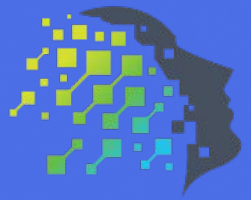
Making, not Taking, The Best-of-N

Ammar Khairi, Daniel D'souza, Marzieh Fadaee, Julia Kreutzer

Cohere, Cohere Labs

ICLR 2026 @ Rio de Janeiro, Brazil



ICLR
International Conference On Learning Representations

Agenda

01 Objective

02 Setup

03 Results

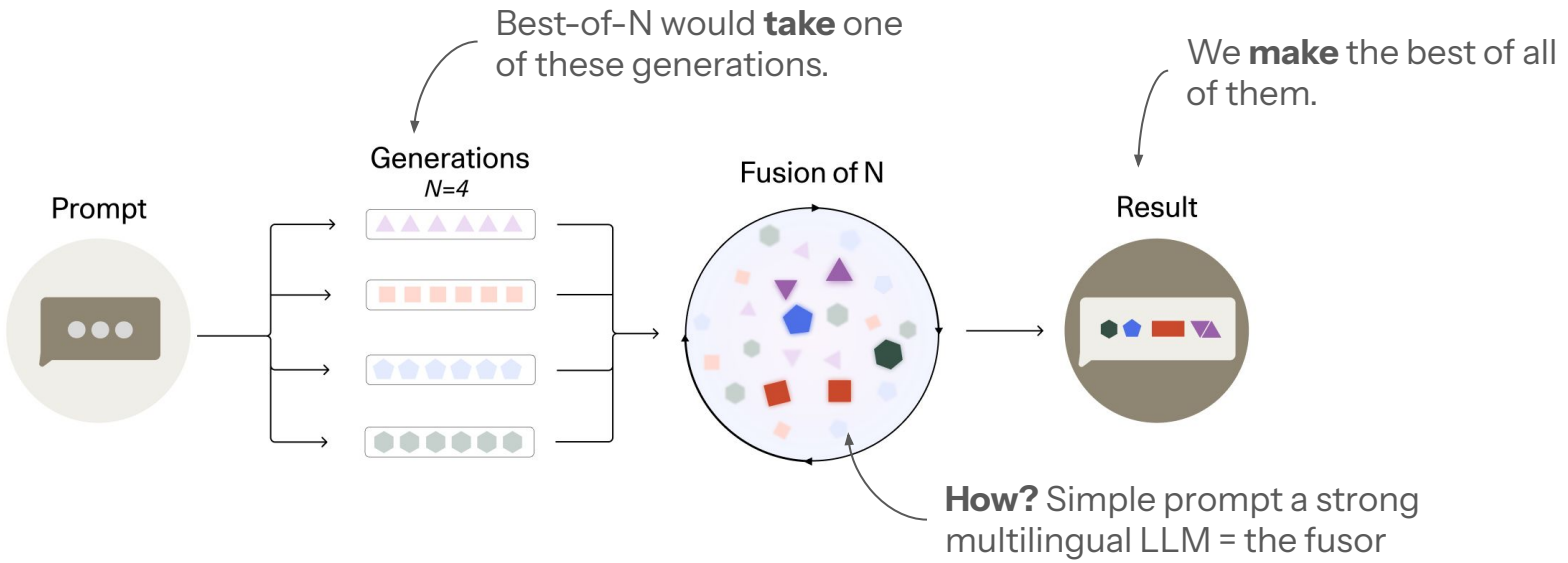
04 Conclusion

01

Objective

Synthesis Over Selection

Ensembling with synthesis instead of selection



Idea: **synthesize** multiple generations into a better one with FusioN.

02

Setup

Comprehensive Evaluation

Experimental Setup

Tasks

Open Ended, Translation, Math, Factual QA, Reasoning Quality

11 Langs.

Arabic, Chinese, English, French, German, Italian, Japanese, Korean, Portuguese, Russian and Spanish

