Universität Stuttgart
Institut für Maschinelle Sprachverarbeitung

SS24
Advanced
Deep
Learning

Tana Deeg
Momo Takamatsu
Sarthak Singh

Improving QA Scenarios using Prompting

Motivation

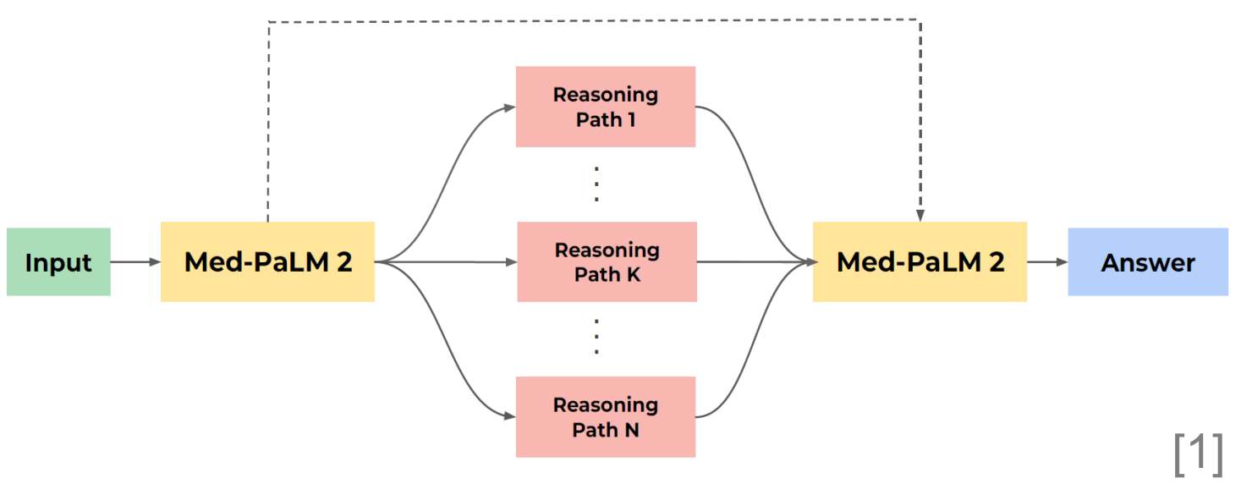
Tuning large models can be expensive. Prompting is an efficient method to enhance the model's performance without the cost involved.

Which zero- & few-shot **QA scenarios** can be improved using **prompting**, and **how**?

Improved QA Tasks

Clinical Decision Support

Prompting to improve **medical QA**

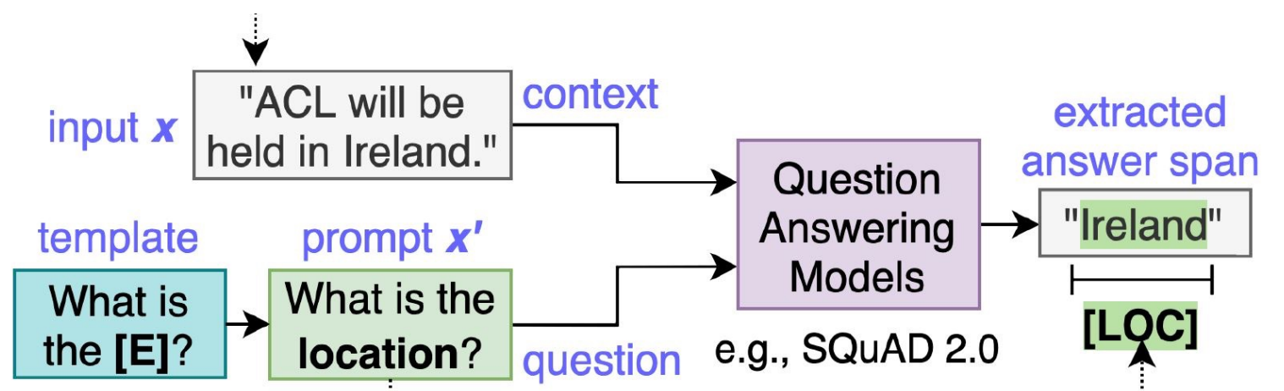


[1]

