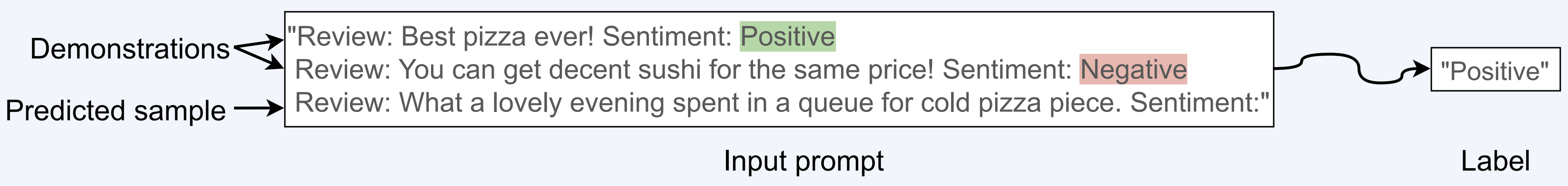
Concept-aware Data Construction Improves In-context Learning of Language Models

Michal Štefánik♣ Marek Kadlčík♣ Petr Sojka♣
♣ Masaryk University, Czech Republic

ICLR
Workshop on Mathematical and Empirical Understanding of Foundation Models

In-context learning: Ability of the model to perform previously unseen task solely from the input context

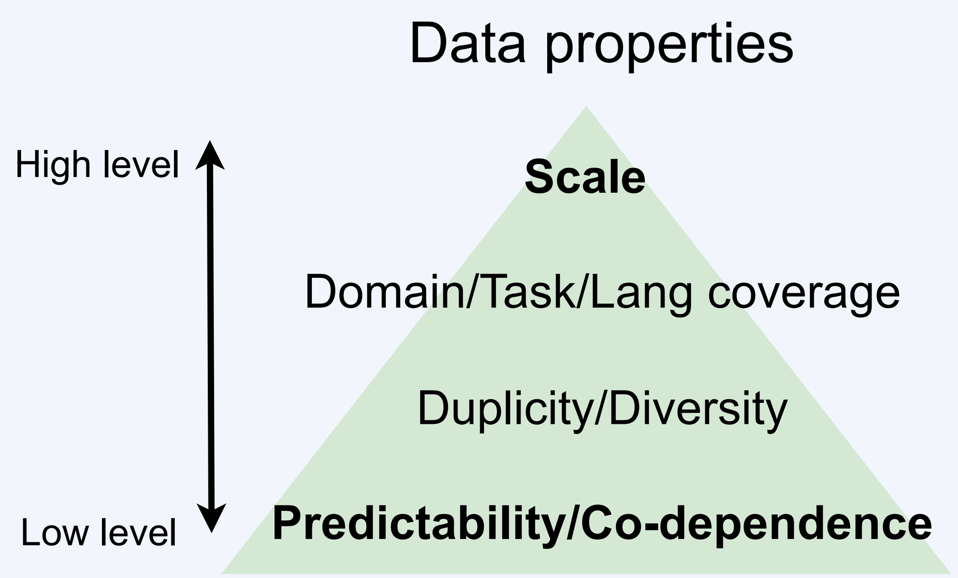


Previous work

Theory: **In-context learning emerges from specific data properties!**

- Hahn & Goyal [1]: ICL emerges in Language models thanks to a shared compositionality of languages
- Chan+ [2]: ICL needs statistical burstiness of data, (a co-occurrence of same concepts in clusters)
- Xie+ [3]: ICL requires training data that condition correct prediction on shared latent concepts

All these works train small models able of ICL in synthetic, small-data settings



Practice: **In-context learning emerges with scale!**

- Brown+ [4] (GPT3) first uncovers ICL ability by scaling model size
- Min+ [5], Sanh+ [6], Wang+ [8] (, ...) scale a diversity of tasks and promts in instruction format
- Wei+ [9] (FLAN) extend training with Chain-of-Tought tasks

What skills can the model learn from pre-training samples?

1. "Some sorts of [MASK]" *fruit* — ambiguous
2. "Some sorts of fruit overripe faster than [MASK]" *others* — pattern matching
3. "(+) Species of banana pertain over two months, but most apples will [MASK]" *last* — conditioned by contextual **concept** (*overripe*)
4. "(+) but most apples will last less than two [MASK]" *weeks* — pattern matching + cond. by contextual **concept**

→ Some samples are more useful than others
→ Most next-token prediction samples are trivial or ambiguous

Can the upscaling of concept-dependent data improve the quality of ICL?

