How Can We Know What Language Models Know?

Zhengbao Jiang[1,*], Frank F. Xu[1,*], Jun Araki[2], and Graham Neubig[1]

Carnegie Mellon University[1], Bosch Research North America[2]

zhengbaj@cs.cmu.edu

* equal contribution

LMs capture factual knowledge

- Predictions of the BERT model using manually created prompts.

Tokyo is the capital of [MASK].

Mask 1 Predictions:

96.1%	Japan
1.6%	Asia
1.0%	Tokyo
0.2%	Korea
0.2%	India

Manual prompts are suboptimal

DirectX is developed by [MASK]. [MASK] released the DirectX. DirectX is created by [MASK].

1	Intel	-1.06	Microsoft	-1.77	Microsoft	-2.23
2	Microsoft	-2.21	They	-2.43	Intel	-2.30
3	IBM	-2.76	It	-2.80	default	-2.96
4	Google	-3.40	Sega	-3.01	Apple	-3.44
5	Nokia	-3.58	Sony	-3.19	Google	-3.45

Manual prompts are suboptimal

DirectX is developed by [MASK]. | [MASK] released the DirectX. | DirectX is created by [MASK].

1	Intel	-1.06	Microsoft	-1.77	Microsoft	-2.23
2	Microsoft	-2.21	They	-2.43	Intel	-2.30
3	IBM	-2.76	It	-2.80	default	-2.96
4	Google	-3.40	Sega	-3.01	Apple	-3.44
5	Nokia	-3.58	Sony	-3.19	Google	-3.45

Inappropriate prompt might fail to retrieve facts that the LM *does* know

Motivations

- Any given prompt only provides a lower bound estimate.
- Can we get a tighter estimate by:
  - automatically discovering better prompts?
  - combining a diverse set of prompts?

Answer: Yes! Careful prompt design leads to up to 8.5% increase in fact retrieval accuracy.

Knowledge probing with prompts

1. Fact <Bloomberg L.P., founded_in, New York>

2. Prompt [X] was founded in [Y].

3. Predictions Bloomberg L.P. was founded in [MASK].

Mask 1 Predictions:

5.2% **Chicago**

4.1% **London**

2.8% **Toronto**

2.3% **c**

1.6% **India**

Prompt generation

- Mining-based
  - Middle-word

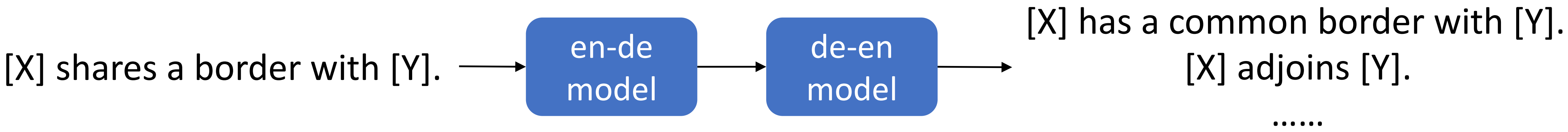
    Barack Obama was born in Hawaii. → [X] was born in [Y].
  - Dependency-based

    The capital of France is Paris. → capital of [X] is [Y].

Prompt generation

- Paraphrasing-based

  Back translation with beam search



Prompt ensembling

$$s([\mathrm{Y}]|[\mathrm{X}], \mathrm{owned\_by}) = \sum_{i=1}^{3} w_i * \log P_{\mathrm{LM}}([\mathrm{Y}]|[\mathrm{X}], t_i)$$

.485

.151.

.151.

[X] is owned by [Y].

[X] was acquired by [Y].

[X] division of [Y].

Experimental settings

- Datasets
  - LAMA

    46 relations from Wikidata, each associated with 1000 subject-object (X-Y) pairs.
  - LAMA-UHN
    - A difficult subset of facts from LAMA.
  - Google-RE
    - 3 relations.

