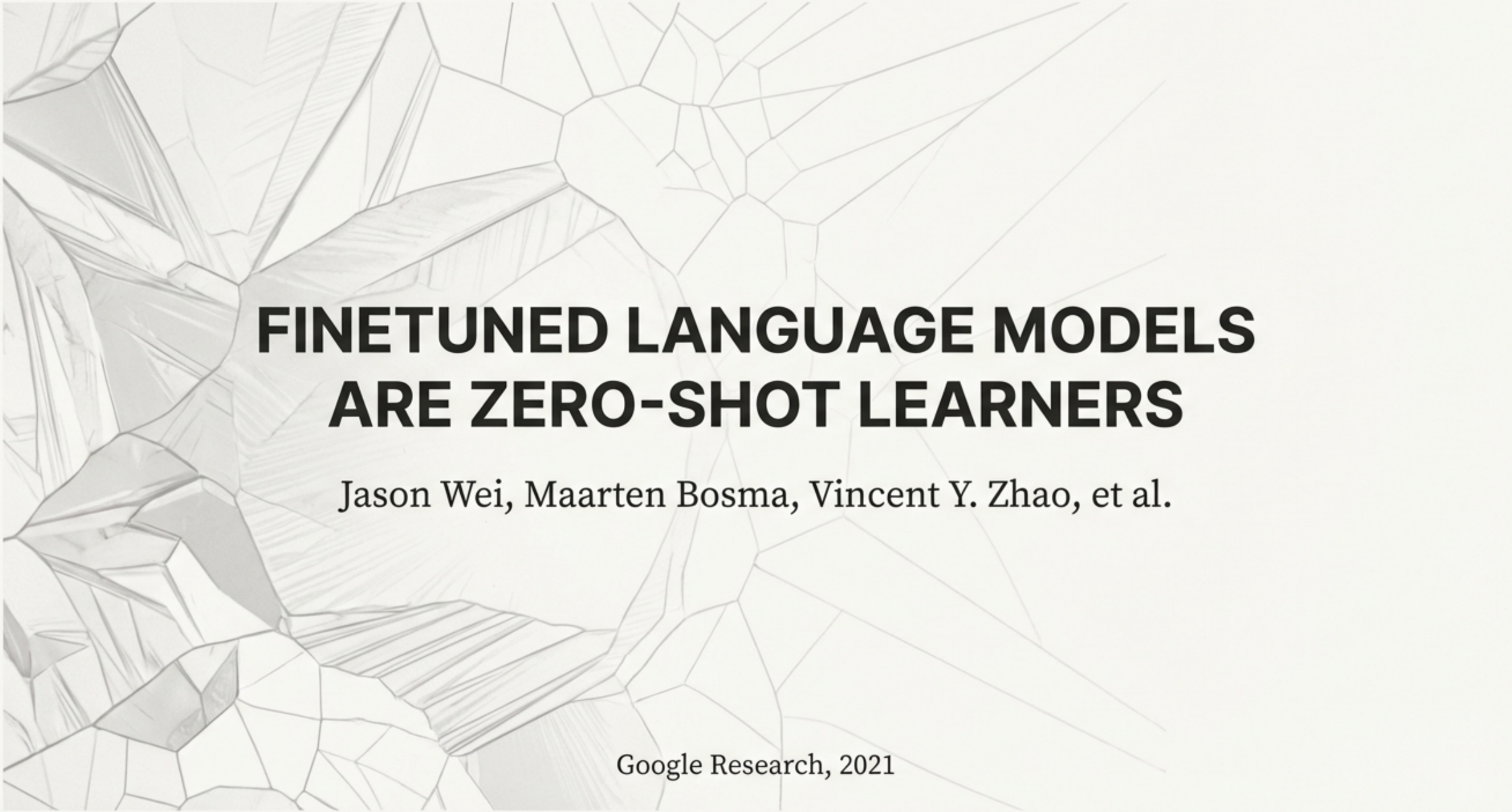


FINETUNED LANGUAGE MODELS ARE ZERO-SHOT LEARNERS

Jason Wei, Maarten Bosma, Vincent Y. Zhao, et al.

Google Research, 2021

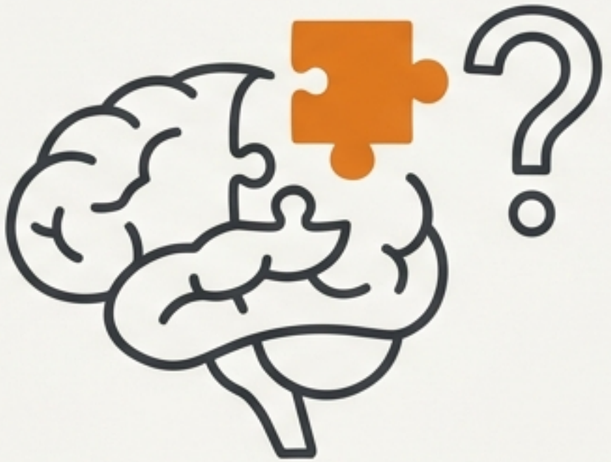
NotebookLM

Large Language Models Face a Paradox

They excel with examples but often fail with simple instructions.

