FiD-ICL: A Fusion-in-Decoder Approach for Efficient In-Context Learning





Qinyuan Ye Iz Beltagy Matthew E. Peters Xiang Ren Hannaneh Hajishirzi

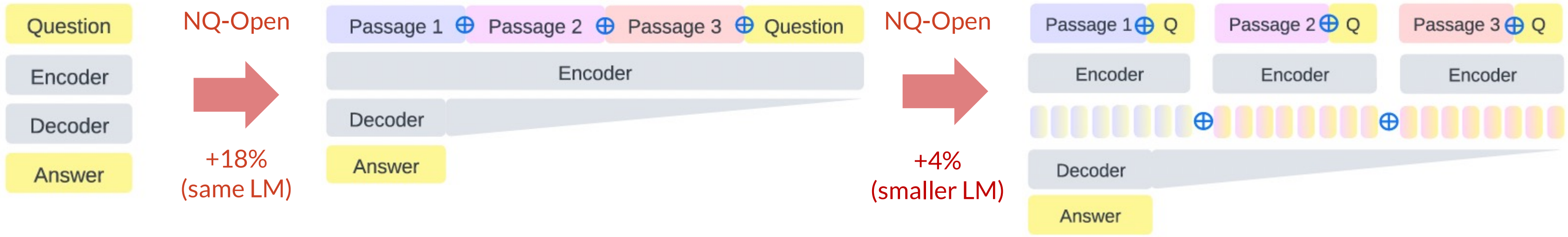
USC University of Southern California
Ai2

Background: QA vs. ICL

Closed-book QA
(Roberts et al., 2020)

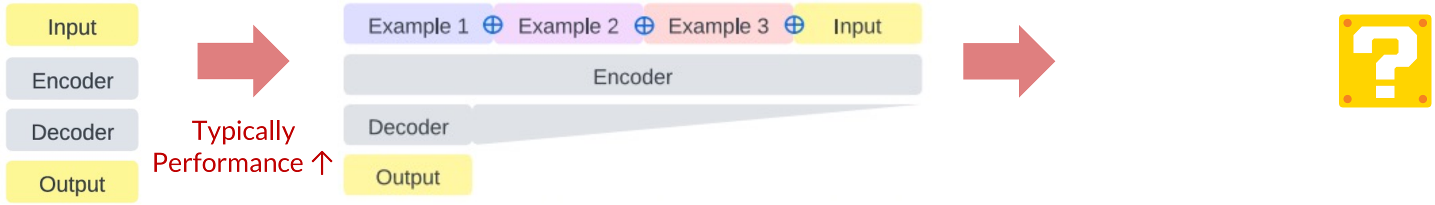
Retrieval-Augmented Generation
(Lewis et al., 2020)

Fusion-in-Decoder
(Izacard et al., 2020)



