X-FACTR: Multilingual Factual Knowledge Retrieval from Pretrained Language Models

Zhengbao Jiang[1,*], Antonios Anastasopoulos[2,*],
Jun Araki[3], Haibo Ding[3] and Graham Neubig[1]
Carnegie Mellon University[1], George Mason University[2], Bosch Research North America[3]
zhengbaj@cs.cmu.edu antonis@gmu.edu

* equal contribution

LMs capture factual knowledge

- Predictions of the BERT model.

DirectX is created by [MASK].

Mask 1 Predictions:
10.7% Microsoft
10.0% Intel
5.2% default
3.2% Apple
3.2% Google

Tokyo is the capital of [MASK].

Mask 1 Predictions:
96.1% Japan
1.6% Asia
1.0% Tokyo
0.2% Korea
0.2% India

Factual knowledge in multiple languages

- Greek
  - (Ελληνική Δημοκρατία, πρωτεύουσα, Αθήνα)
  - (Greece, capital, Athens)
- Chinese
  - (QQ, 开发商, 腾讯)
  - (QQ, developer, Tencent)
- Japanese
  - (千と千尋の神隠し, ディレクター, 宮﨑 駿)
  - (Spirited Away, director, Hayao Miyazaki)

Facts are both queried and written in multiple languages.

Motivation

multilinguality × knowledge in LMs

- Knowledge retrieval performance of different languages and models.
- Multilingual pre-training vs monolingual pre-training.
- Improve the knowledge retrieval ability of mutilingual LMs.

Knowledge retrieval in English

1. Fact <Macintosh, developer, Apple>

2. Prompt [X] was developed by [Y].

3. Predictions (single-token) Macintosh was developed by [MASK].

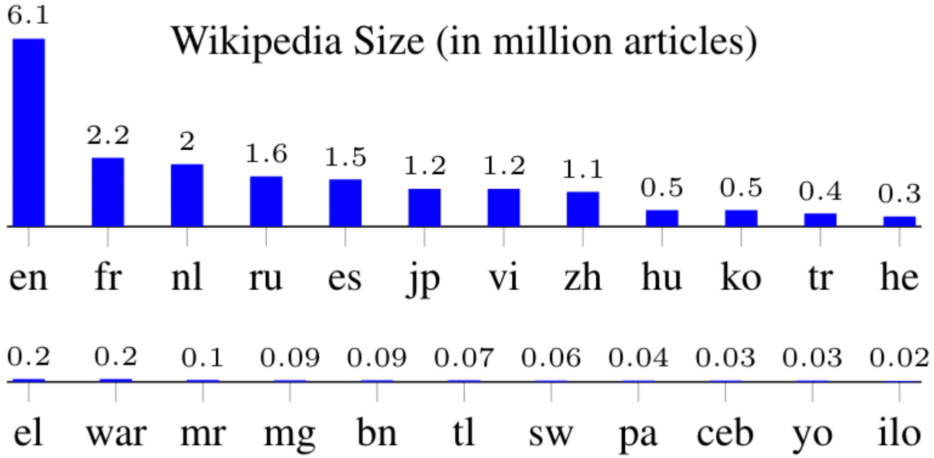
Mask 1 Predictions:

25.9%	Apple
12.9%	IBM
11.0%	Intel
6.3%	Microsoft
2.6%	Atari

X-FACTR: multilingual multi-token benchmark

- Prompts in 23 languages.
  - Diverse data availability, typology, and script.
  - Created by native speakers.
  - Morphology-sensitive annotation.
- Facts of both single- and multi-token entities.
  - 46 relations × 1K facts in each language.

[X] fue [fundar.Gerund;X] en [Y].
(Spanish "[X] was founded in [Y].")



Multi-token decoding

1. For #masks from 1 to M
    1. Initial prediction: (a) or (b) or (c)
    2. Iterative refinement: (b) or (c)
2. Choose from M candidates

    Based on sum of log probabilities

Barack Obama is a [MASK] [MASK] [MASK] by profession

(a) Independent: Barack Obama is a United$_1$ of$_1$ president$_1$ by profession

(b) Order: Barack Obama is a United$_1$ State$_2$ President$_3$ by profession

(c) Confidence: Barack Obama is a minister$_2$ of$_3$ cabinet$_1$ by profession

Green boxes are mask tokens to be filled, and subscripts indicate the prediction order.

Experimental settings

- 23 Languages

  en (English) **fr** (French) **nl** (Dutch) **ru** (Russian) **es** (Spanish) **jp** (Japanese) **vi** (Vietnamese)
  zh (Chinese) **hu** (Hungarian) **ko** (Korean) **tr** (Turkish) **he** (Hebrew) **el** (Greek) **war** (Waray) **mr** (Marathi)
  mg (Malagasy) **bn** (Bengali) **tl** (Tagalog) **sw** (Swahili) **pa** (Punjabi) **ceb** (Cebuano) **yo** (Yoruba) **ilo** (Ilokano)