Powerful Few-Shot Learners

- Models like GPT-3 show remarkable ability to adapt when given a few examples (few-shot prompting).

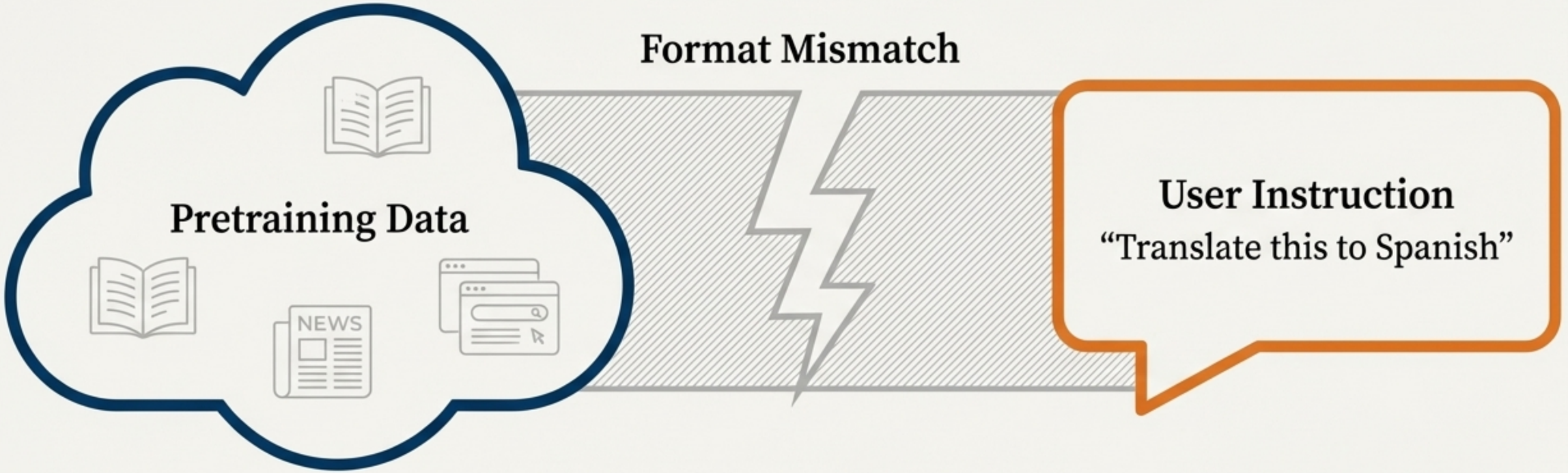


Brittle Zero-Shot Learners

- Without examples, zero-shot performance drops dramatically, even on tasks with clear instructions.

NotebookLM

The Core Issue is a Format Mismatch



- **Pretraining Objective:** Language models are trained to predict the next word in a massive text corpus (e.g., web pages, books). Their fundamental skill is text completion.
- **User’s Goal:** Users want to give direct, natural language instructions to perform a specific task (e.g., ‘Translate this to Spanish’).
- **The Gap:** An instruction doesn’t look like typical pretraining data, making it difficult for the model to “continue” the sequence in the intended way.

NotebookLM

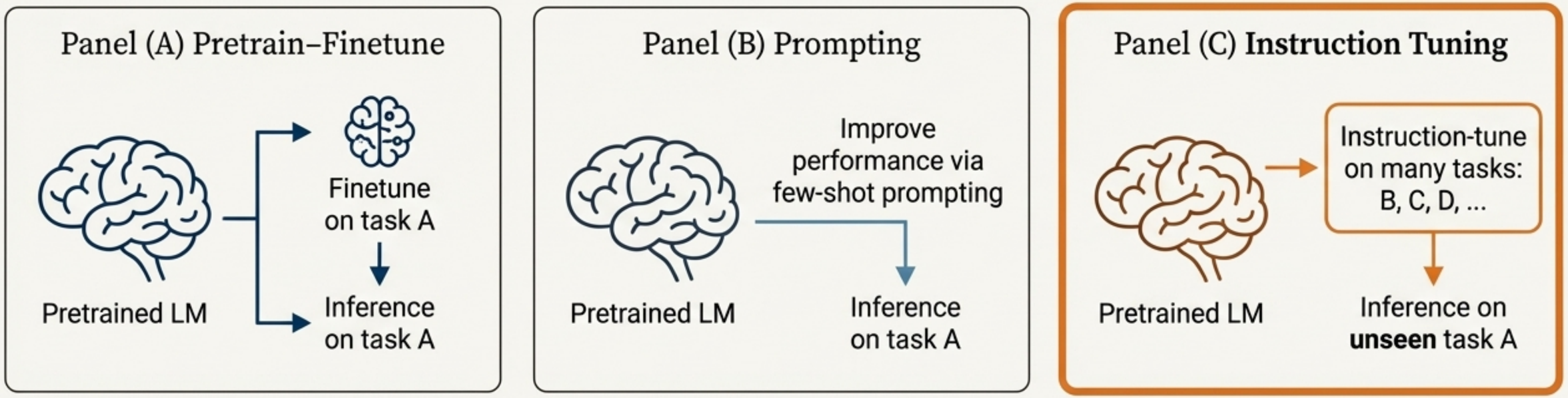
Can We Teach a Language Model to Follow Instructions?