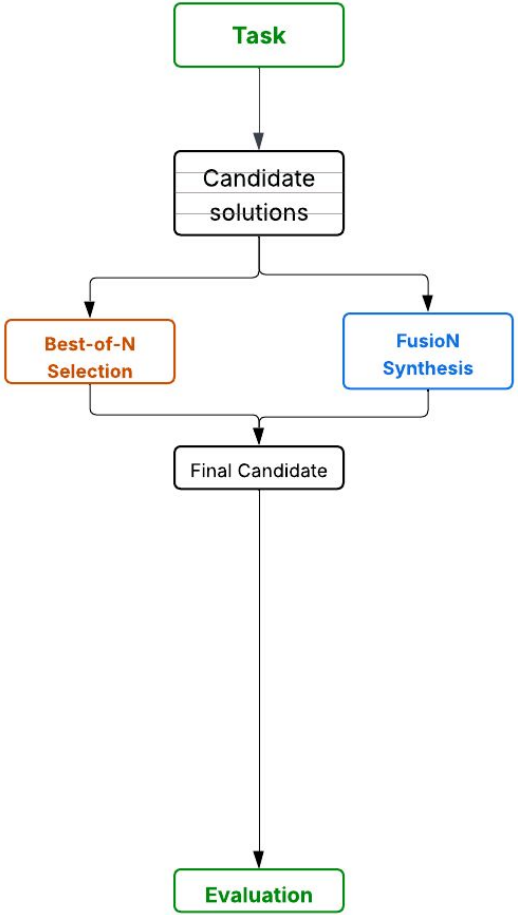
Test Time Scaling

2 Candidate Models: Aya-8b, Command-A
5 Samples per prompt

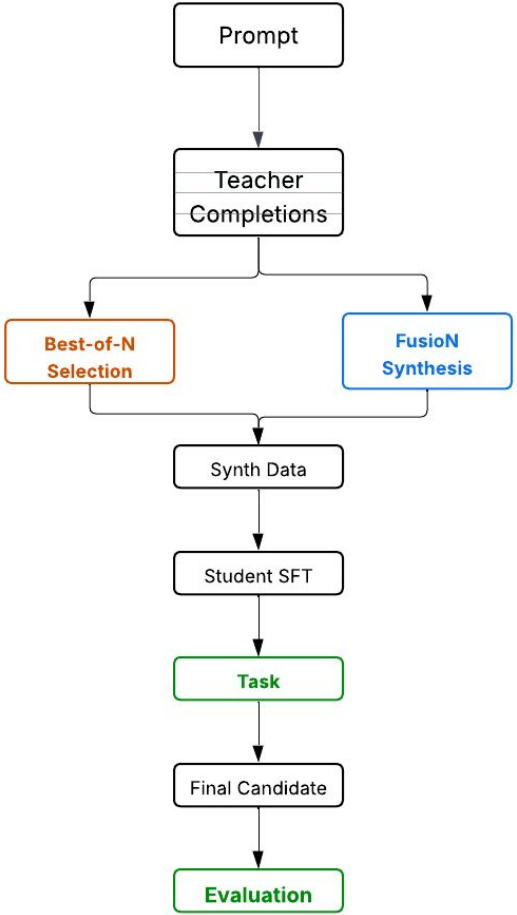
Synth Data Gen.

