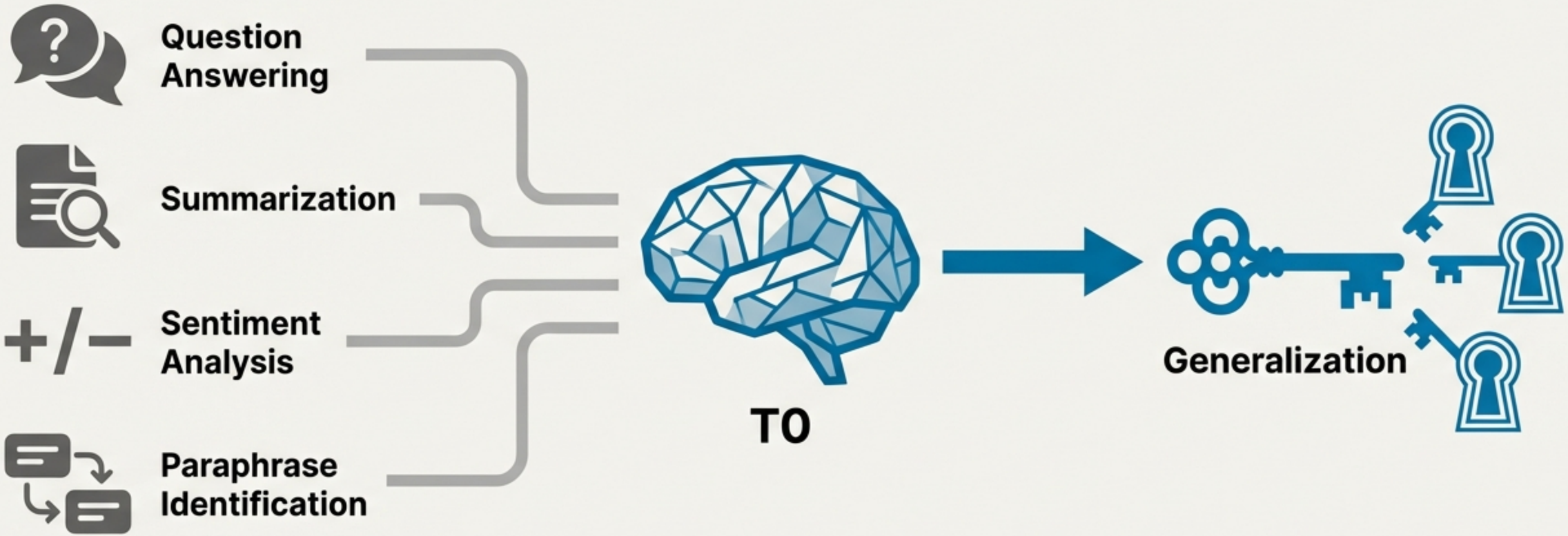
Multitask Prompted Training Enables Zero-Shot Task Generalization

Victor Sanh, Albert Webson, Colin Raffel, et al.

A collaborative effort from the BigScience workshop, involving Hugging Face, Brown University, Snorkel AI, and contributing institutions.



Large Language Models have amazing zero-shot skills, but this ability is brittle and costly.

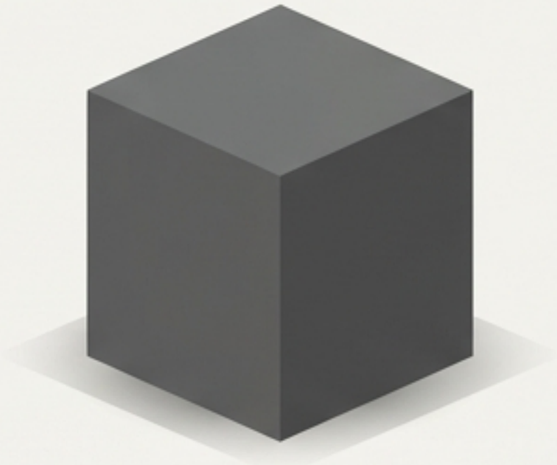
The Magic

Giant models like GPT-3 (175B parameters) can perform new tasks without specific training (zero-shot), a capability hypothesized to be an *implicit* side effect of web-scale pre-training.

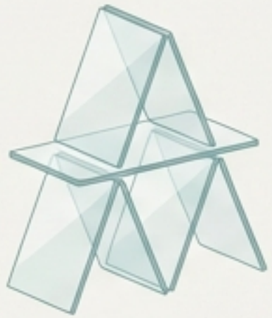
The Problem

This ability is unreliable and inefficient.

- **Brittle:** Performance is highly sensitive to the exact wording of the prompt.
- **Costly:** Requires enormous model scale, making it inaccessible and expensive to train and deploy.