Experimental setup

- Few-shot **instruction training** format:

$$[x_1, y_1, \langle sep \rangle, \ldots, x_k, y_k, \langle sep \rangle, x_{\text{pred}}] \rightarrow y_{\text{pred}}$$

- Demonstrations sharing a specific reasoning concept (*Informativeness* condition)
- Diverse demonstrations (*Non-triviality* condition)
- Baselines:
  - Uncontrolled demonstrations selection (Tk-Random)
  - Previous Instruction-tuned models (T0, Flan, Tk-Instr)

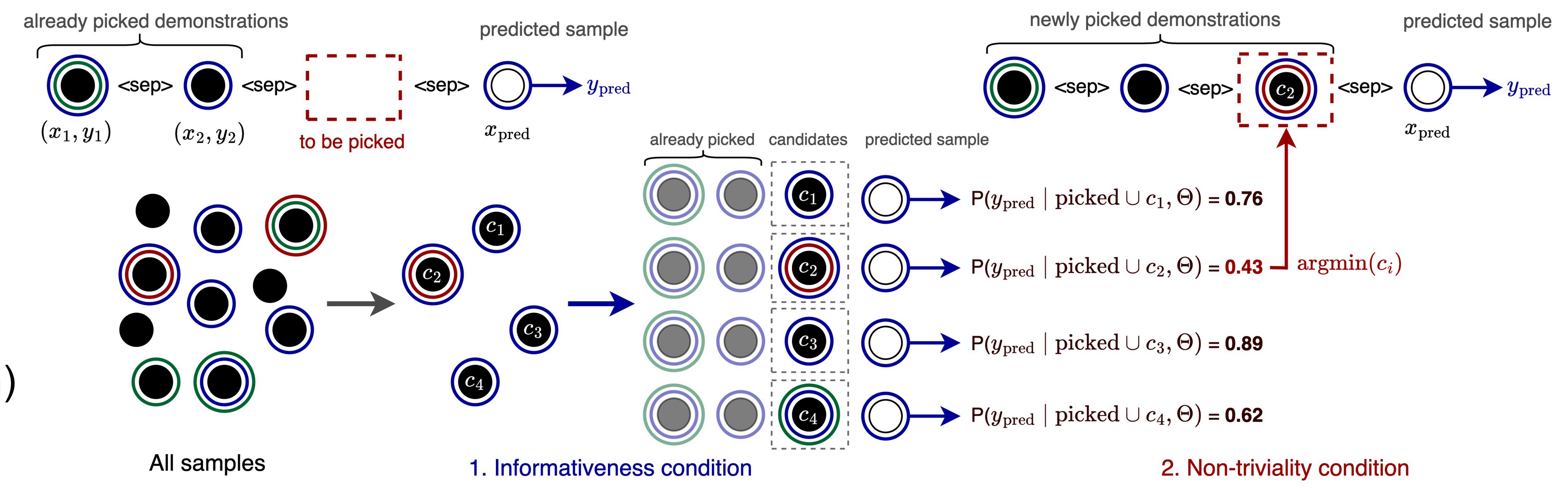
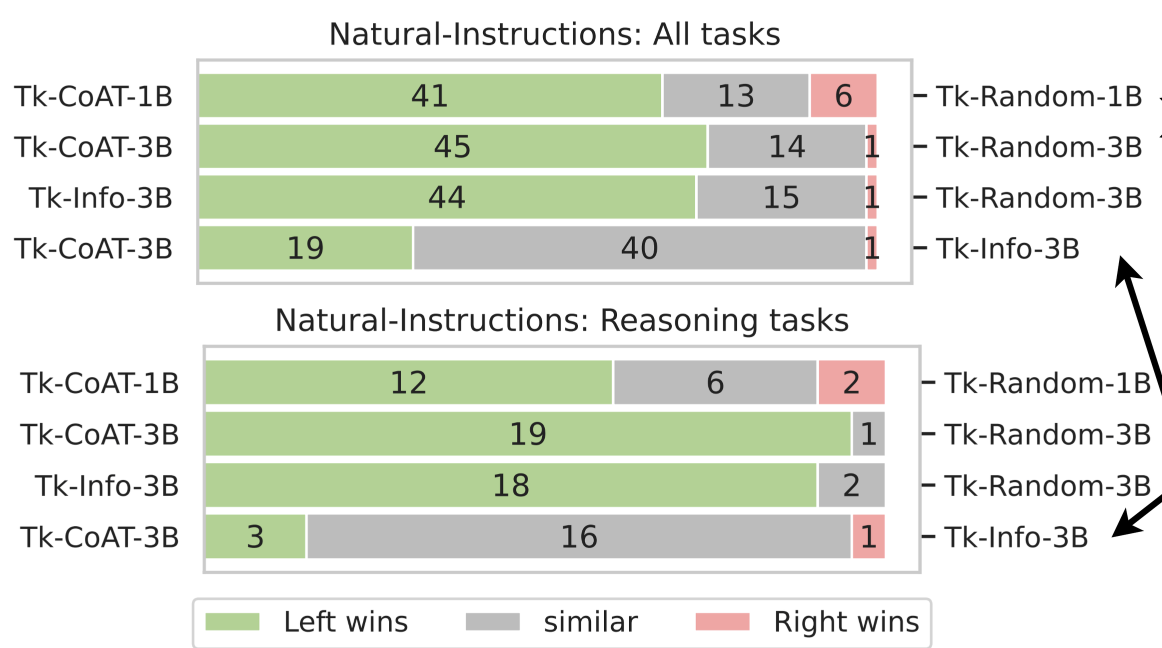