5 teacher models: Deepseek, Gemma, Qwen, Command, Kimi

Test-Time Scaling

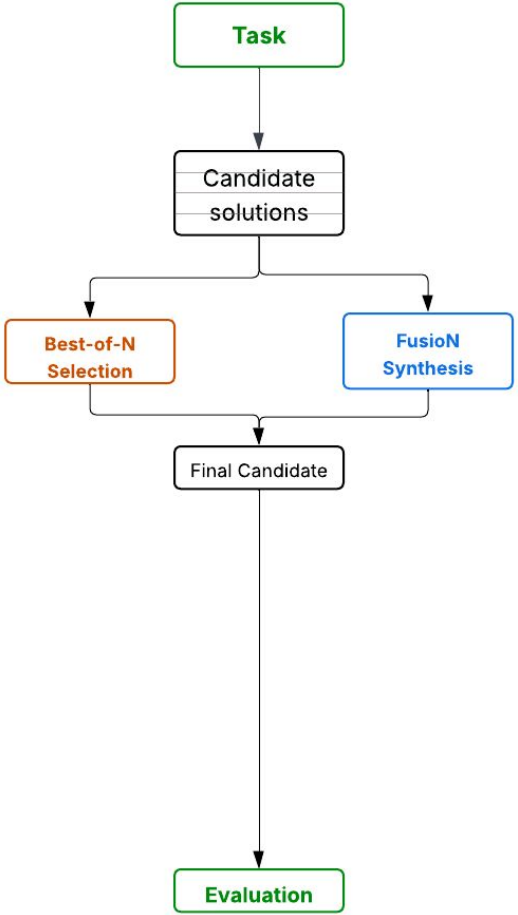


Synth Data Generation

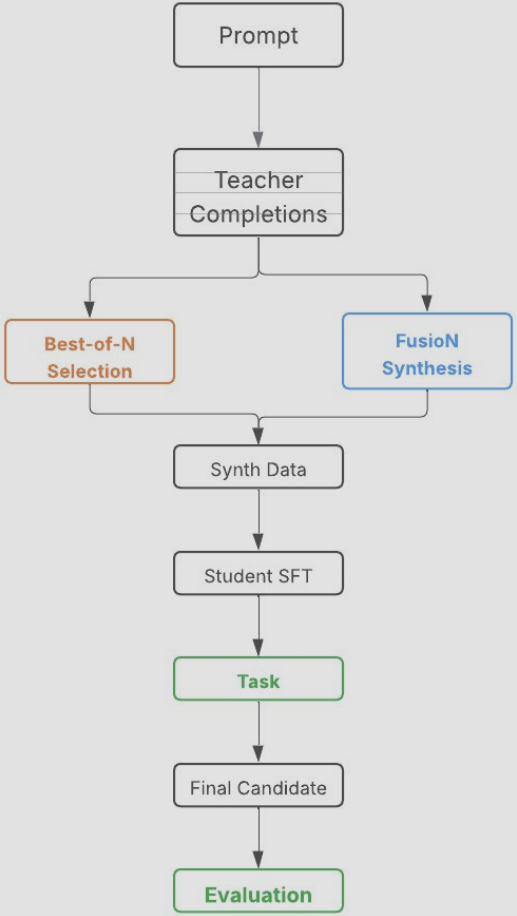


Cohere Labs

Test-Time Scaling

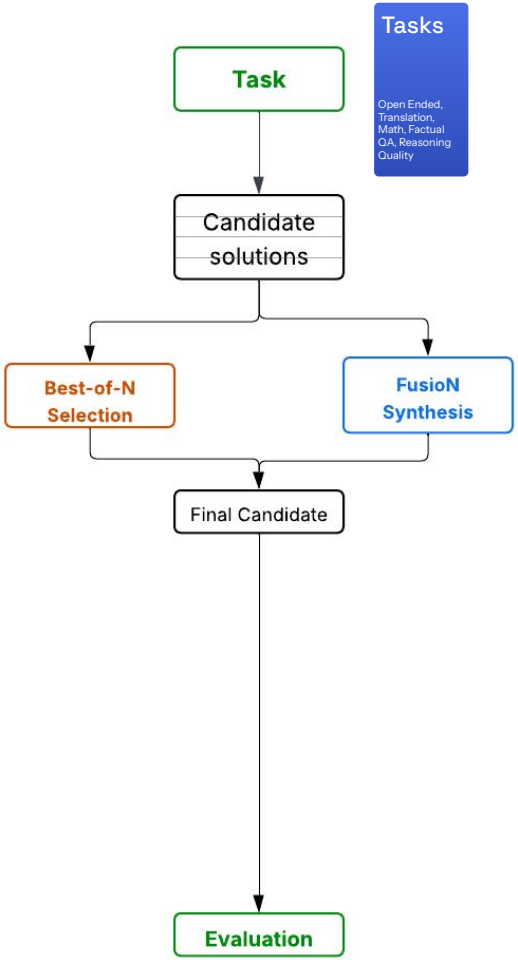