Prompting strategy: Ensemble Refinement

Named Entity Recognition

Prompting to improve **QA for NER**

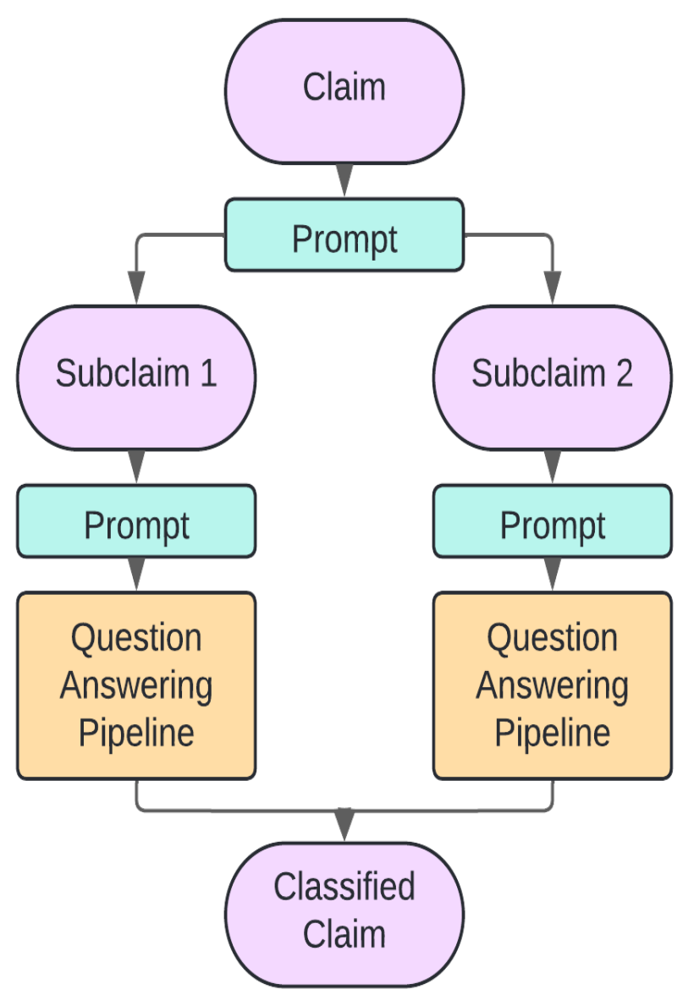


[2]