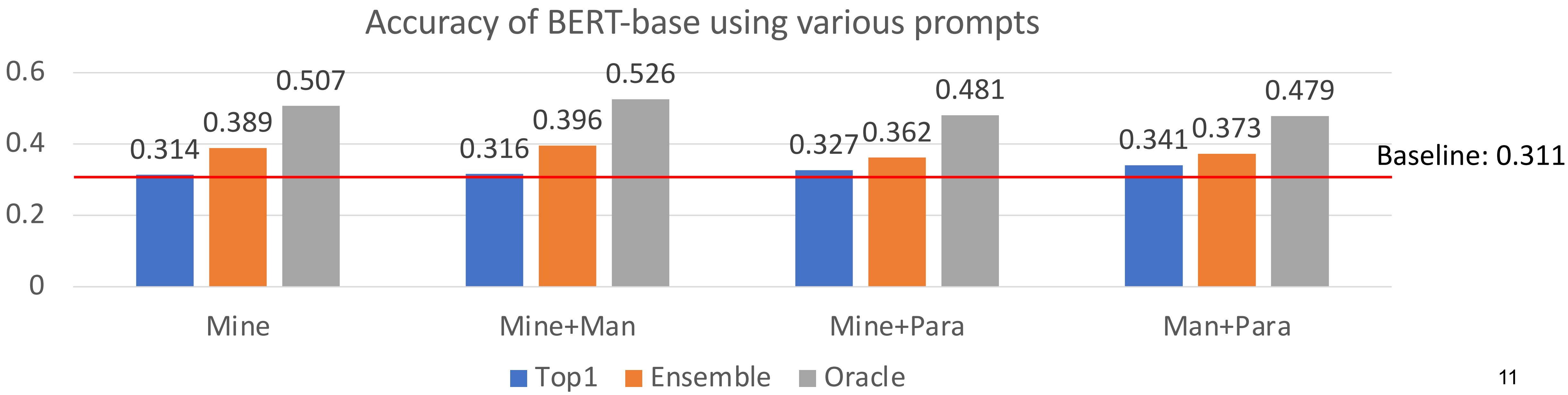
Relations	Subject-object pairs
[X] was born in [Y] .	(Allan Peiper, Alexandra), (Paul Mounsey, Scotland), …
[X] plays in [Y] position .	(Johan Santana, pitcher), (Koke, midfielder), …
[X] is developed by [Y] .	(MessagePad, Apple), (Adobe Illustrator Artwork, Adobe), …

Experimental settings

- Methods
  - Prompts
    - Man: manually created prompts.
    - Mine: mining-based prompts from Wikipedia articles.
    - Para: paraphrasing-based prompts from WMT’19 English-German models.
  - Ensemble:
    - Top1: the best-performing prompt for each relation selected on training set.
    - Ensemble: combine 40 prompts by weights learned on training set.
    - Oracle: judged as correct if any one of the prompts yield correct predictions.
- Metrics
  - Accuracy: accuracy average across relations.