Synth Data Generation

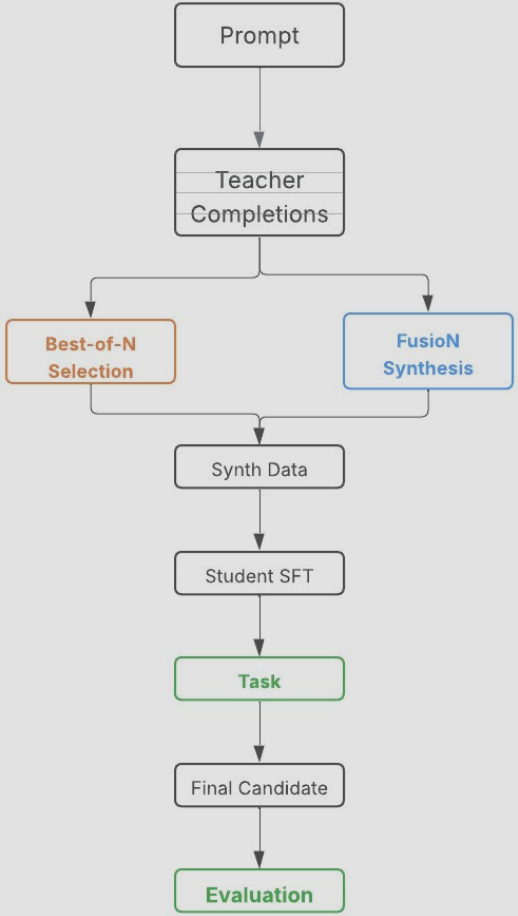


Cohere Labs

Test-Time Scaling

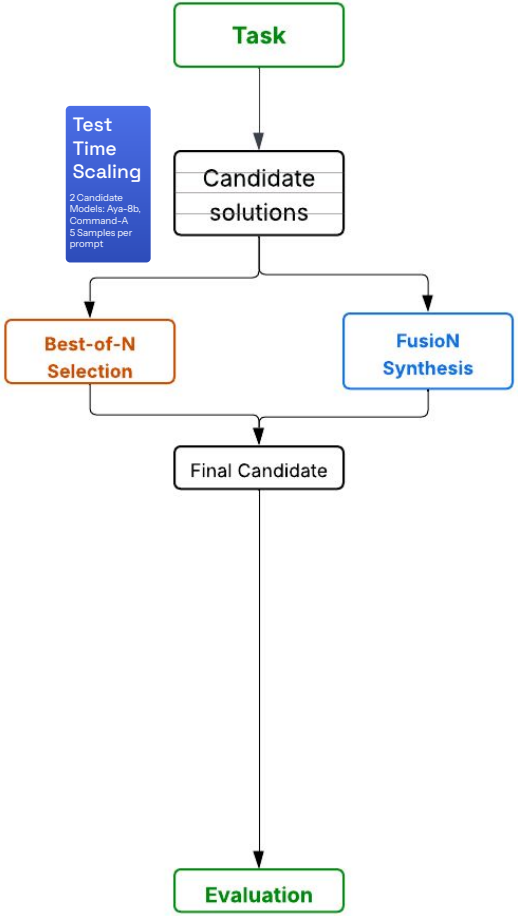


Synth Data Generation

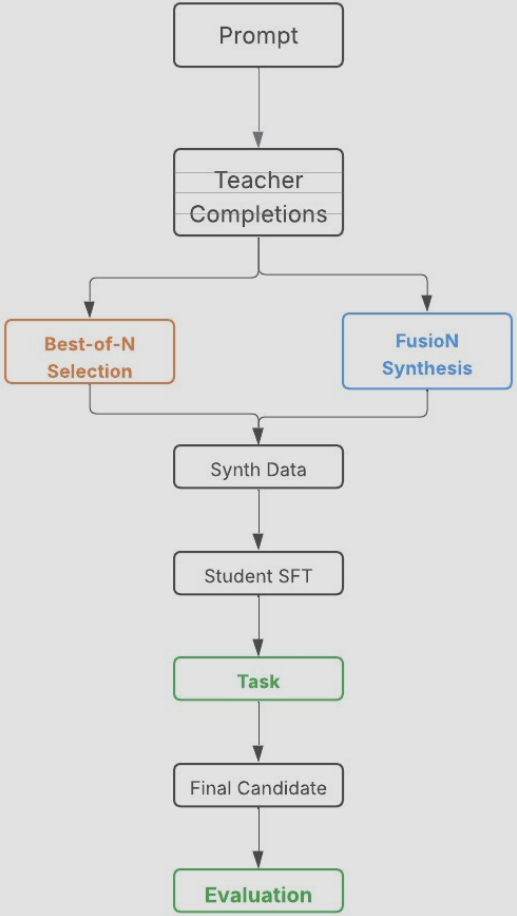


Cohere Labs

Test-Time Scaling