- Hypothesis: By finetuning a model on a wide variety of tasks framed as natural language instructions...
- ...can we teach it the general skill of instruction-following?
- The Test: Will this enable the model to generalize and follow instructions for new, unseen tasks in a zero-shot setting?

NotebookLM

The Innovation: Instruction Tuning

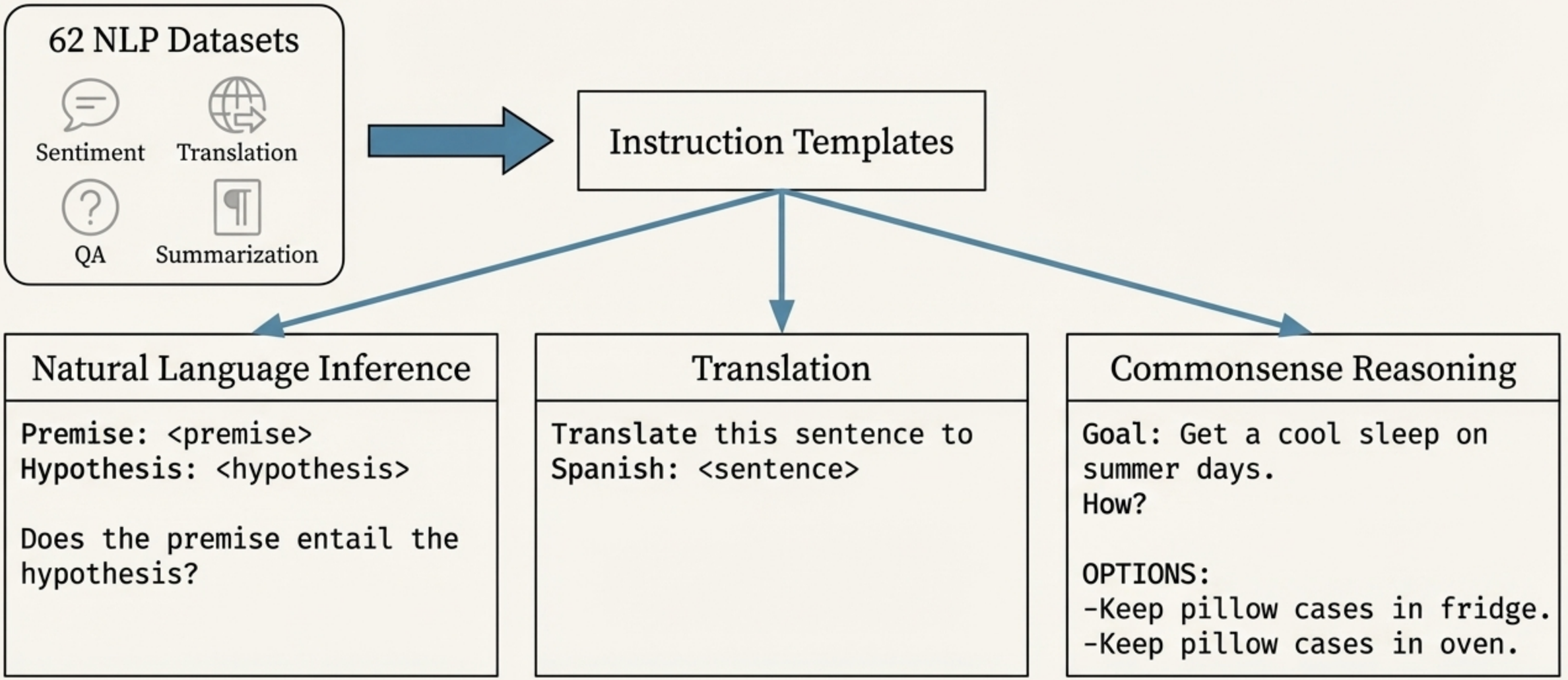
A new paradigm that teaches a pretrained LM to be a **general-purpose instruction-follower.**



- A pretrained LM is **finetuned** on a massive collection of existing NLP datasets (>60).
- Each example from every dataset is reformatted into an **instructional template.**
- The resulting model is named **FLAN: F**inetuned **L**anguage **N**et.