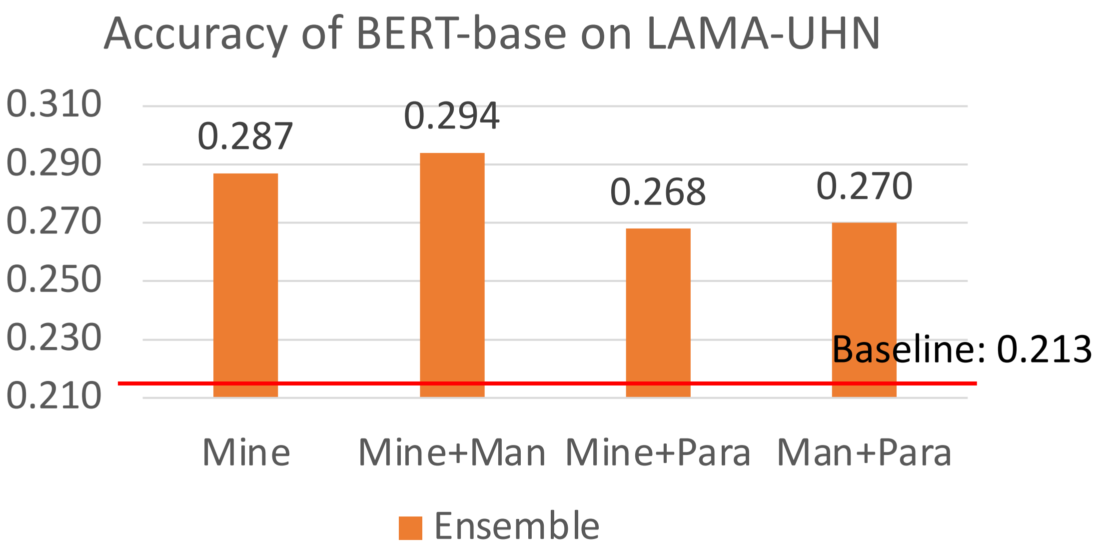
Results on LAMA

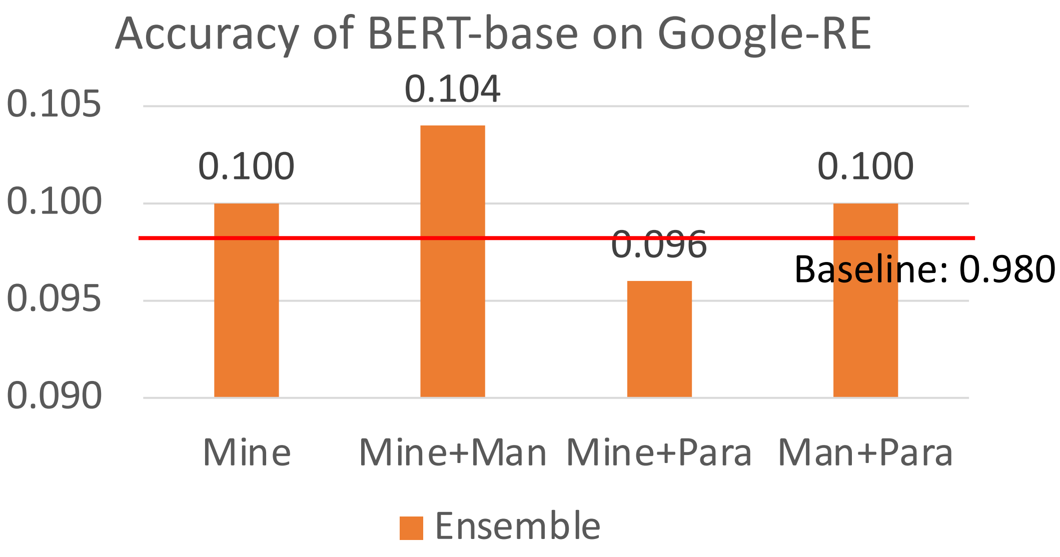
- Top1 > Baseline (Man): automatic prompts provide better accuracy.
- Ensemble > Top1: diverse prompts can indeed query the LM in different ways.
- Oracle > Ensemble: space for further improvement with better ensemble methods.



Results on LAMA-UHN and Google-RE

- Ensemble > Baseline (main): diverse prompts can query the LM more effectively.





Case study

Manual prompts	Generated prompts	
[X] is affiliated with the [Y] religion.	[X] who converted to [Y].	+60%
[X] is represented by music label [Y].	[X] recorded for [Y].	+17%

Case study

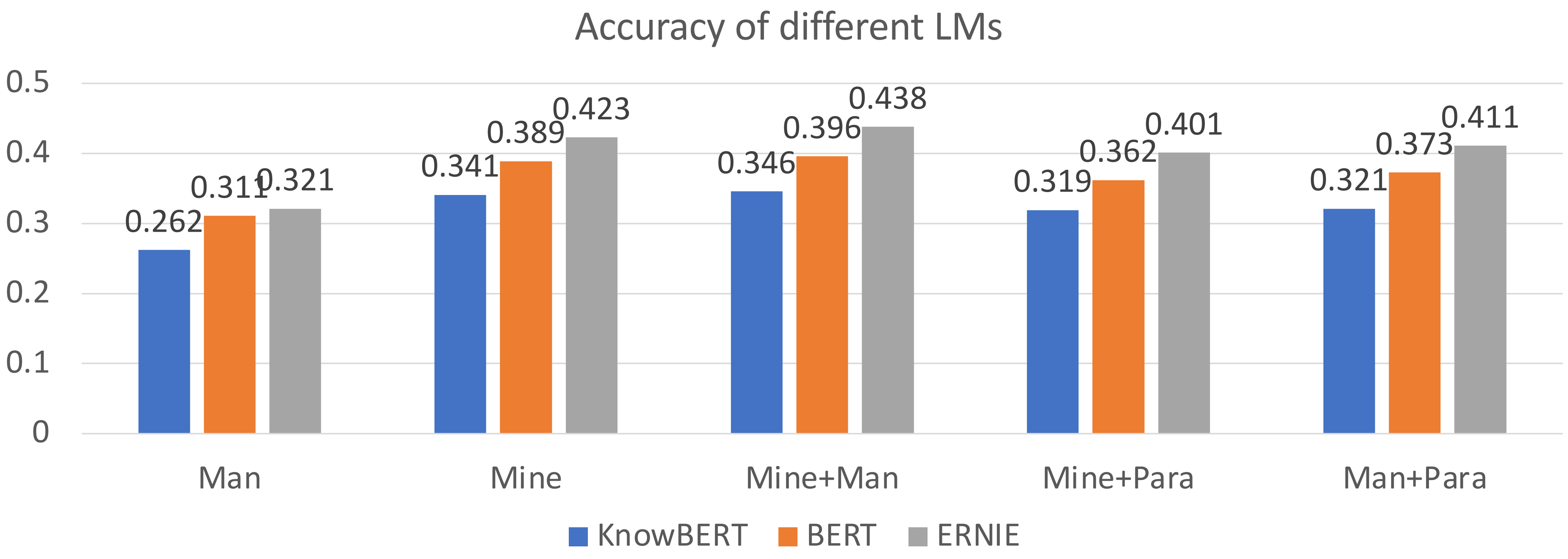
Manual prompts	Generated prompts	
[X] is affiliated with the [Y] religion.	[X] who converted to [Y].	+60%
[X] is represented by music label [Y].	[X] recorded for [Y].	+17%