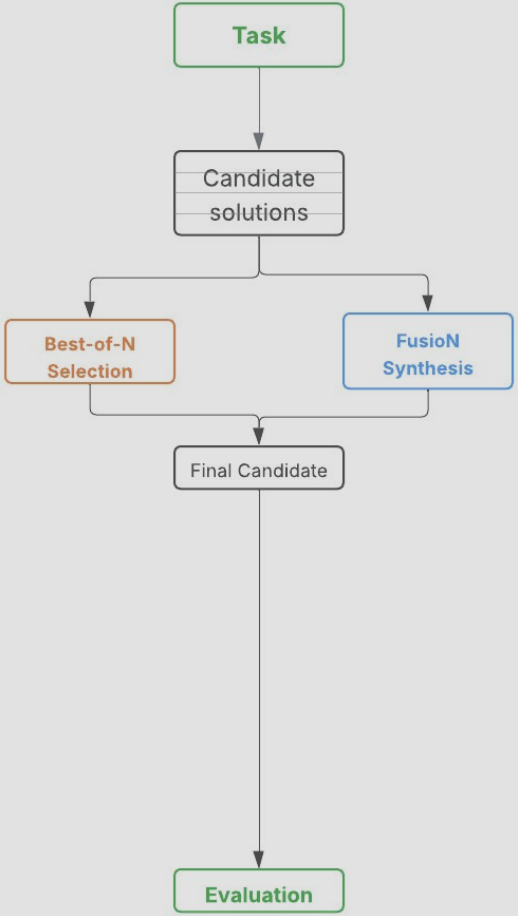


Synth Data Generation



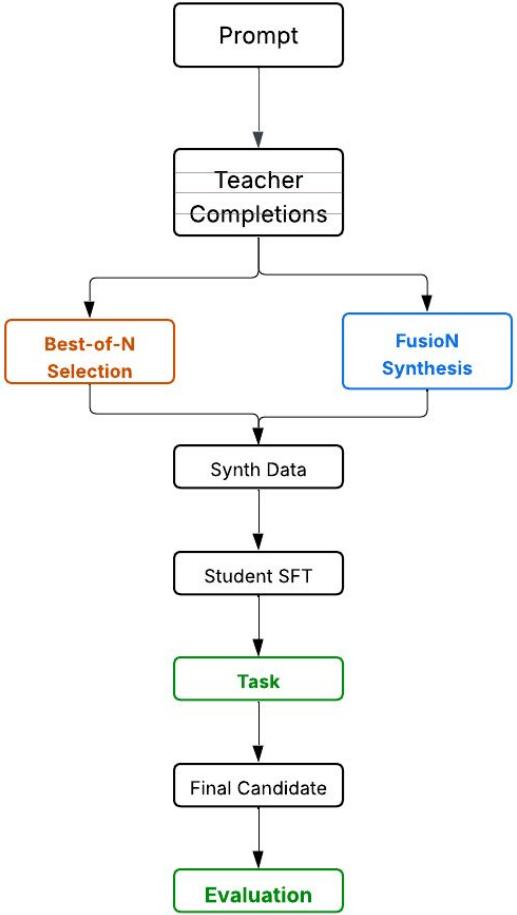
Cohere Labs

Test-Time Scaling

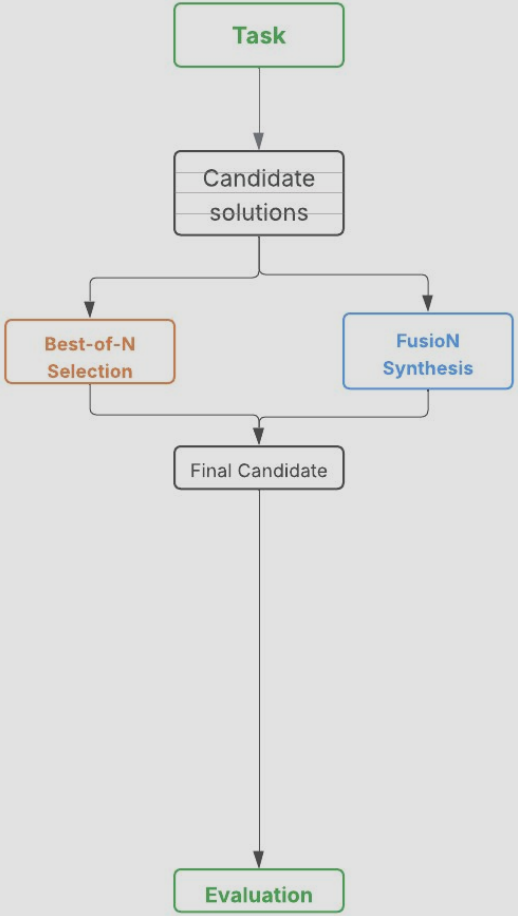


Cohere Labs

Synth Data Generation