Implicit Learning at Massive Scale



Prompt Sensitivity

High Cost

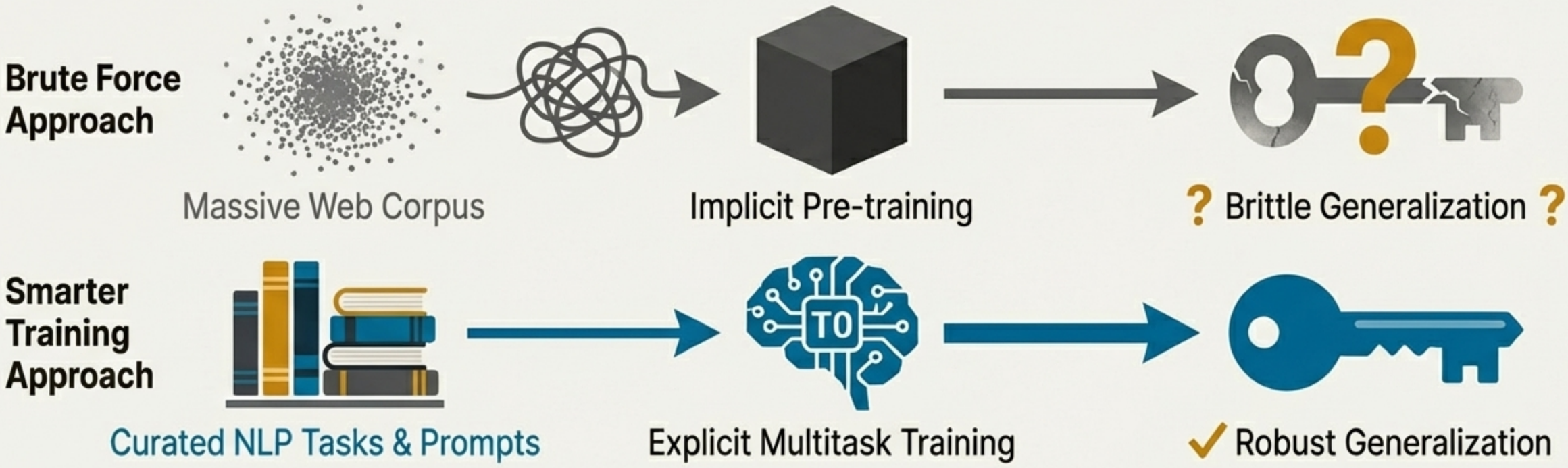
Can we directly induce zero-shot ability with explicit multitask training?

Our goal is to test if we can fine-tune a model on a massive, diverse collection of supervised tasks to build a more robust and efficient zero-shot learner.

Shift from Implicit to Explicit: Instead of hoping generalization emerges from pre-training, we actively engineer it.

Our Research Questions:

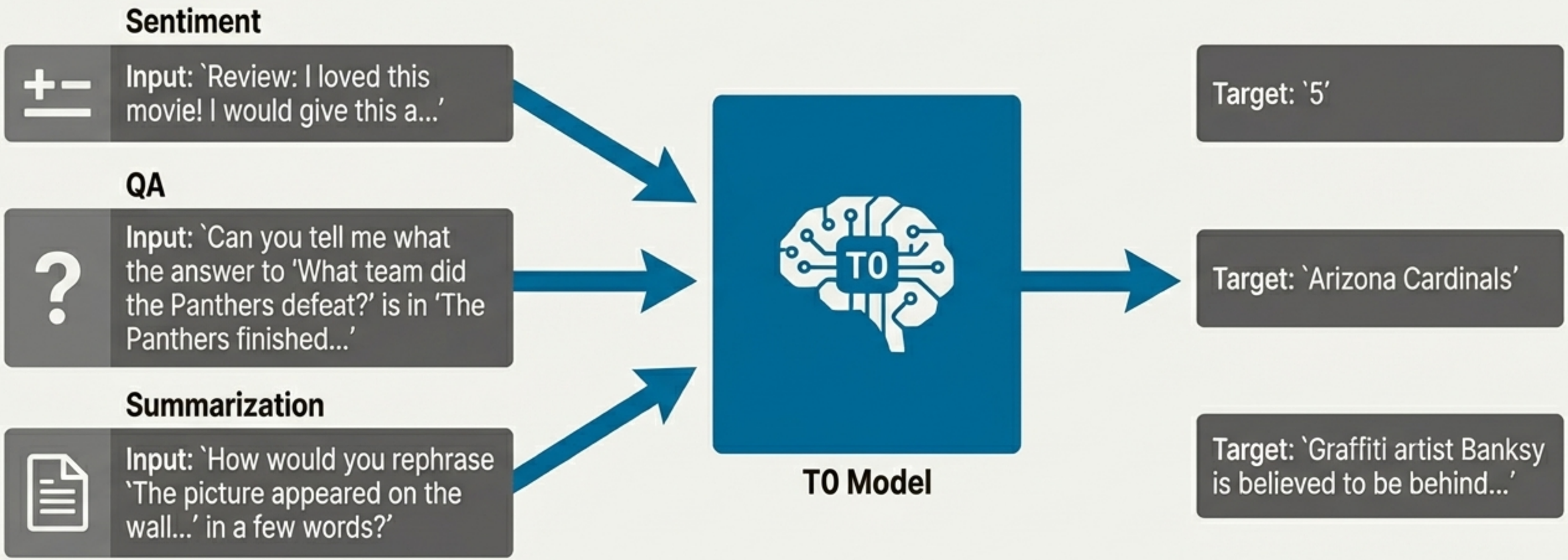
1. Does explicit multitask training improve generalization to held-out tasks?
2. Does training on diverse prompts improve robustness to prompt wording?