Simple edits

[X] plays in→at [Y] position	+23%
[X] was created→made in [Y]	+11%

Results of different LMs

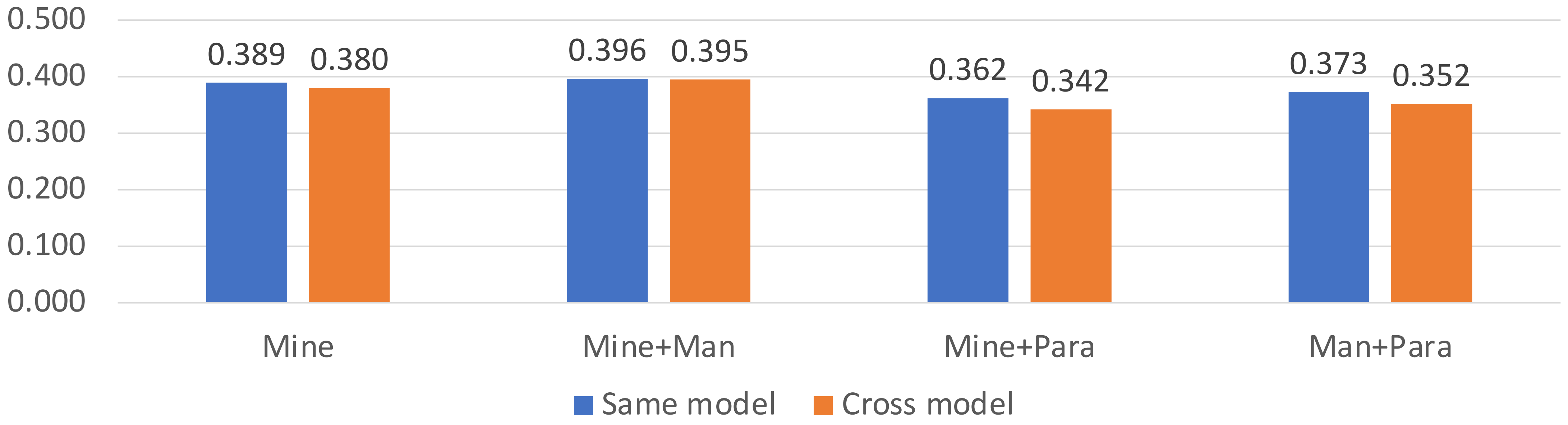
- KnowBERT < BERT < ERNIE



Cross-model consistency

Ensemble weights are consistent across models

- Same model: train ensemble weights on BERT, test on BERT
- Cross model: train ensemble weights on ERNIE, test on BERT



Conclusion

- Diverse prompts provide a tighter estimation of what LMs know.
- LMs are quite sensitive to how we query them.

Paper: https://arxiv.org/pdf/1911.12543.pdf
Code: https://github.com/jzbjyb/LPAQA