NotebookLM

Over 60 Datasets are Unified Through Simple Instructions

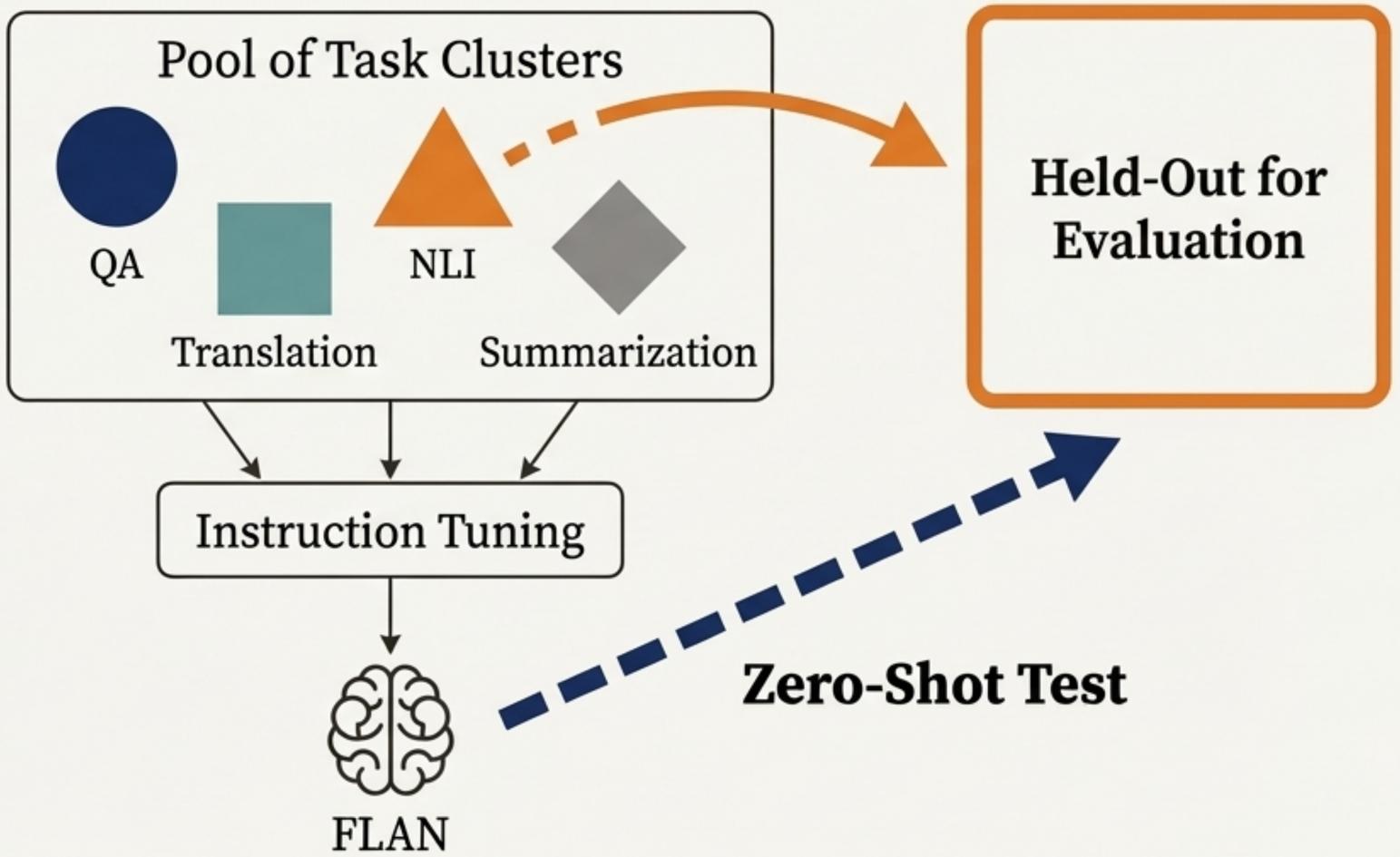


NotebookLM

The Evaluation Ensures a True Zero-Shot Test

To prove generalization, entire categories of tasks were held out from tuning.