We frame every NLP task as a text-to-text problem using a unified prompt format.

This simple format allows us to train a single model on a vast mixture of different tasks simultaneously.

Our Model (T0): An 11-billion parameter encoder-decoder model based on T5+LM, significantly smaller than models like GPT-3.

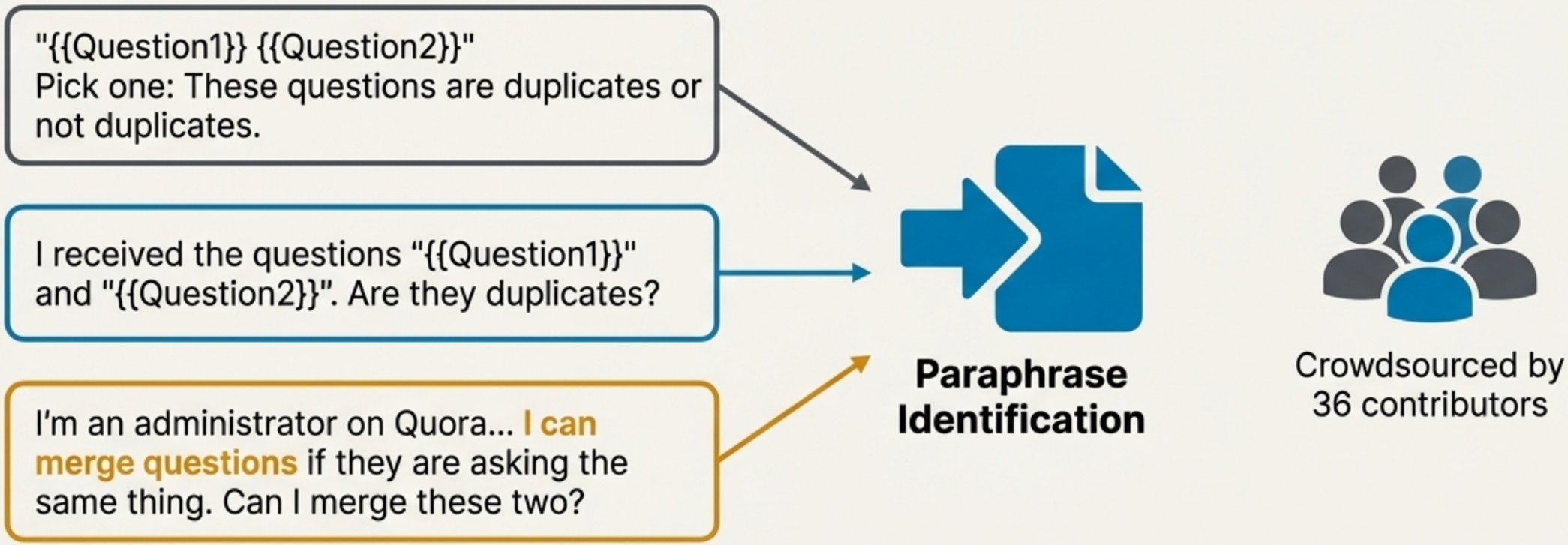


We crowdsourced thousands of diverse prompts to teach the model linguistic variety.

We built **PromptSource**, a tool enabling a global community of researchers to contribute prompts for public datasets.

Diversity by Design: Contributors were encouraged to create creative, formal, and informal phrasing to build a rich and robust training signal.

The Result (P3): The **Public Pool of Prompts** contains **over 2,000 prompts** across **177** datasets.



To prove generalization, we evaluate T0 on entire tasks it has *never* seen during training.