Fact Verification

Prompting to improve **QA for fact verification**

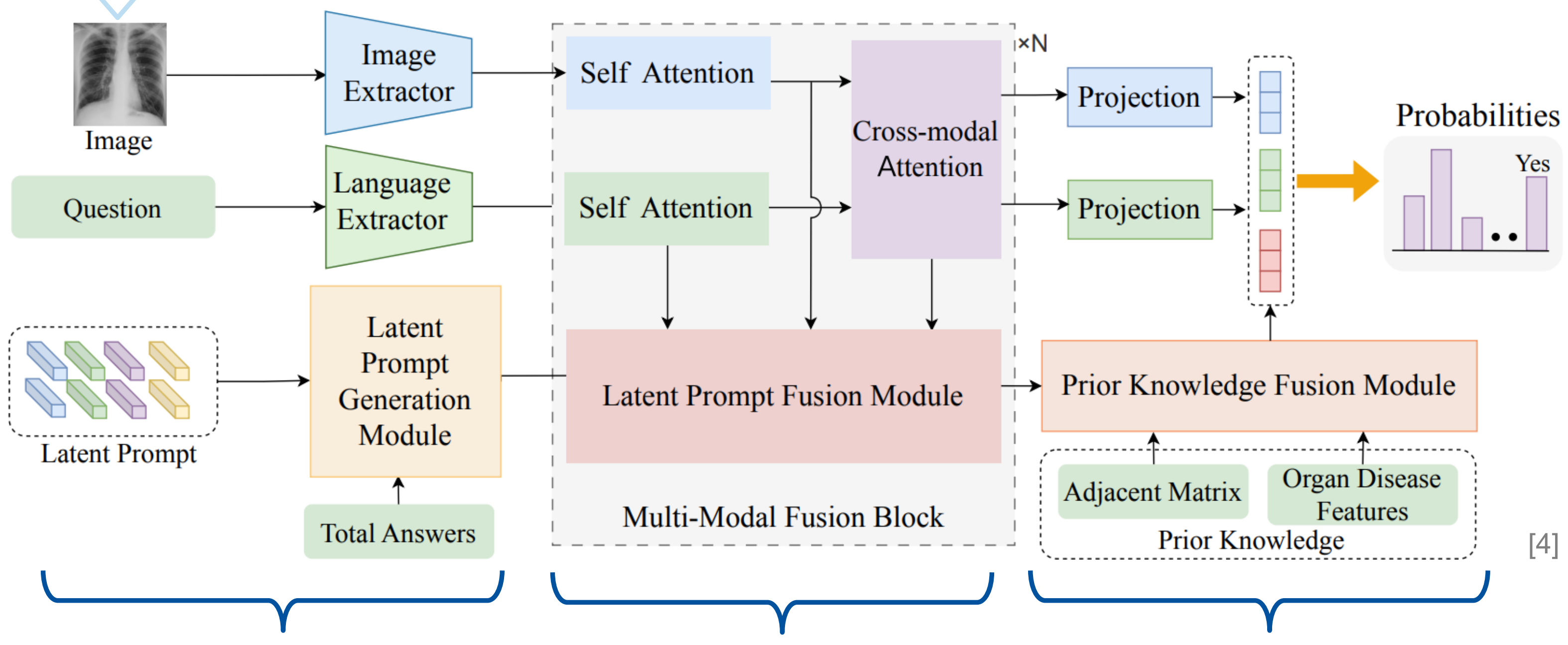
Hierarchical Step-by-Step Prompting



Prompting Techniques

Latent Prompt Assist: LaPA

VQA-RAD, SLAKE, VQA-2019



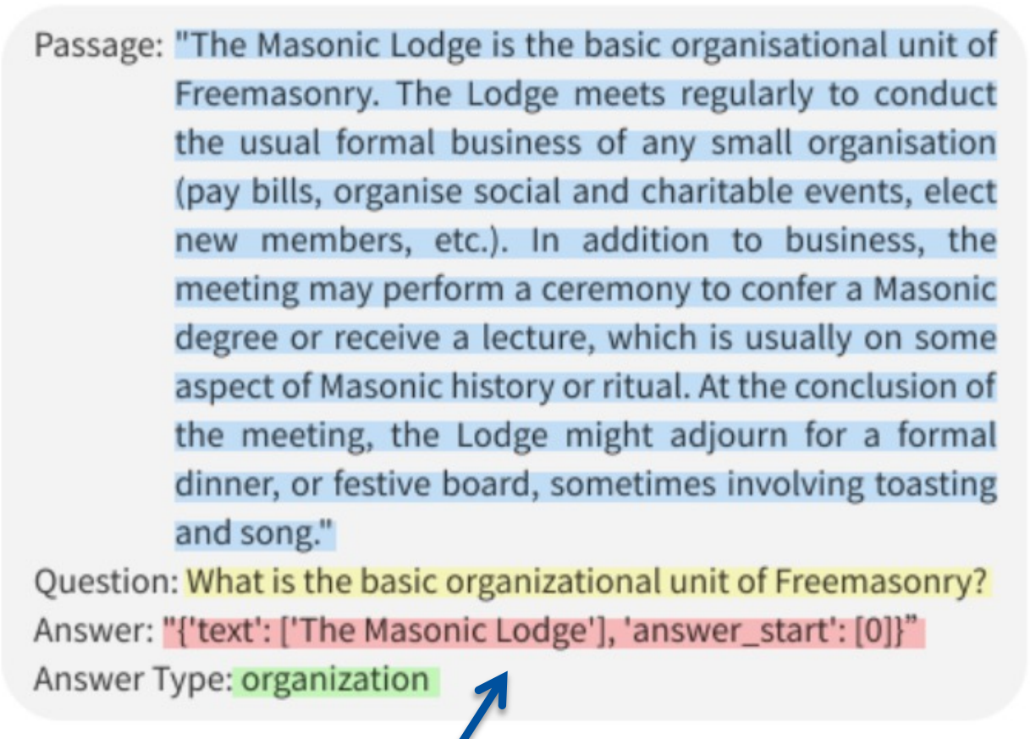
[4]