Fig: Selection of demonstrations in our implementation of Concept-aware Training (CoAT)

Training data setup

- Existing datasets with annotated concepts are not big enough for training
- We pre-train on a synthetic TeaBReAC [10] dataset which annotates **reasoning chains** (train concepts)
- We fine-tune the resulting model on AdversarialQA [11] to restore the ability to work with natural text

Evaluations

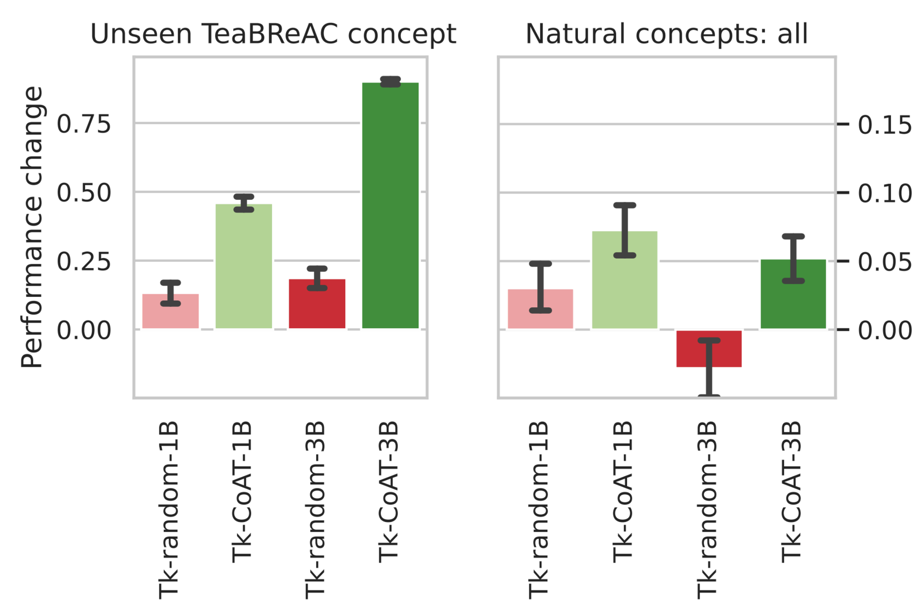


CoAT models win in 68% and 75% of tasks

...mainly thanks to the concept-sharing data construction

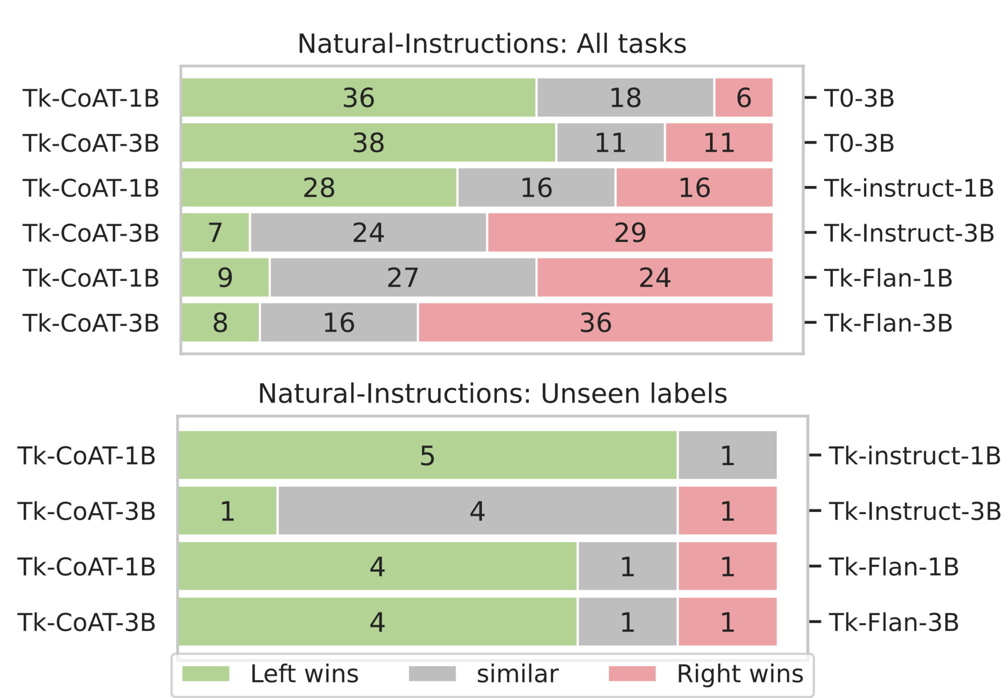
Win rates: Comparison to baselines on **(up)** all and **(down)** reasoning tasks of **Natural Instructions** collection [8]:
(1) Uncontrolled demo construction (Tk-Random) and
(2) selection with only Informativeness condition (Tk-Info).