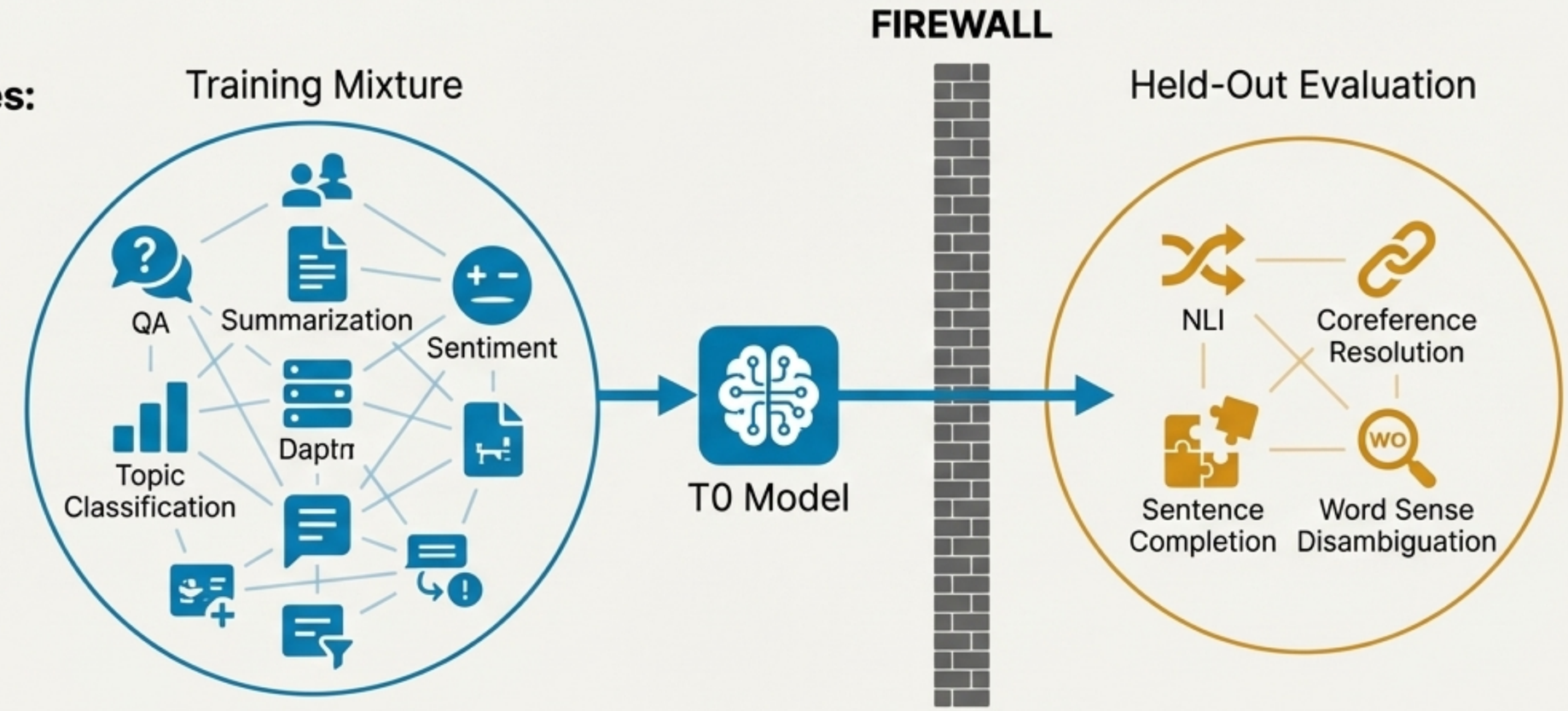
We created a strict “firewall” between our training and evaluation data to ensure a true zero-shot test.

The model was not trained on any examples from the evaluation tasks.

The model was **not trained on any even examples** from the evaluation tasks.

Held-Out Task Categories:

- Natural Language Inference (NLI)
- Coreference Resolution
- Sentence Completion
- Word Sense Disambiguation