Generation of learnable latent prompt (LP) to assist model

Extraction of clinically relevant information with LP

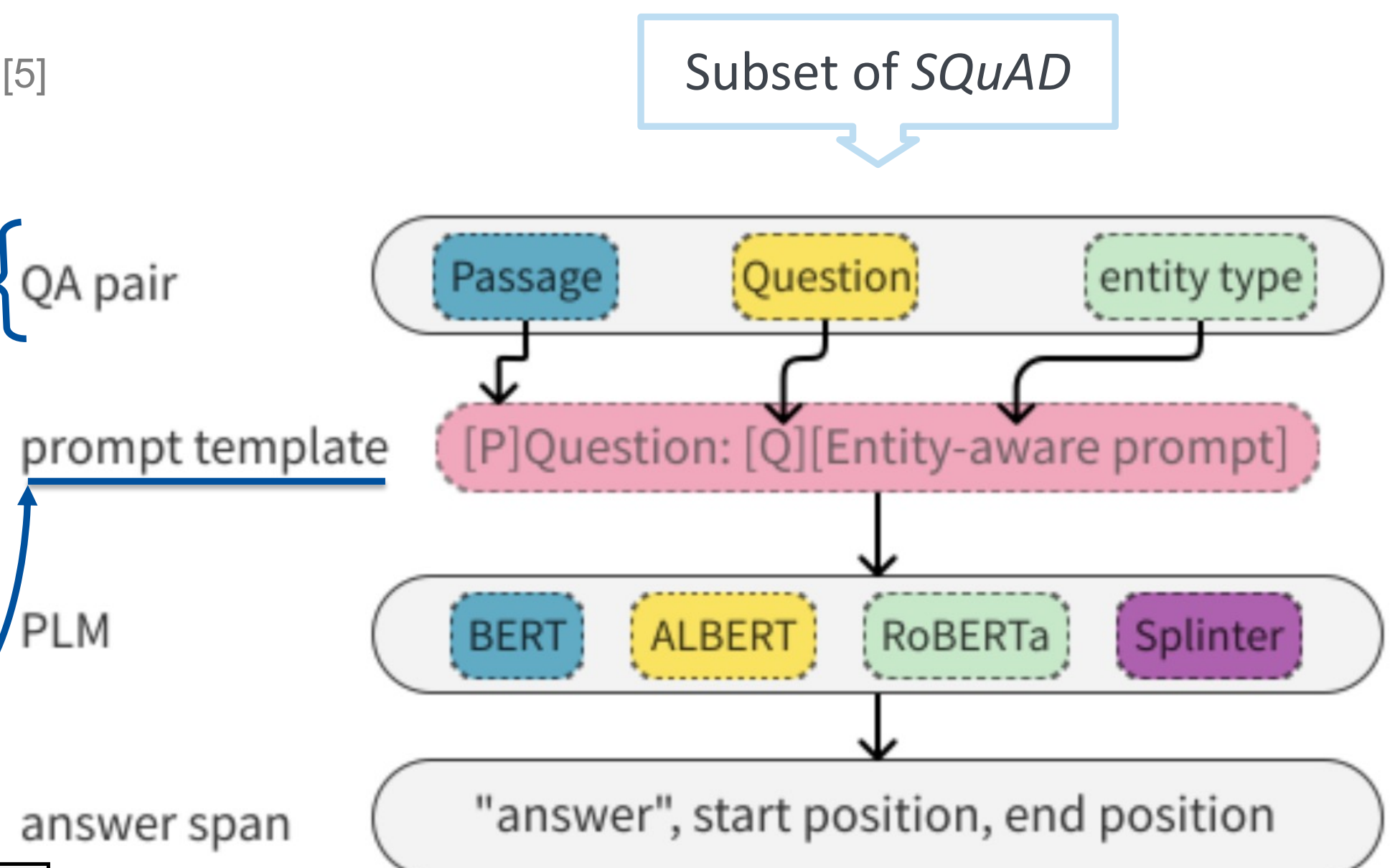
Capturing relationships between organ and disease