Zero-shot Learning

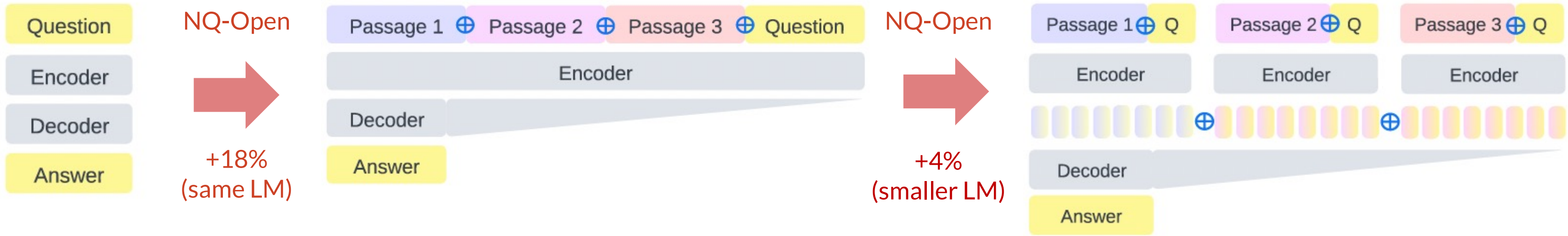
Few-shot In-Context Learning



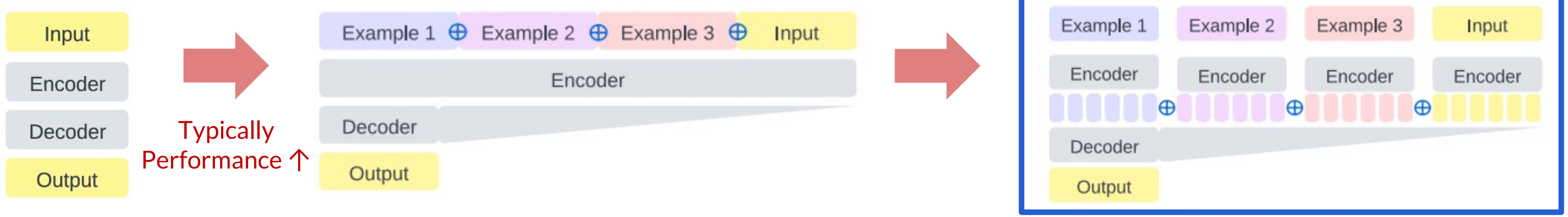
USC University of Southern California
AI2

Background: QA vs. ICL

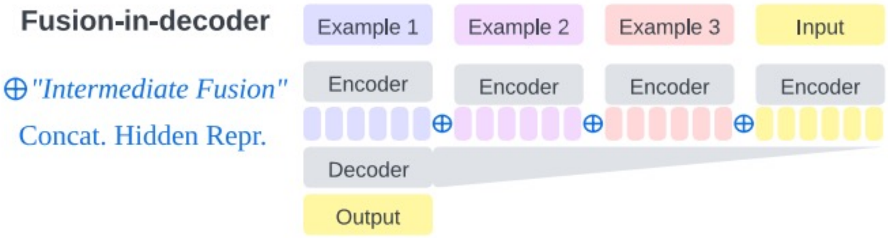
Closed-book QA (Roberts et al., 2020) · **Retrieval-Augmented Generation** (Lewis et al., 2020) · **Fusion-in-Decoder** (Izacard et al., 2020)