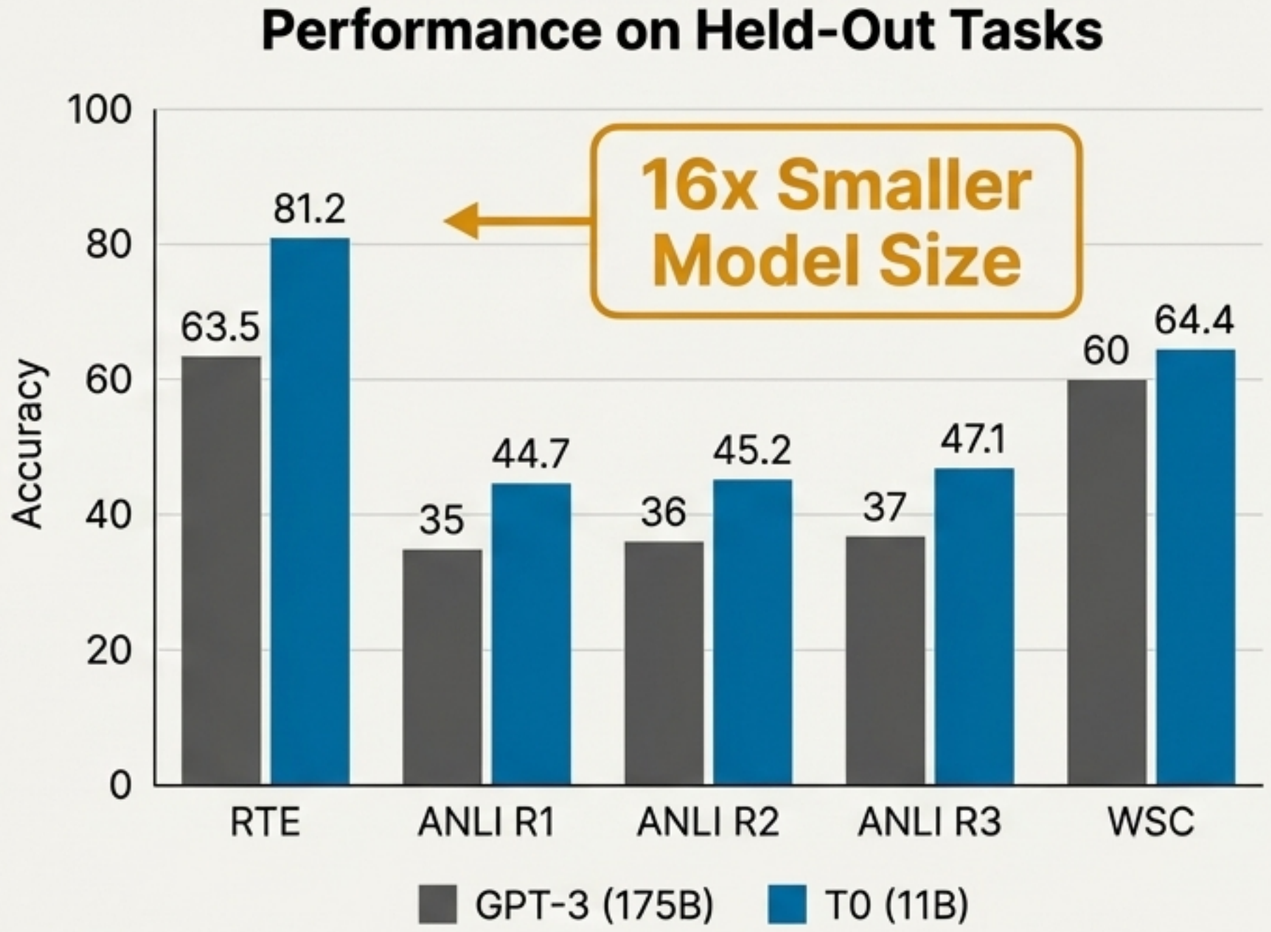


Despite being 16x smaller, T0 outperforms GPT-3 on most held-out tasks.

The Matchup: T0 (11B parameters) vs. GPT-3 (175B parameters).

The Verdict: T0 matches or exceeds GPT-3's performance on **9 out of 11** standard evaluation datasets.

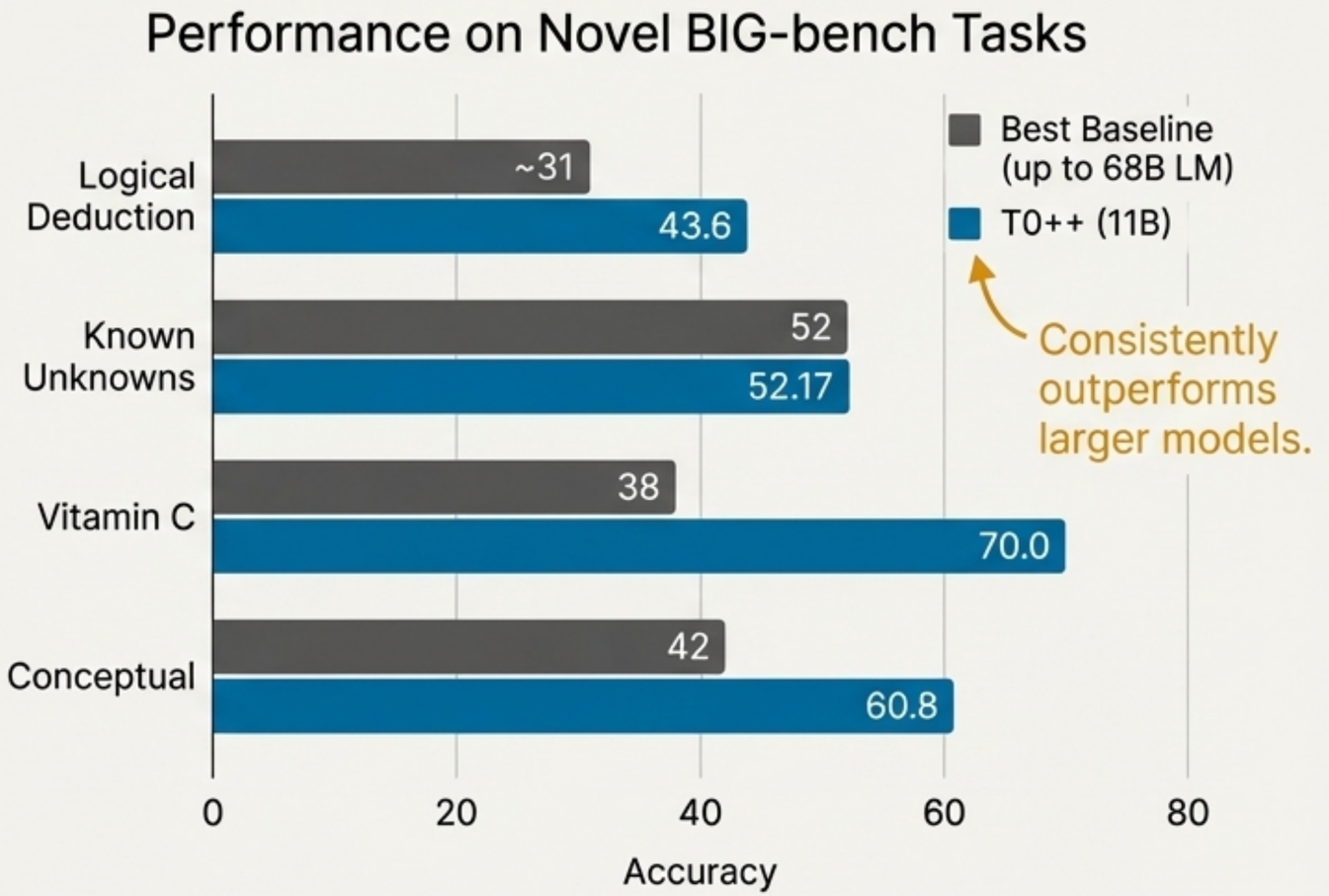
Key Win: T0 shows particularly strong generalization on Natural Language Inference (ANLI datasets), a complex reasoning task it was never trained on.