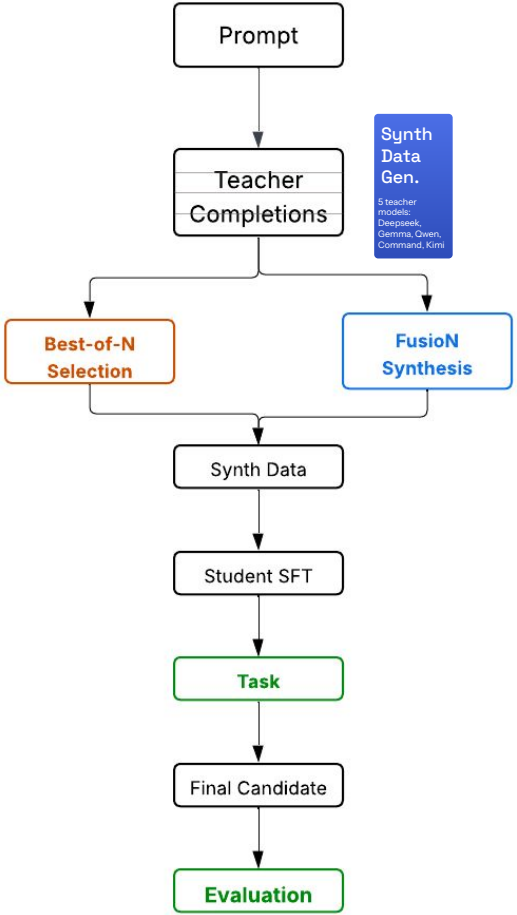


Test-Time Scaling

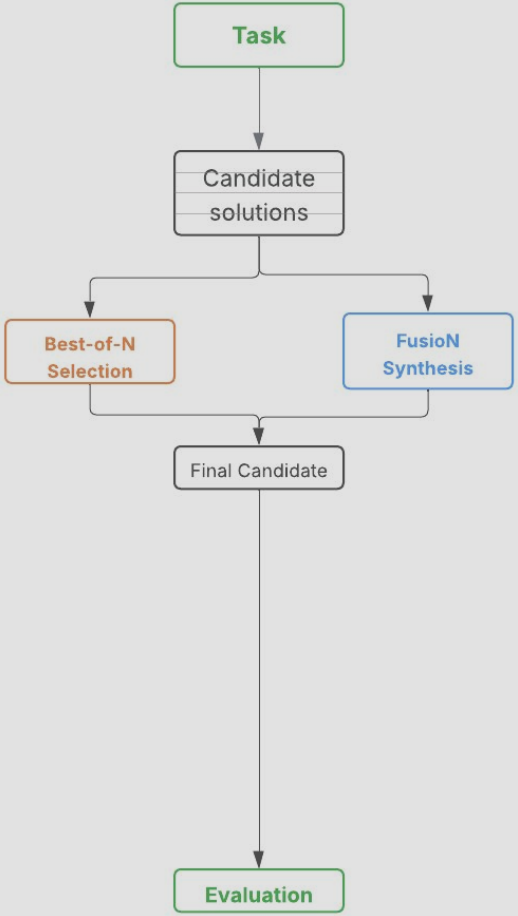


Cohere Labs

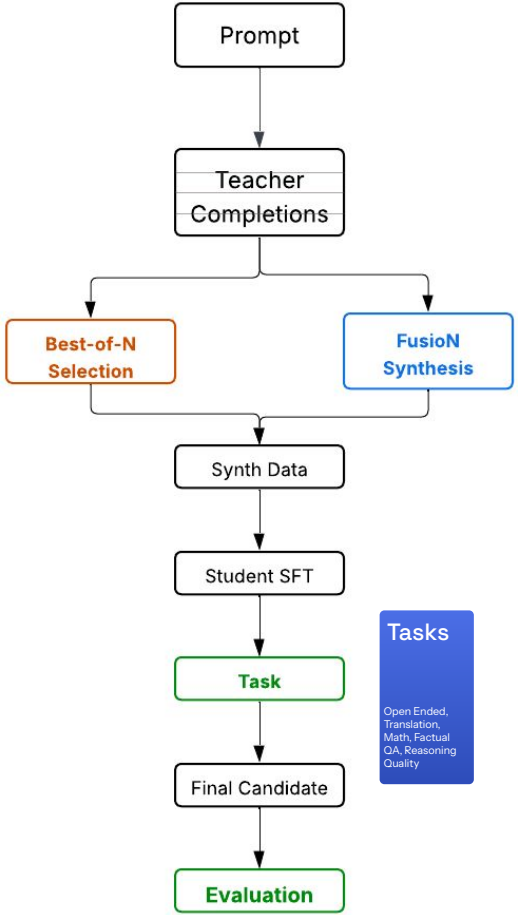
Synth Data Generation



Test-Time Scaling



Synth Data Generation



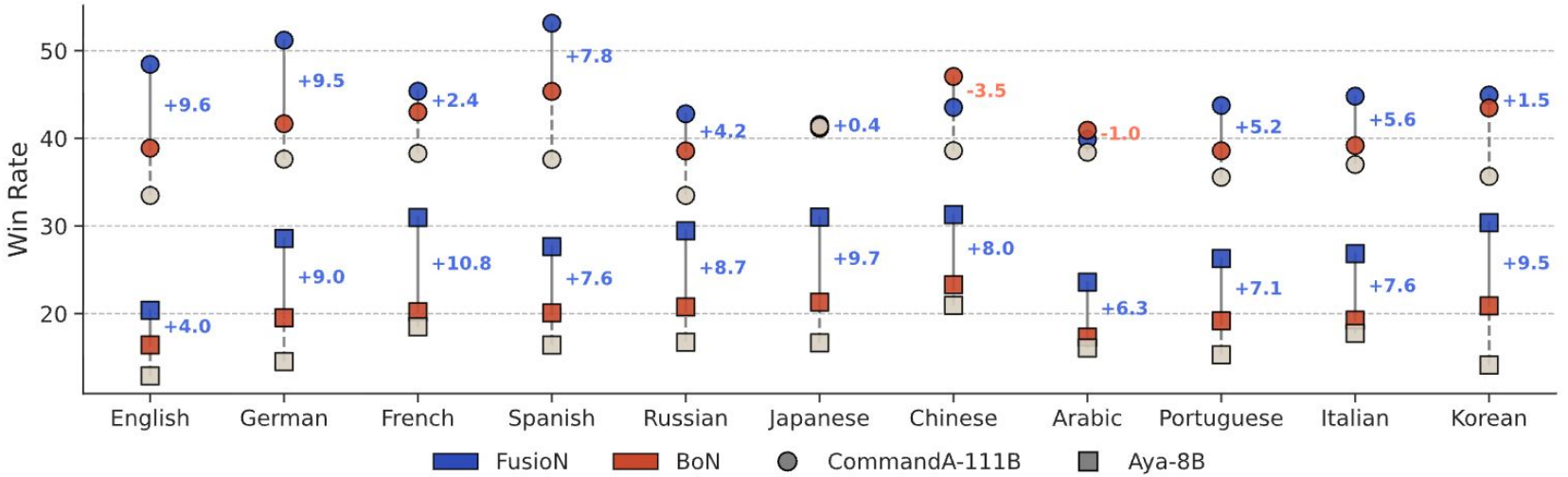