Entity-aware Prompting


Passage: "The Masonic Lodge is the basic organisational unit of Freemasonry. The Lodge meets regularly to conduct the usual formal business of any small organisation (pay bills, organise social and charitable events, elect new members, etc.). In addition to business, the meeting may perform a ceremony to confer a Masonic degree or receive a lecture, which is usually on some aspect of Masonic history or ritual. At the conclusion of the meeting, the Lodge might adjourn for a formal dinner, or festive board, sometimes involving toasting and song."
Question: What is the basic organizational unit of Freemasonry?
Answer: "{'text': ['The Masonic Lodge'], 'answer_start': [0]}"
Answer Type: organization

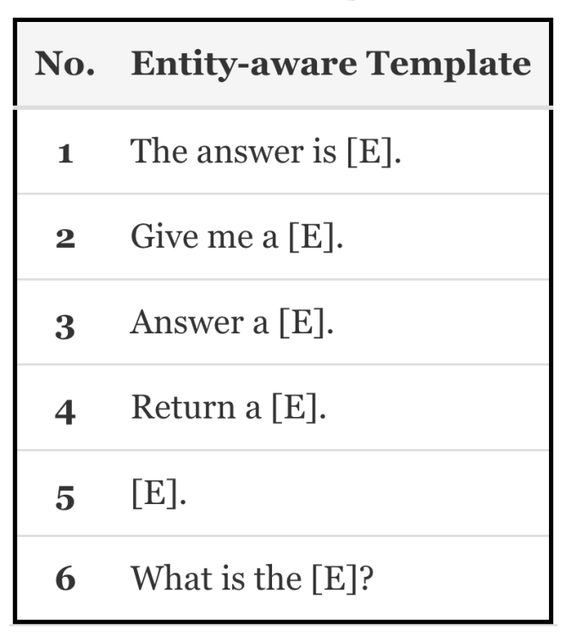
[5]

Rule matching:
Mapping of question word to entity type



[5]