Zero-shot Learning · **Few-shot In-Context Learning** · **This work: FiD-ICL**

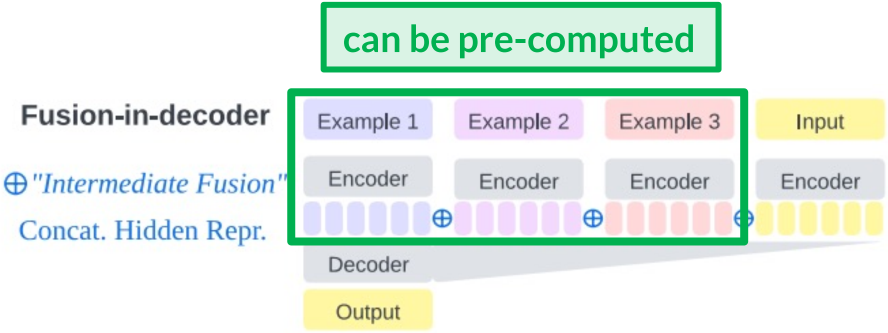


FiD-ICL



* *We compared two more FiD variations in the appendix.*

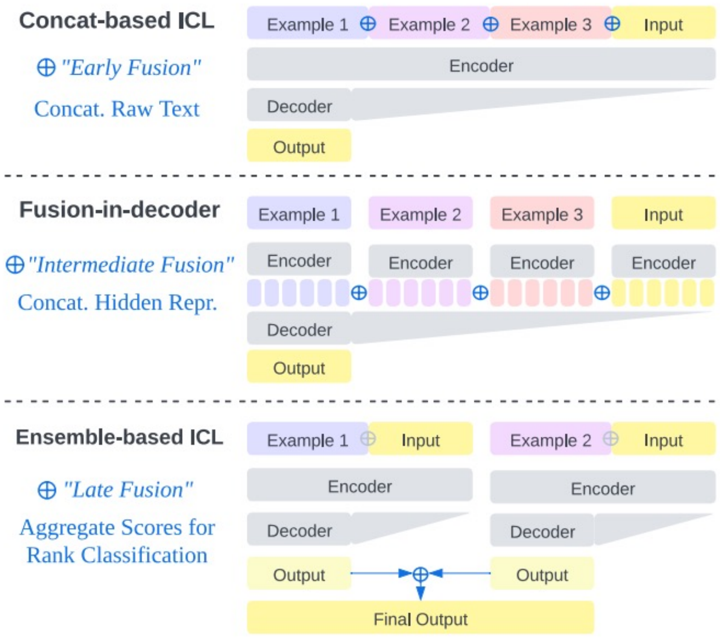
FiD-ICL



* *We compared two more FiD variations in the appendix.*

Compared Methods

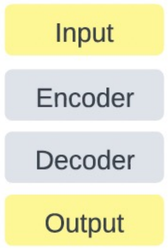
Referred to as “fusion” methods for ICL