T0 also excels on novel BIG-bench tasks, outperforming models up to 6x larger.

The Challenge: BIG-bench is a community benchmark with difficult, non-standard NLP tasks designed to test the limits of model reasoning (e.g., logic puzzles, identifying misconceptions).

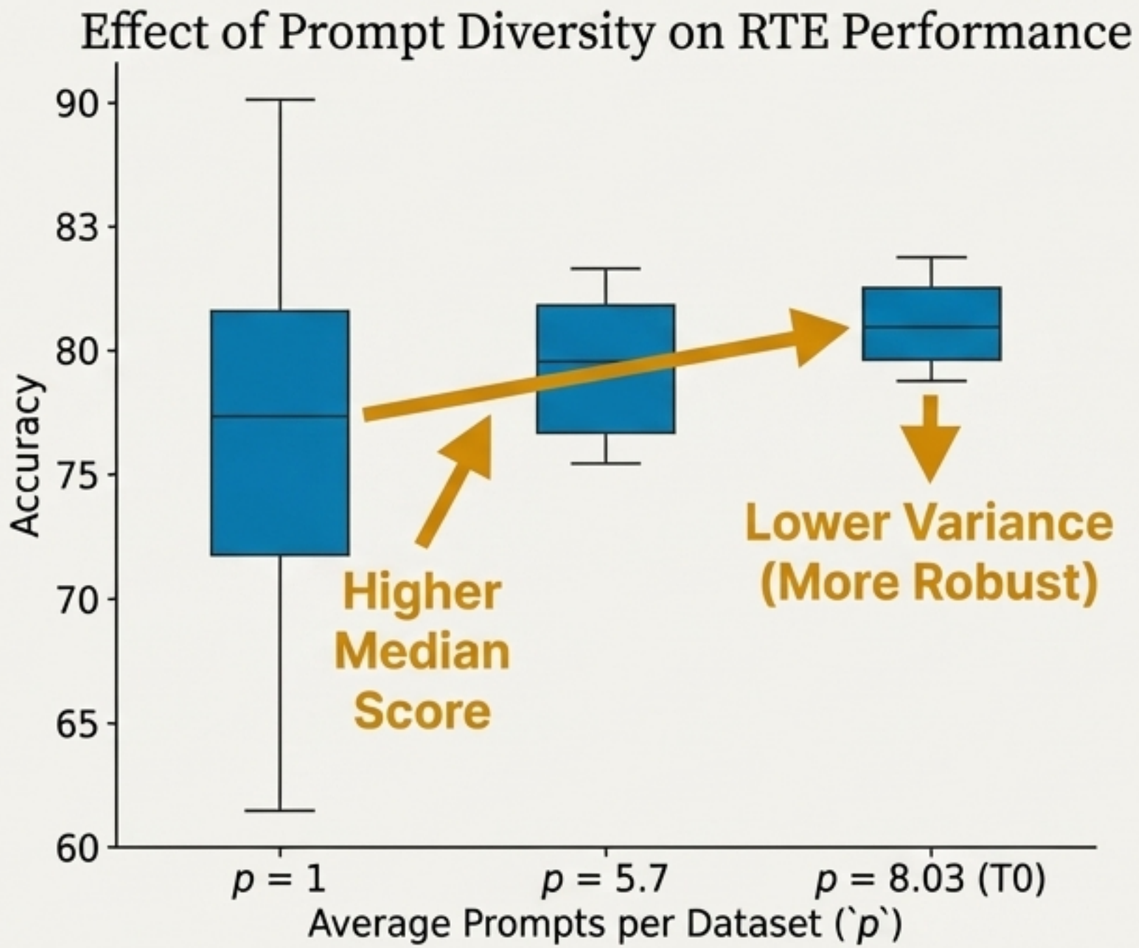
The Outcome: At least one T0 variant outperformed all baseline language models (up to 68B parameters) on **13 out of 14** tested tasks.