person, place, date, organization, other



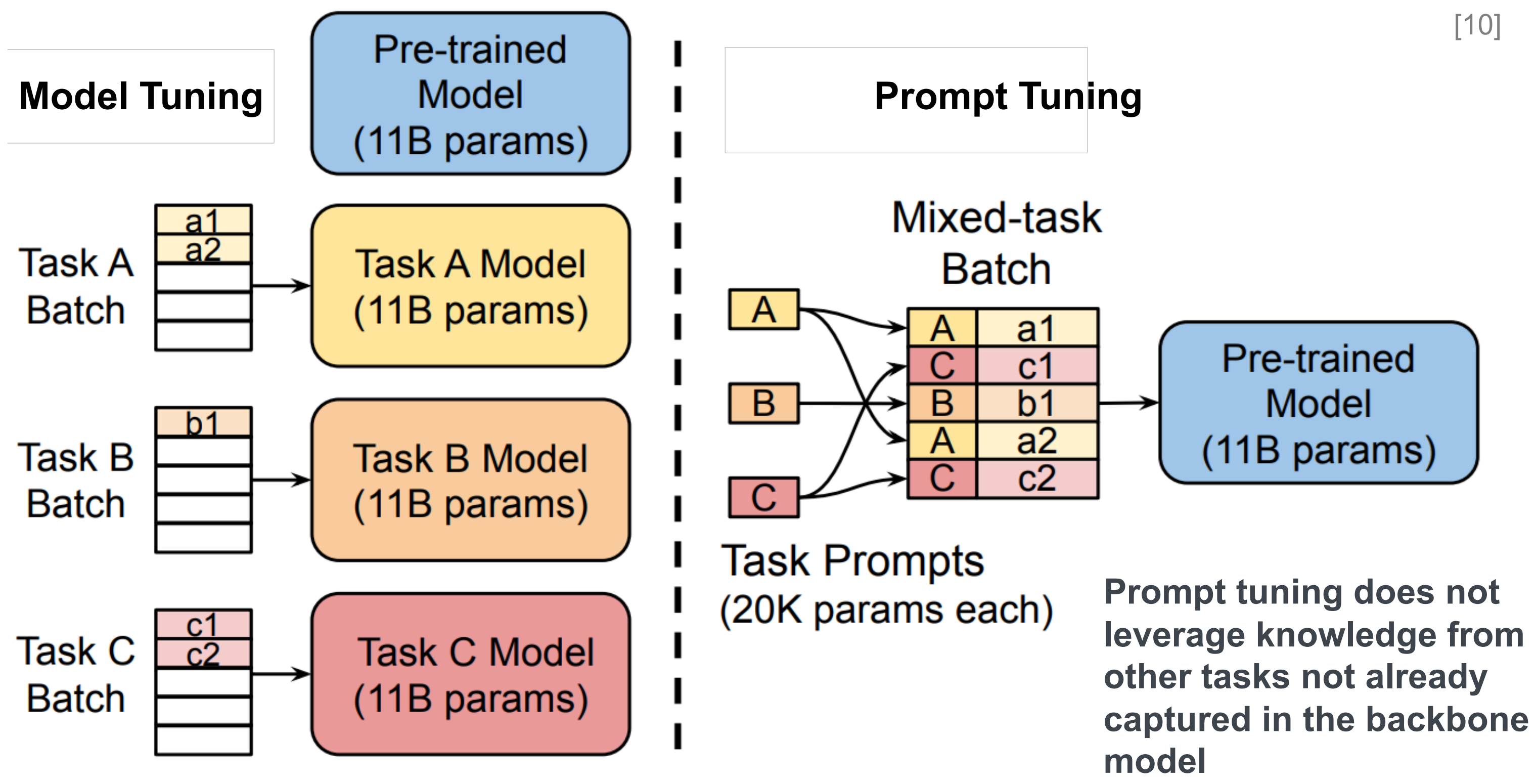
No.	Entity-aware Template
1	The answer is [E].
2	Give me a [E].
3	Answer a [E].
4	Return a [E].
5	[E].
6	What is the [E]?

[5]

Prompt Tuning

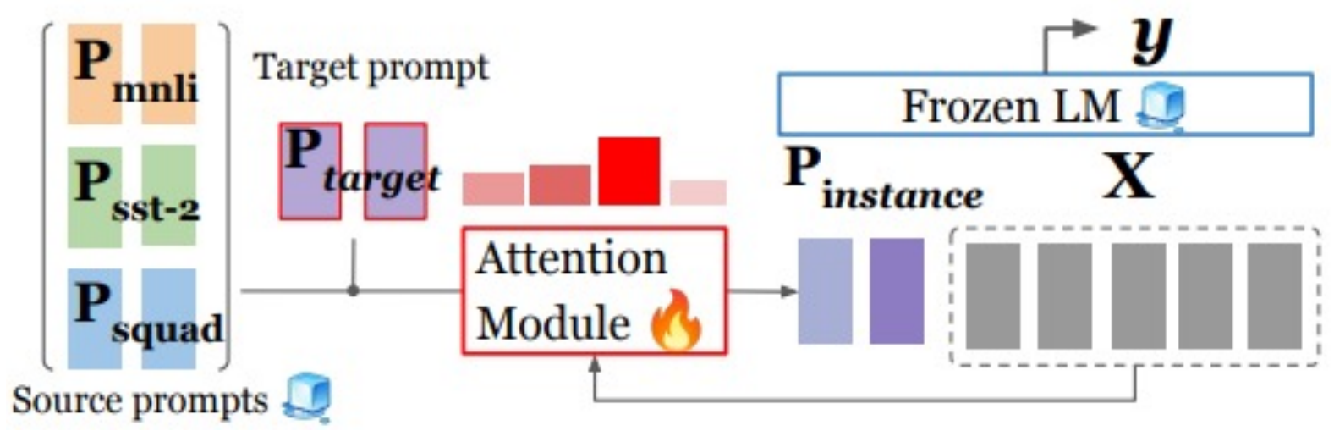
Unified QA by Prompt Tuning

[10]