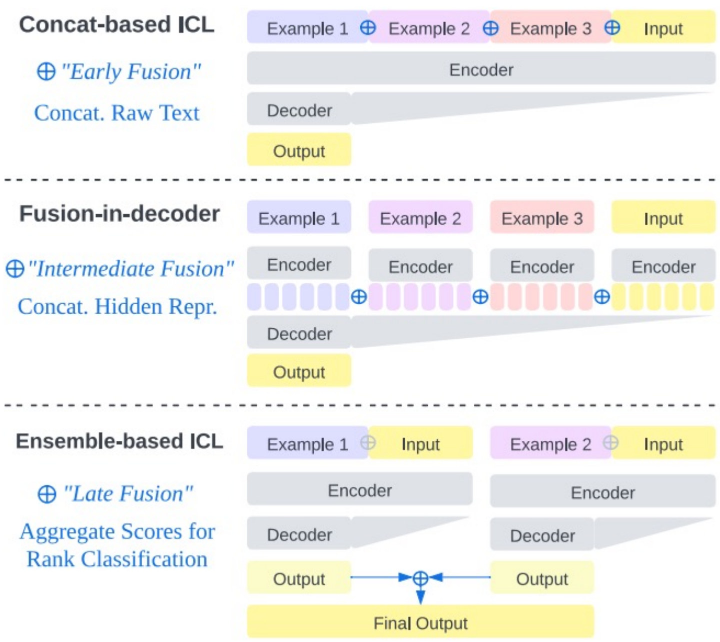
USC University of Southern California

Compared Methods

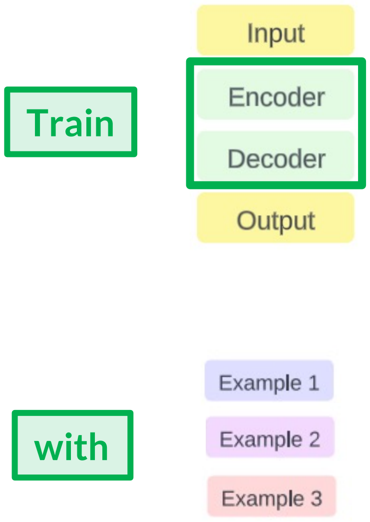
Zero-shot Learning



Few-shot In-Context Learning

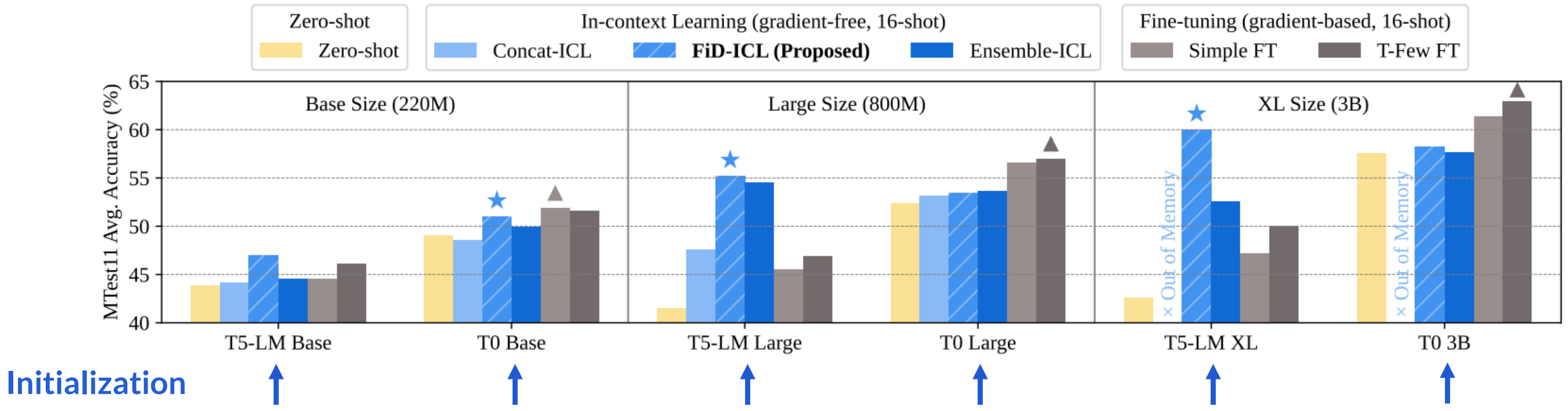


Fine-tuning



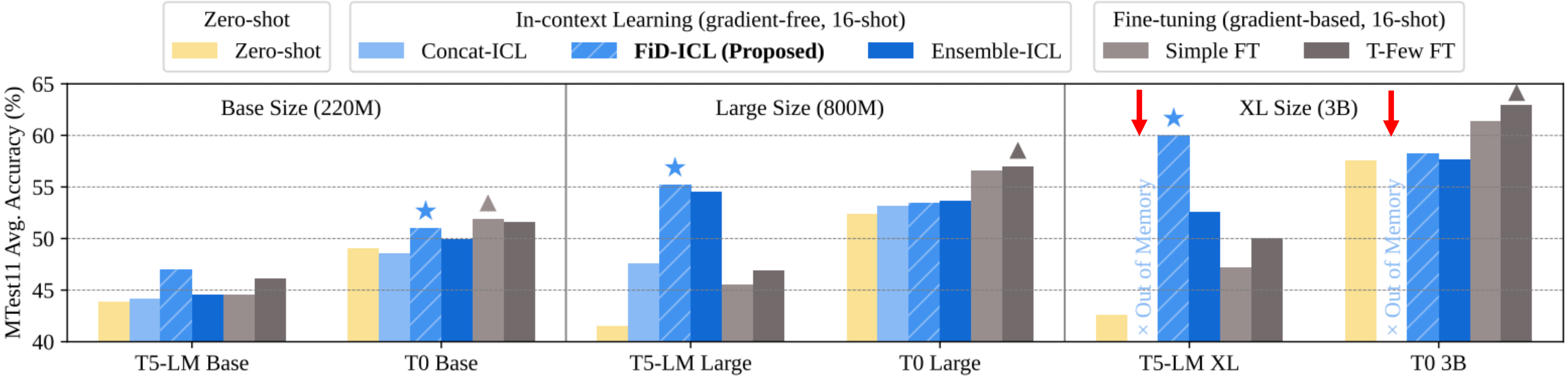
Main Results

Using *Public Pool of Prompts (P3)* dataset
Using a *meta-training* setting



USC University of Southern California

Main Results