Cohere Labs

03

Results

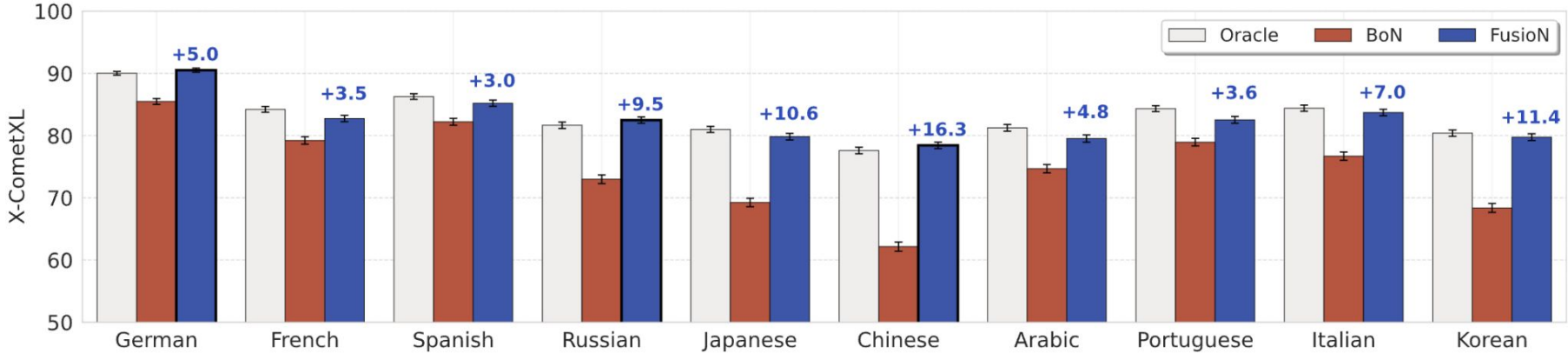
Consistent Gains

Test-Time-Scaling: m-ArenaHard v2 vs Gemini 2.5 Pro



Large impact on generative tasks performance compare to BoN across languages and model sizes.