Model Tuning
Pre-trained Model (11B params)
Task A Batch
a1
a2
Task A Model (11B params)
Task B Batch
b1
Task B Model (11B params)
Task C Batch
c1
c2
Task C Model (11B params)
Prompt Tuning
Mixed-task Batch
A
B
C
A a1
C c1
B b1
A a2
C c2
Pre-trained Model (11B params)
Task Prompts (20K params each)

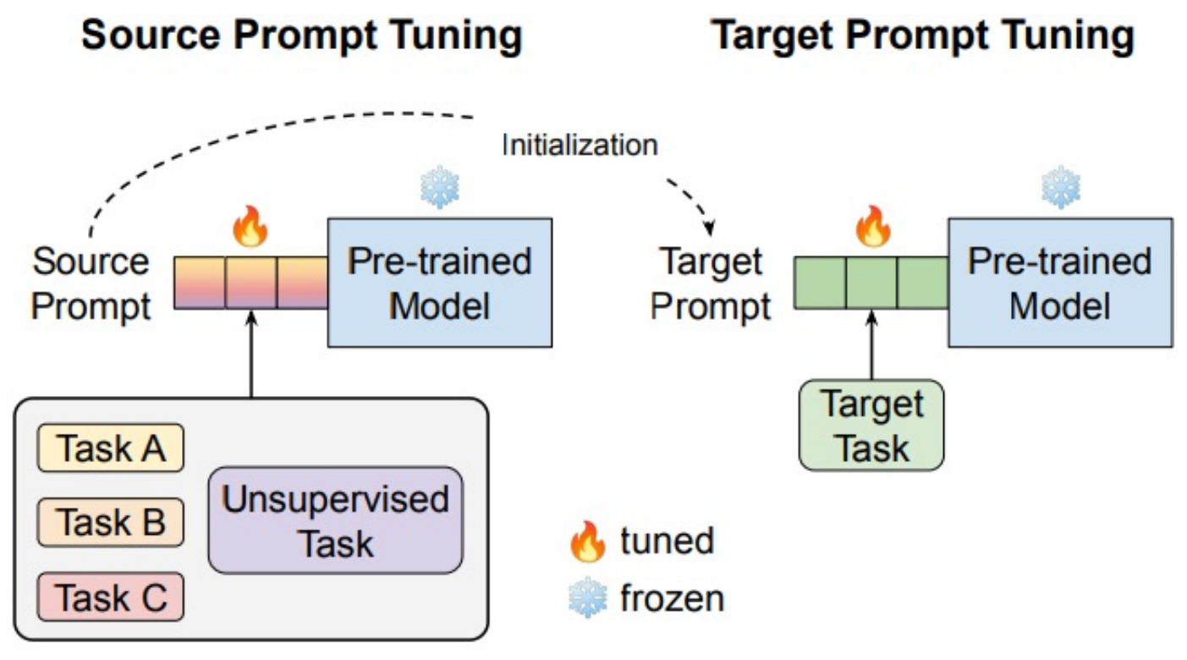
Prompt tuning does not leverage knowledge from other tasks not already captured in the backbone model