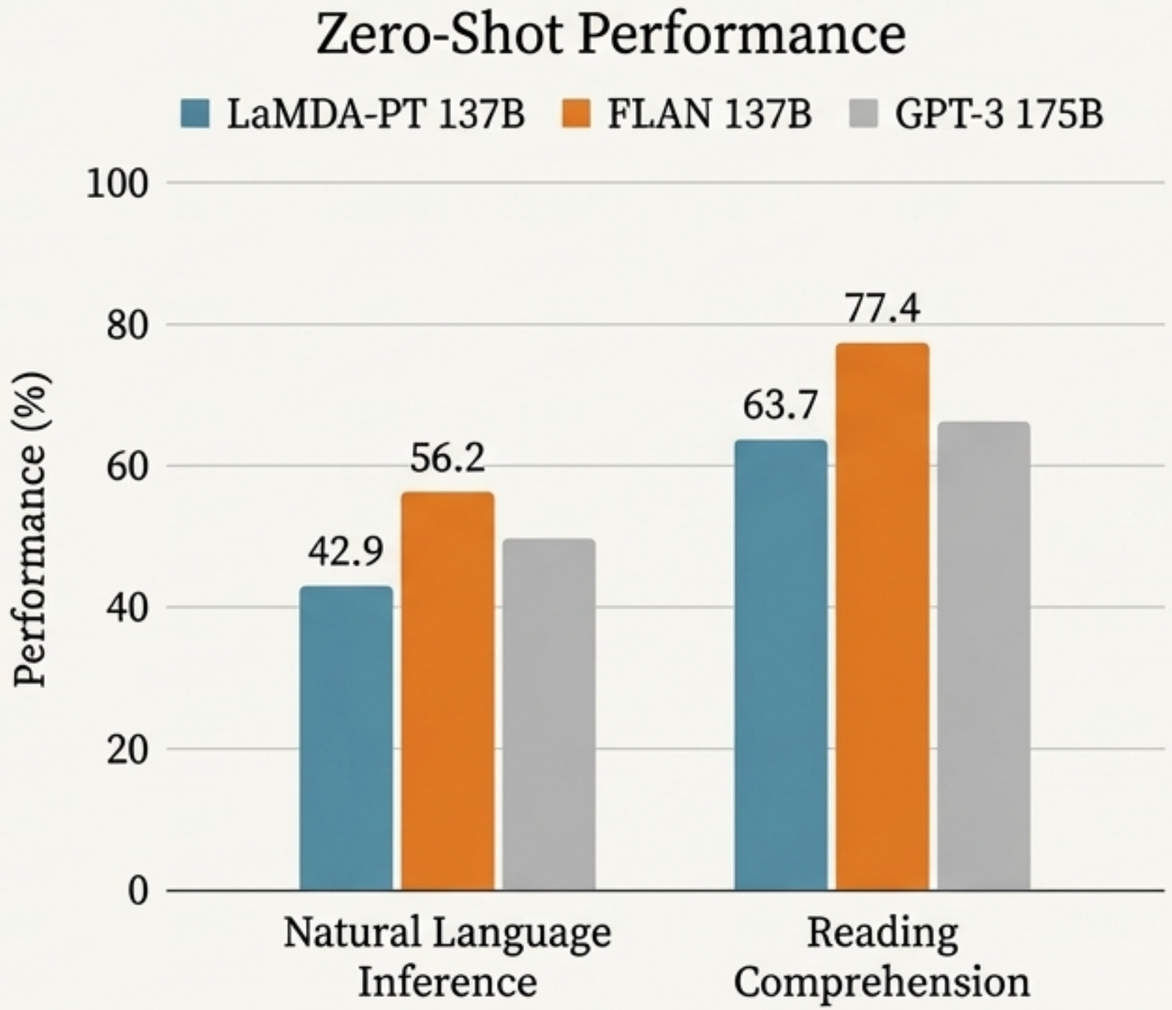
- Datasets were grouped into **task clusters** (e.g., NLI, Reading Comp, Translation).
- To evaluate on NLI, a model was instruction-tuned on **all other clusters**, having never seen an NLI task.
- This prevents the model from simply memorizing formats for a known task type and forces true generalization.



NotebookLM

FLAN Dramatically Outperforms Zero-Shot Baselines

- Zero-shot FLAN (137B) substantially improves over its untuned 137B parameter base model (LaMDA-PT).
- Zero-shot FLAN surpasses zero-shot GPT-3 (175B) on 20 out of 25 evaluated datasets.
- Gains are especially strong on tasks like Natural Language Inference (NLI), which are awkwardly phrased as text completion for standard models.