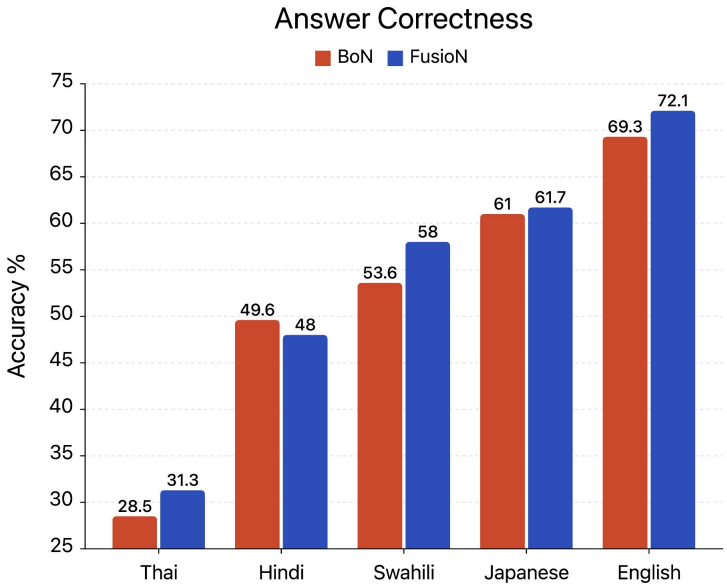
Test-Time-Scaling: Machine Translation on WMT24++

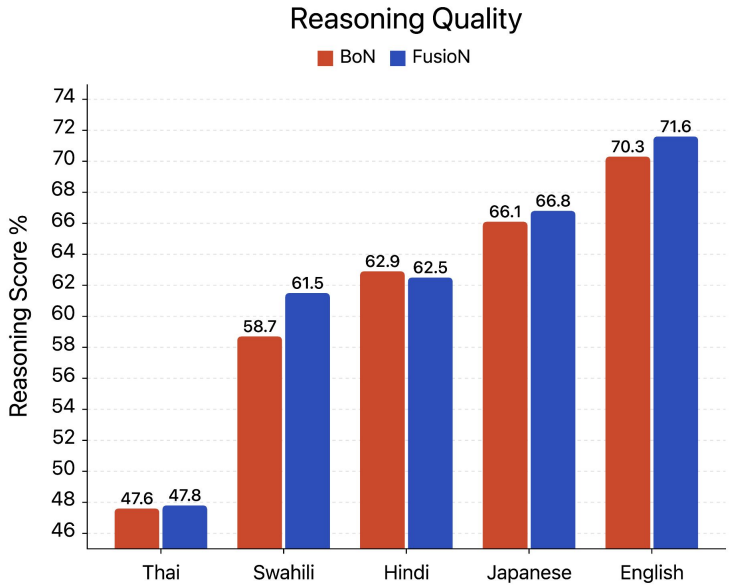


Strong wins over **BoN**, **FusioN** sometimes outperforming the **oracle**.*

*Oracle = pick the sample with the best eval metric score (here: XCometXL).

Synth Data Generation: Factual Reasoning on GeoFactX





Consistent downstream gains in knowledge tasks and reasoning quality.

*Thai and Swahili are ood for the Fusor Model.

04

Conclusion

Replace BoN → FusioN

Findings and Limitations

- **FusioN** consistently outperforms traditional winner-takes-all approaches like **BoN**.
-

Findings and Limitations

- **FusioN** consistently outperforms traditional winner-takes-all approaches like **BoN**.
- FusioN leverages the strengths of multiple models, even when some are weaker, showing robustness and adaptability.
-

Findings and Limitations

- **FusioN** consistently outperforms traditional winner-takes-all approaches like **BoN**.
- FusioN leverages the strengths of multiple models, even when some are weaker, showing robustness and adaptability.
- Much more analysis including robustness ablations, contribution analysis, efficiency tradeoffs ...

Findings and Limitations

- **FusioN** consistently outperforms traditional winner-takes-all approaches like **BoN**.
- FusioN leverages the strengths of multiple models, even when some are weaker, showing robustness and adaptability.
- Much more analysis including robustness ablations, contribution analysis, efficiency tradeoffs ...

Full Paper

Datasets & Evals

Credits to the team!

Daniel D'souza

Julia Kreutzer

Marzieh Fadaee

And Thank YOU all for listening !

Cohere Labs