Training on more prompts per dataset improves both performance and robustness.

We ran an ablation study, varying the average number of prompts per dataset (p) used during training.

- **Finding 1 (Higher Score):** Increasing p consistently **improves the median performance** on held-out tasks.
- **Finding 2 (More Reliable):** It also **reduces the performance variance**, making the model less sensitive to prompt wording.

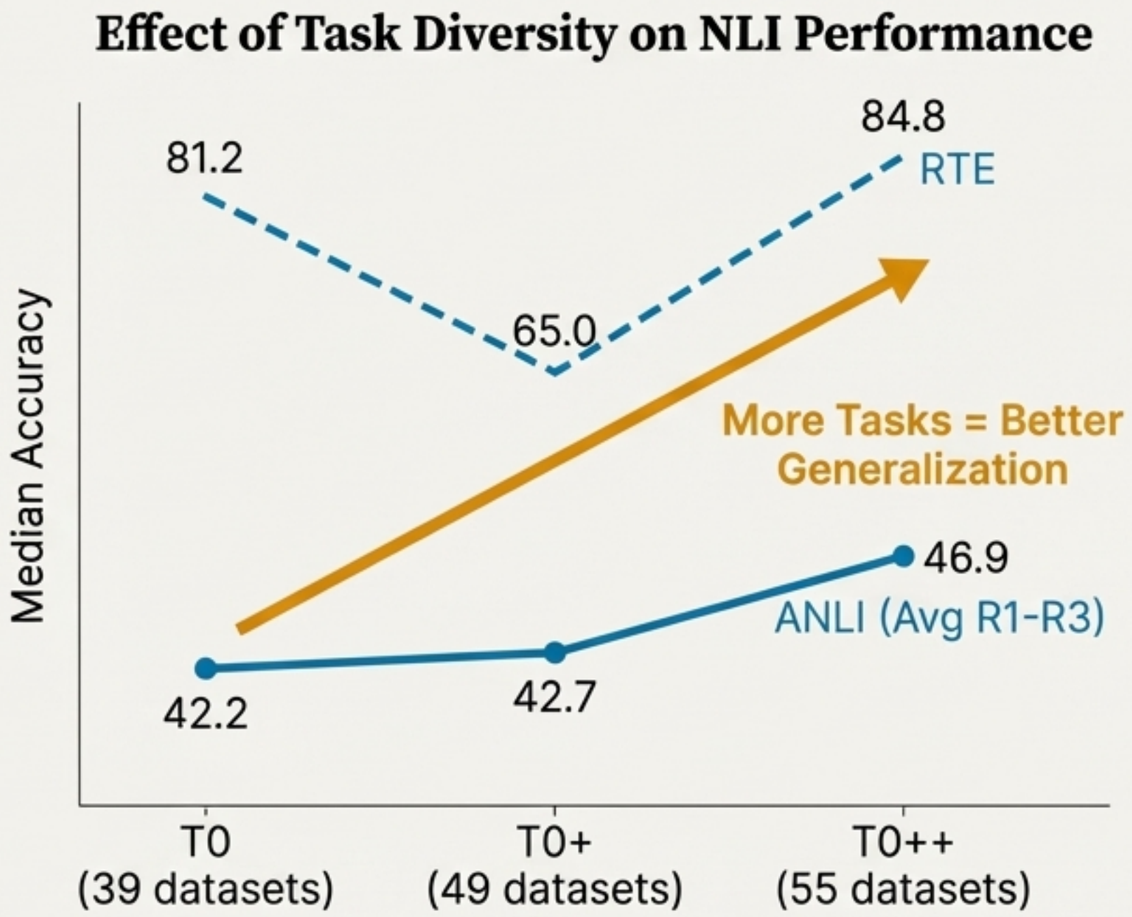


Training on a wider variety of tasks also boosts generalization performance.