Knowledge Transfer



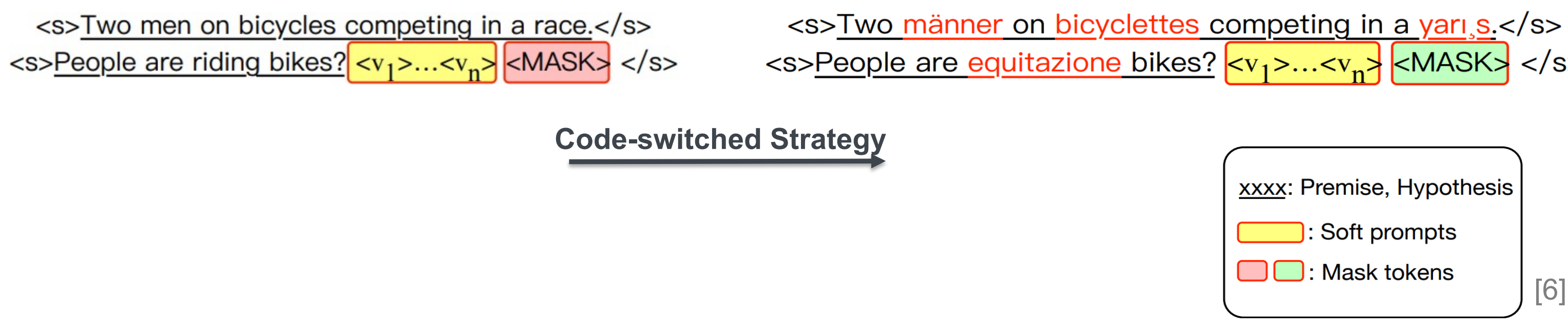
[9]

Prompt Initialization



[8]

Enhancing XNLI by Soft Prompting


<s>Two men on bicycles competing in a race.</s>
<s>People are riding bikes? <v1>...<vn> <MASK> </s>
Code-switched Strategy
<s>Two männer on bicyclettes competing in a yarış.</s>
<s>People are equitazione bikes? <v1>...<vn> <MASK> </s>
xxxx: Premise, Hypothesis
: Soft prompts
: Mask tokens

[6]

Conclusion

Benefiting Tasks:

- Clinical Decision Support (Medical QA)
- Named Entity Recognition
- Fact Verification

Techniques:

- Latent Prompts
- Prompt Templates and Rule Matching

Advantages of Prompt Tuning:

- Efficiency: Fewer parameter adjustments needed
- Generalization: Better across tasks and languages

Models with prompting outperform promptless state-of-the-art models for various tasks [1-3] with various techniques [4, 5] and show more efficiency in few- and zero-shot scenarios [6-10].

References:

1. Singhal, Karan, et al. "Towards expert-level medical question answering with large language models." *arXiv preprint arXiv:2305.09617* (2023).
2. Liu, Andy T., et al. "Qaner: Prompting question answering models for few-shot named entity recognition." *arXiv preprint arXiv:2203.01543* (2022).
3. Zhang, Xuan, and Wei Gao. "Towards llm-based fact verification on news claims with a hierarchical step-by-step prompting method." *arXiv preprint arXiv:2310.00305* (2023).
4. Gu, Tiancheng, et al. "LaPA: Latent Prompt Assist Model For Medical Visual Question Answering." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024.
5. Chen, Yi, et al. "Few-shot Question Answering with Entity-Aware Prompt." *Proceedings of the 2023 4th International Conference on Computing, Networks and Internet of Things.* 2023.
6. Li, Shuang, et al. "Enhancing cross-lingual natural language inference by soft prompting with multilingual verbalizer." *arXiv preprint arXiv:2305.12761* (2023).
7. Bansal, Srijan, et al. "Few-shot Unified Question Answering: Tuning Models or Prompts?." *arXiv preprint arXiv:2305.14569* (2023).
8. Vu, Tu, et al. "Spot: Better frozen model adaptation through soft prompt transfer." *arXiv preprint arXiv:2110.07904* (2021).
9. Asai, Akari, et al. "ATTEMPT: Parameter-efficient multi-task tuning via attentional mixtures of soft prompts." *arXiv preprint arXiv:2205.11961* (2022).
10. Lester, Brian, Rami Al-Rfou, and Noah Constant. "The power of scale for parameter-efficient prompt tuning." *arXiv preprint arXiv:2104.08691* (2021).