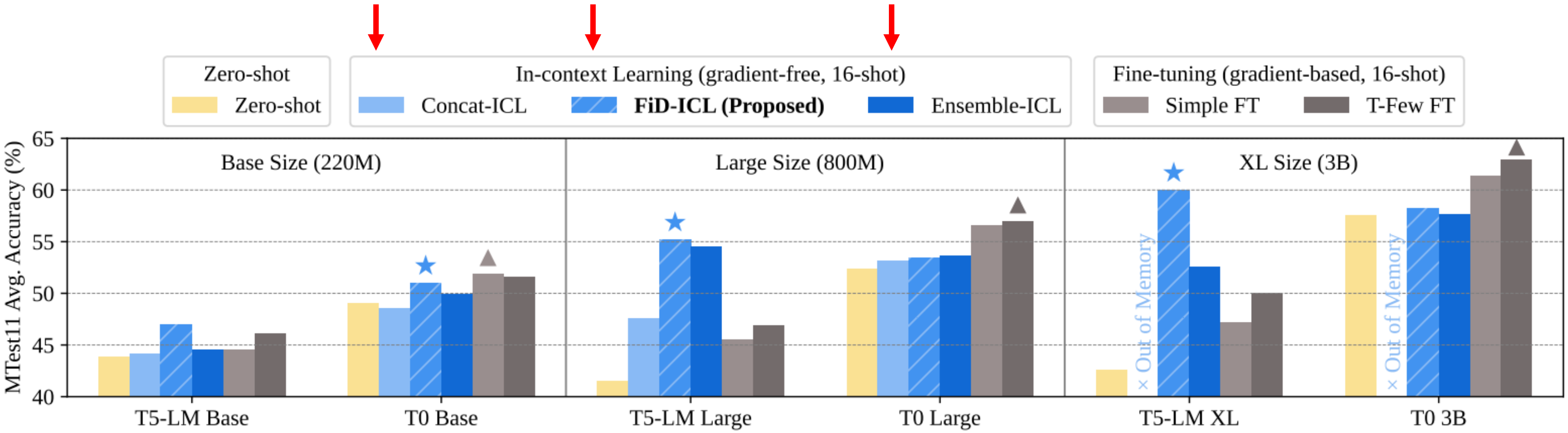


FiD-ICL enables efficient meta-training (Concat-ICL would fail at 3B)

USC University of Southern California

Ai2

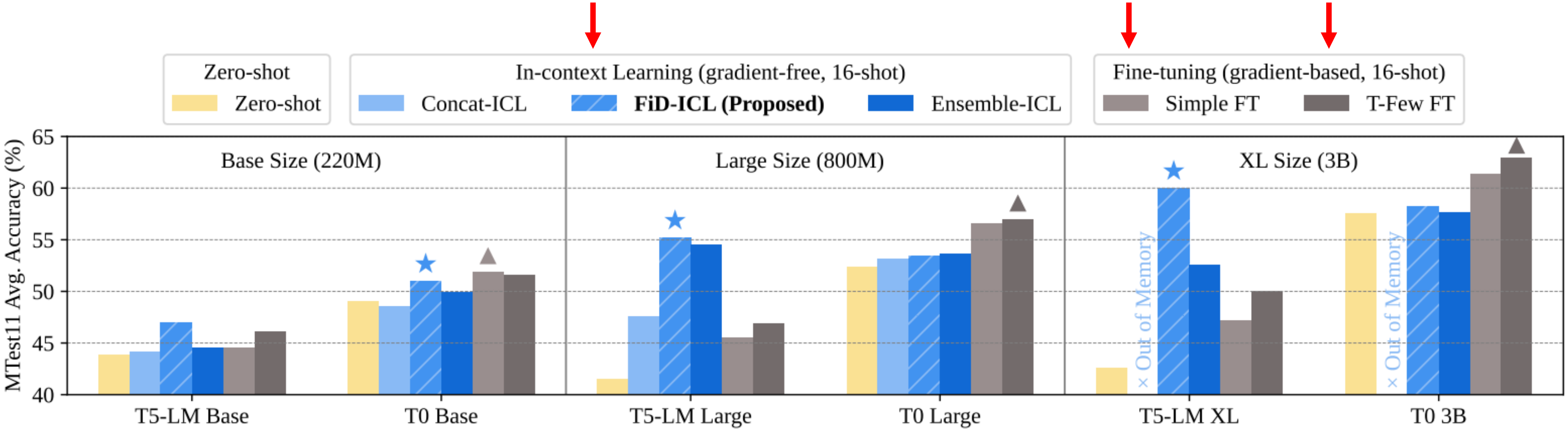
Main Results



FiD-ICL outperforms the other two fusion methods (Concat and Ensemble)

USC University of Southern California AI2

Main Results



The gap between FiD-ICL (★ gradient-free) and fine-tuning (▲ gradient-based) is <3%.