NotebookLM

The Surprise: Zero-Shot FLAN Beats Few-Shot GPT-3

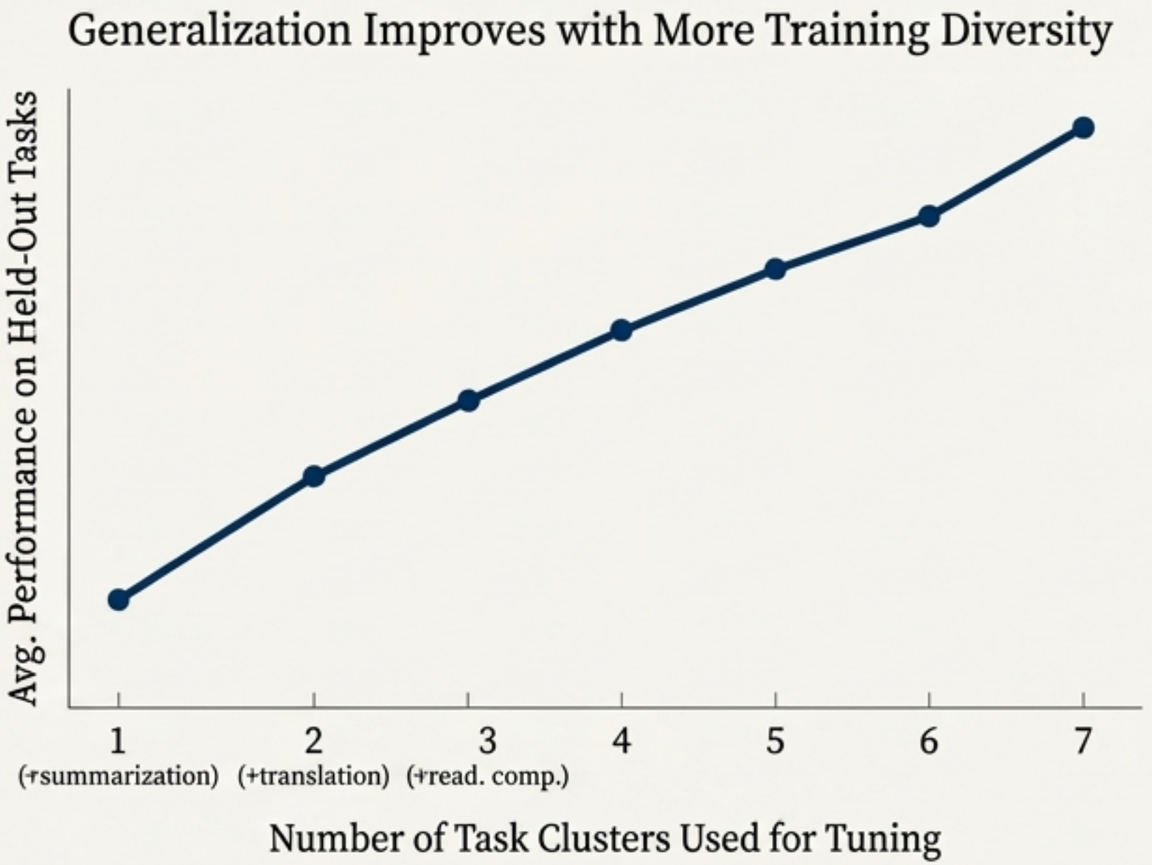
On several challenging benchmarks, a direct instruction to FLAN is more effective than giving multiple examples to GPT-3.

- FLAN's zero-shot performance exceeds few-shot GPT-3 by a large margin on:
  - ANLI (Adversarial NLI)
  - RTE (Recognizing Textual Entailment)
  - BeolQ (Reading Comprehension)
  - OpenbookQA (Question Answering)

NotebookLM

Ingredient #1: Performance Scales with Task Diversity

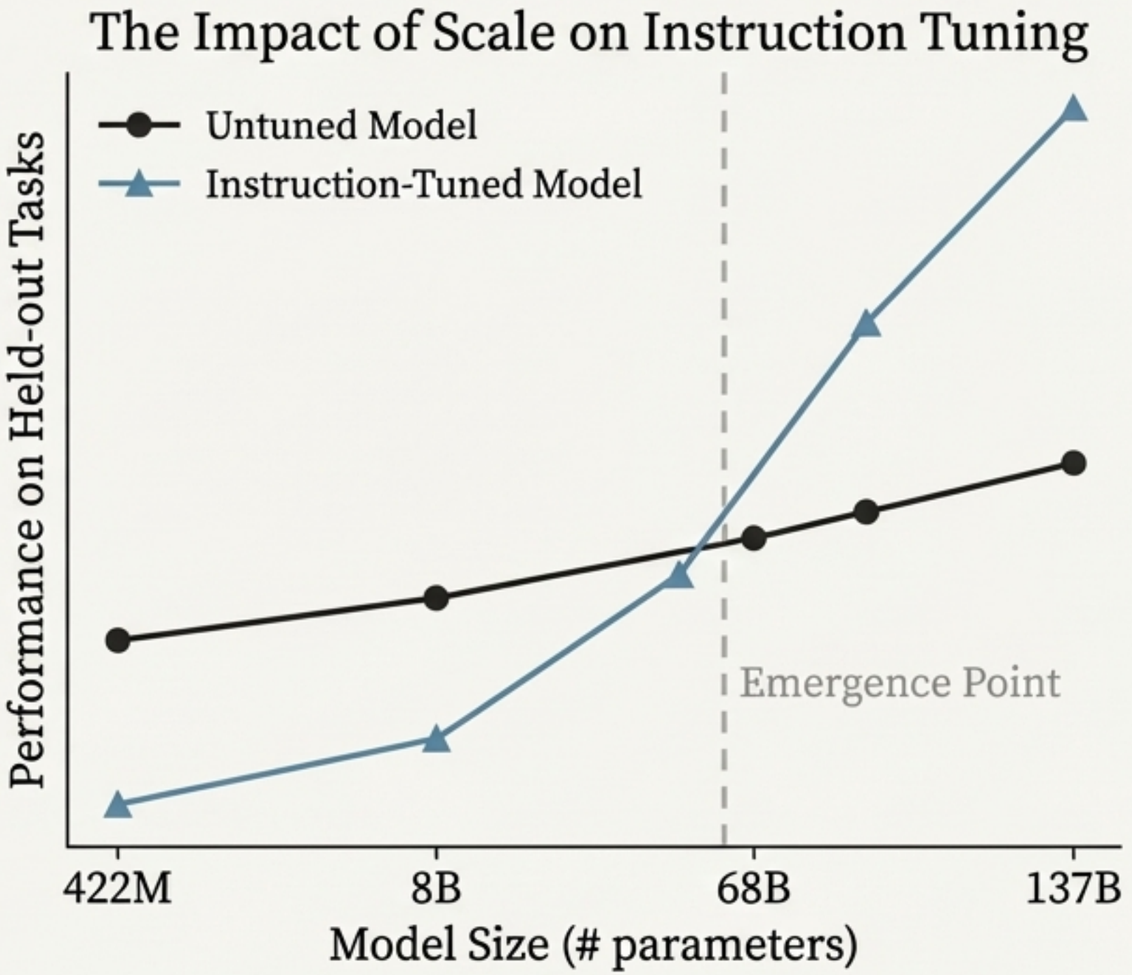
- Ablation studies show that as more task clusters are added to the instruction tuning mix, zero-shot performance on unseen clusters consistently improves.
- The performance curve does not saturate, suggesting that adding even more task diversity could lead to further gains.
- This confirms the model is learning a general skill of instruction-following, not just task-specific knowledge.