We trained three model variants on progressively larger training mixtures:

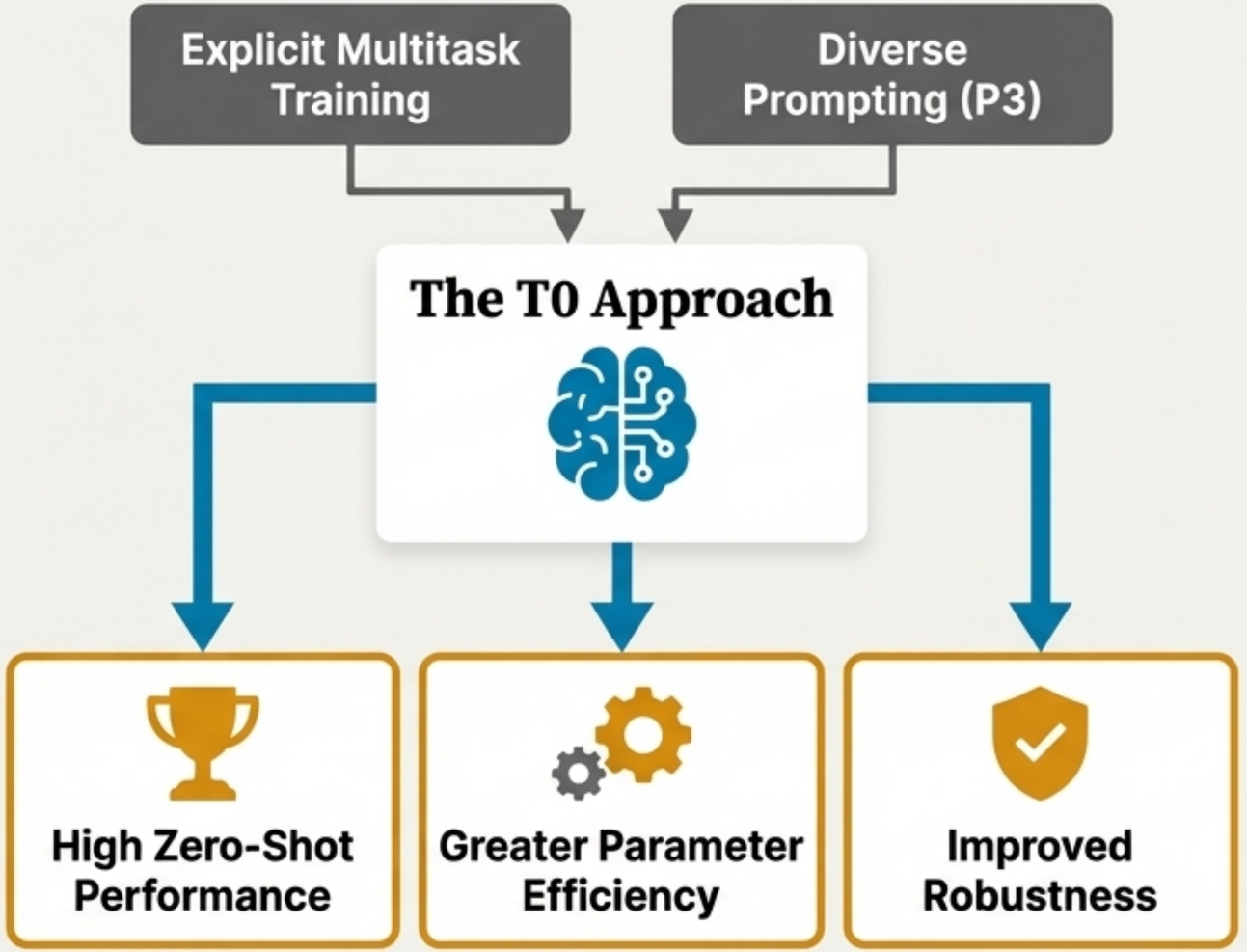
- **T0:** Standard training mixture (39 datasets)
- **T0+:** Adds GPT-3 evaluation datasets (49 total)
- **T0++:** Adds SuperGLUE datasets (55 total)

The Finding: As the number of training datasets increases, median performance on held-out NLI tasks consistently improves.



Explicit multitask training offers a more efficient and robust path to generalization.

- **Engineered Generalization**: We can directly train models for zero-shot capabilities, rather than waiting for them to emerge from scale alone.
- **Parameter Efficiency**: This approach enables smaller models to outperform giants, democratizing access to powerful AI.
- **Prompt Robustness**: Training on diverse, human-written prompts is crucial for building models that are less sensitive to phrasing.



Open questions remain on prompt engineering, task taxonomy, and architecture.

Limitations

- Underperforms on certain task types (e.g., Winogrande, HellaSwag) where pure language modeling from decoder-only models may be advantageous.
- The optimal mix and categorization of training tasks is still an open question.

Future Directions

- Systematically study what makes a prompt effective.
- Scale this approach to even more tasks, datasets, and languages.
- Explore why encoder-decoder models seem to benefit more from this approach at this scale.

All models, prompts, and tools are publicly available to build upon.

Models (T0, T0+, T0++)

github.com/bigscience-workshop/t-zero

Prompts (The P3 Collection)

github.com/bigscience-workshop/promptsource

Data (Materialized Prompts)

huggingface.co/datasets/bigscience/P3

Thank You.
Questions?