Analyses



←Are concept-aware models better at benefiting from both synthetic and natural-language concepts?

Fig: Change in performance between in-context learning with random and concept-sharing demonstrations: **(left)** synthetic reasoning chains, **(right)** different natural-language concepts [7]



Win rates: Comparison to previous models on **(up)** all tasks, and **(down)** tasks with the labels unseen in the training.

↓Are concept-aware models more robust in in-context learning functional relations?

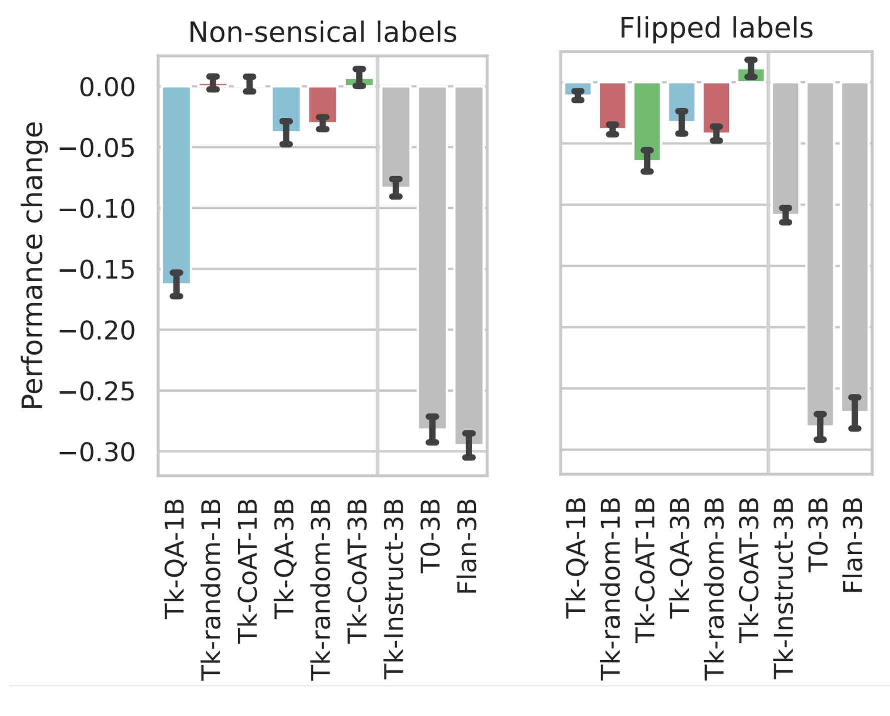


Fig: Performance change of in-context learning with **(left)** non-sensical labels ("foo", "bar" instead of "positive", "negative"), or **(right)** flipped labels.

References

[1] Hahn & Goyal: **A Theory of Emergent In-Context Learning as Implicit Structure Induction.** Arxiv 2023.
[2] Chan et al: **Data Distributional Properties Drive Emergent In-Context Learning in Transformers.** NeurIPS 2022.
[3] Xie et al: **An Explanation of In-context Learning as Implicit Bayesian Inference.** ICLR 2022.
[4] Brown et al: **Language Models are Few-Shot Learners.** NeurIPS 2020.
[5] Min et al: **MetaICL: Learning to learn in context.** NAACL 2022.
[6] Sanh et al: **Multitask Prompted Training Enables Zero-Shot Task Generalization.** ICLR 2022.
[7] Štefánik & Kadlčík: **Can In-context Learners Learn a Reasoning Concept from Demonstrations?** ACL NLRSE 2023
[8] Wang et al: **Super-NaturalInstructions: Generalization via Instructions on 1600+ NLP Tasks.** EMNLP 2022
[9] Wei et al: **Larger language models do in-context learning differently.** Arxiv 2023
[10] Trivedi et al: **Teaching Broad Reasoning Skills for Multi-Step QA by Generating Hard Contexts.** EMNLP 2022
[11] Bartolo et al: **Improving QA Model Robustness with Synthetic Adversarial Data Generation.** EMNLP 2021