Efficiency

Zero-shot	In-context Learning (gradient-free, 16-shot)			Fine-tuning (gradient-based, 16-shot)	
Zero-shot	Concat-ICL	**FiD-ICL (Proposed)**	Ensemble-ICL	Simple FT	T-Few FT

Runtime (Pre-inference + Inference)

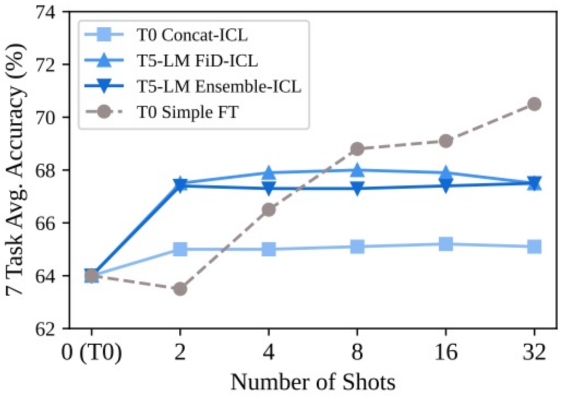
	Zero-shot	Concat-ICL	FiD-ICL (Proposed)	Ensemble-ICL	Simple FT	T-Few FT
RTE	1x	17x	**2x**	26x	26x	>26x
StoryCloze	1x	27x	**3x**	29x	6x	>6x

FiD-ICL is computationally efficient.

** Limitations apply. Fine-tuned models are still more efficient for large-scale inference.*

Analysis (or... surprise?)

Number of Shots

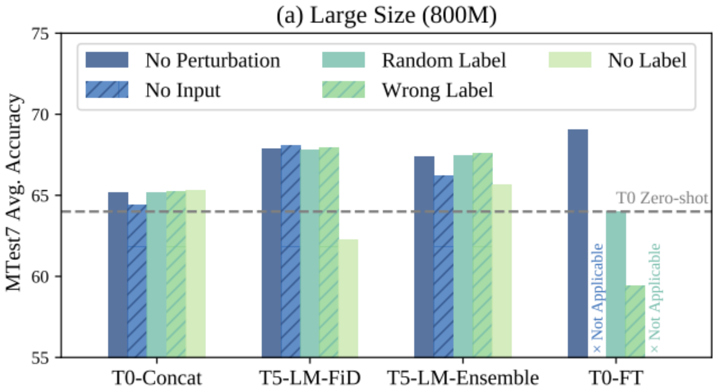


Average performance ***does not*** grow with more shots.

It's ***task-dependent***.

Perturbation to In-context Examples

(Inspired by Min et al., 2022)



Performance is rather ***insensitive*** to perturbations to in-context examples.

Still ***not*** learning effectively.

Conclusion

FiD-ICL, a fusion-in-decoder approach for efficient in-context learning

Performance

It outperforms Concat-ICL and Ensemble-ICL.
The gap between FiD-ICL and fine-tuning is ***<3%*** on P3 meta-test tasks.

Efficiency

FiD-ICL is more efficient than Concat-ICL, Ensemble-ICL.
More efficient than fine-tuning when considering pre-inference + inference time*.

Limitations

FiD-ICL is still ***not perfect***; still has the similar limitations as Concat-ICL.

Implications

Insights and methodologies from ***open-domain QA*** are very useful!
FiD-ICL is related to ***retrieval augmentation***, ***sparse attention***, and ***hypernetworks***.

** Limitations apply. Fine-tuned models are still more efficient for large-scale inference.*