NotebookLM

Ingredient #2: Benefits Emerge Only at Massive Scale

- For models smaller than 8B parameters, instruction tuning **hurts** performance on unseen tasks, likely due to overfitting.
- For models larger than 68B parameters, instruction tuning provides a **substantial boost**, with the benefit increasing with model scale.
- This suggests a certain model capacity is required to learn the meta-skill of instruction-following in addition to the finetuning tasks.

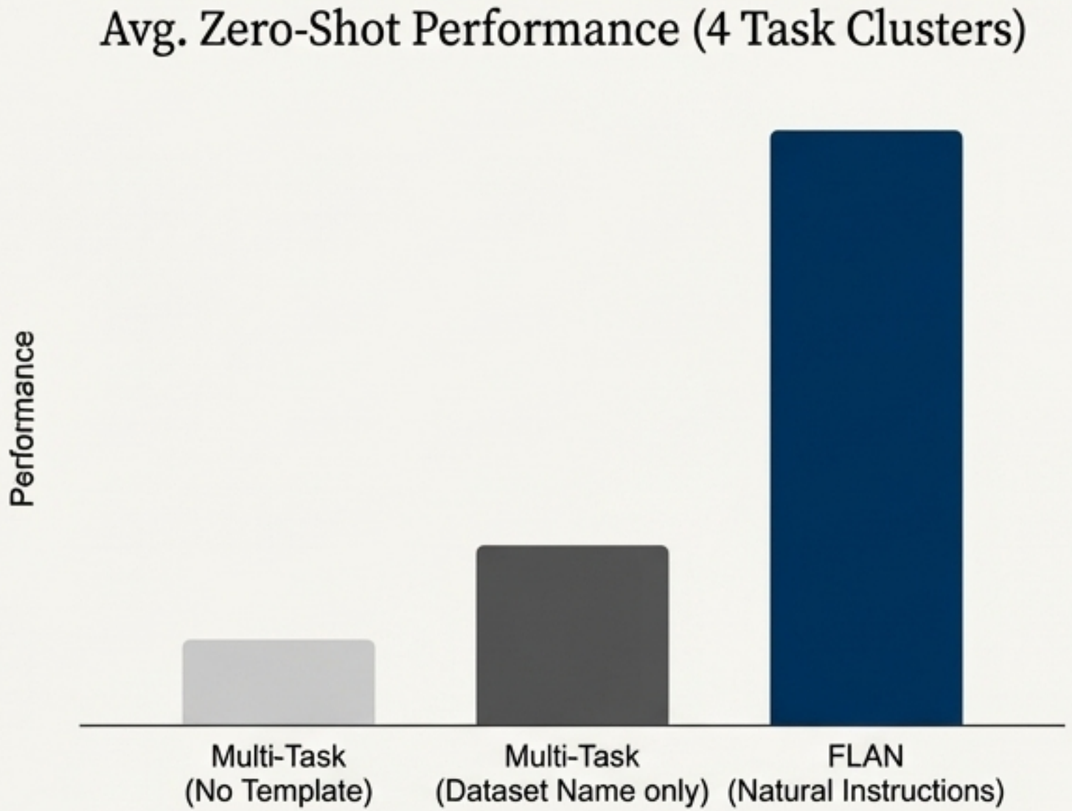


NotebookLM

Ingredient #3: Natural Language Instructions are Crucial

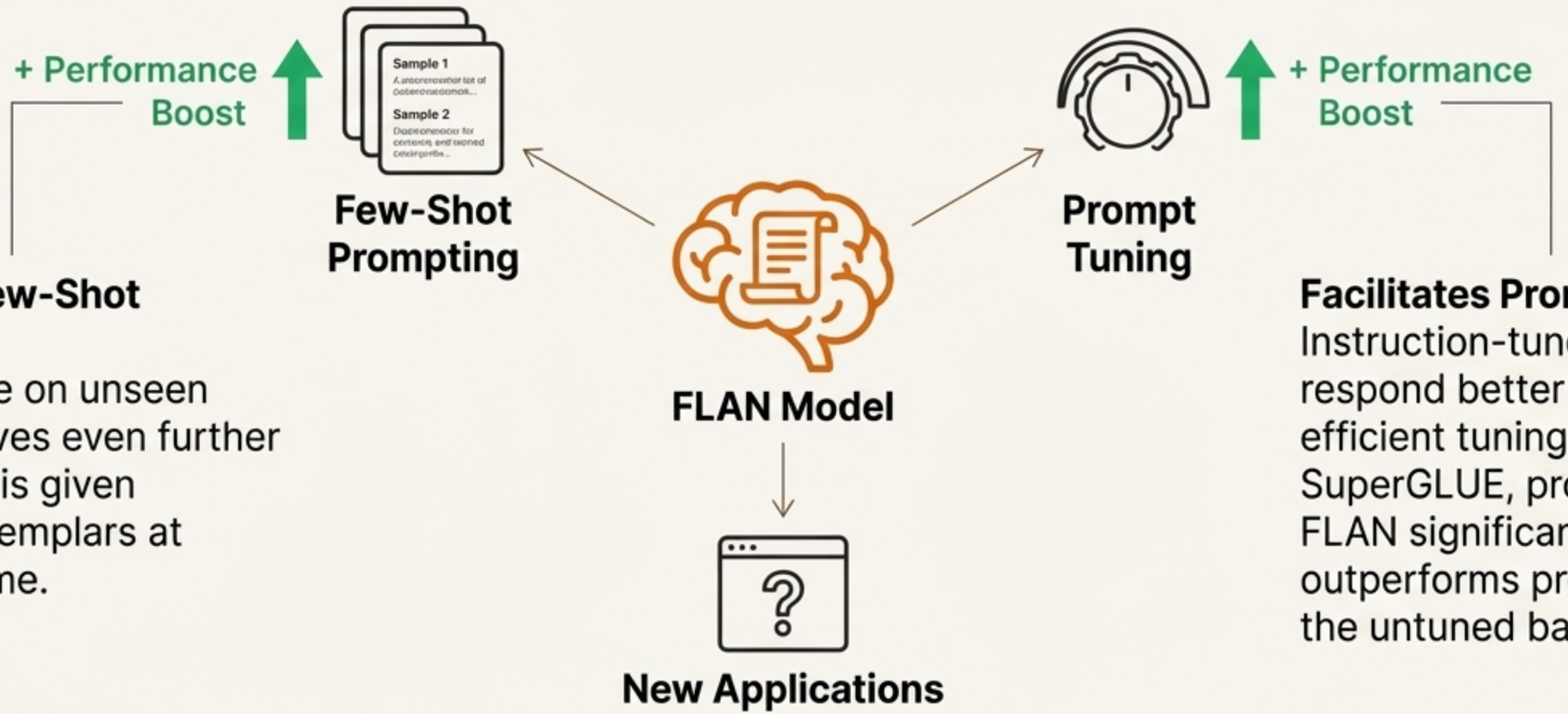
It’s not just multi-task learning; the format matters.

- An ablation that finetuned on the same tasks without instructional templates (just input–output pairs) performed substantially worse.
- Simply providing the dataset name as a prefix instead of a full instruction was also significantly less effective.
- The model must learn from human-like instructions to be able to generalize to new ones.



NotebookLM

Instruction Tuning Creates a More Capable Foundation Model



Improves Few-Shot Prompting: Performance on unseen tasks improves even further when FLAN is given few-shot exemplars at inference time.

Facilitates Prompt Tuning: Instruction-tuned models respond better to parameter-efficient tuning methods. On SuperGLUE, prompt tuning FLAN significantly outperforms prompt tuning the untuned base model.

The result is a single, generalist model that is more adept at performing a wide range of tasks out of the box.

NotebookLM

A New Paradigm for More Generalist Models

Implications

- Shifts focus from specialist models to a single, generalist, instruction-following model.
- Reduces the need for complex prompt engineering for many zero-shot tasks.
- A path to making LLMs more accessible and useful.

Limitations

- Benefits only emerge at massive (and costly) scale.
- Does not improve performance on tasks that already resemble language modeling (e.g., sentence completion).
- Subjectivity in task clustering and template creation.

Future Directions

- Scale even further: more tasks, more languages.
- Use FLAN to generate training data for smaller models.
- Explore instruction tuning to improve model fairness and safety.

NotebookLM

Instruction tuning unlocks the latent zero-shot potential of LLMs.

- ✓ It teaches models the **general skill of following instructions**, moving beyond simple pattern matching.
- ✓ It creates a single, more **versatile model** that outperforms larger competitors in zero-shot settings.
- ✓ It reveals that the combination of **model scale, task diversity,** and **instructional format** is the key to true generalization.

NotebookLM