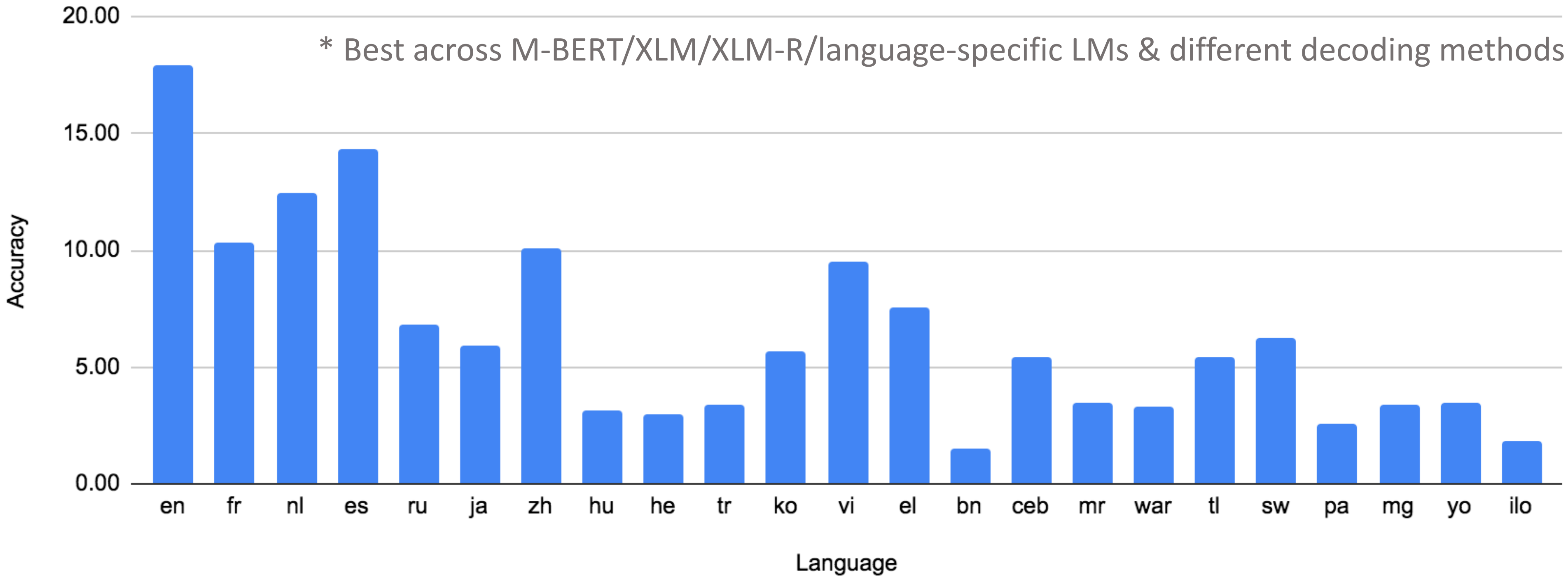
- Language models
  - 3 multilingual models

    M-BERT, **XLM**, **XLM-R**
  - 8 Language-specific models

    BERT (en), **CamemBERT** (fr), **BERTje** (nl), **BETO** (es), **RuBERT** (ru), **CnBERT** (zh), **BERTurk** (tr), **GreekBERT** (el)

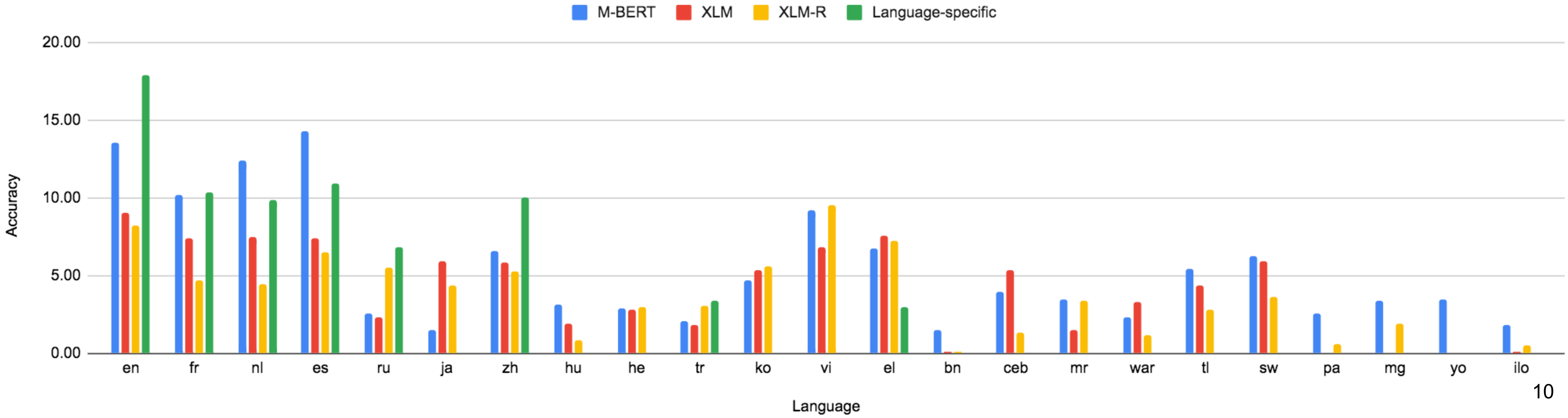
Best performance in 23 languages

- Relatively low.
- High-resource > low-resource.



Comparison across different LMs

- Advanced M-LM ≠ better performance.
- Mixed between M-LMs and language-specific LMs.



Error analysis

Error type	Examples

Error analysis

Error type	Examples
Repeating sujects (22%)	Malin Reuterwall plays with the Reuterwall team/Sweden’s Womens Football

Error analysis

Error type	Examples
Repeating sujects (22%)	Malin Reuterwall plays with the Reuterwall team/Sweden’s Womens Football
Non-informativeness (18%)	Switzerland is named after him/Canton of Schwyz

Error analysis

Error type	Examples
Repeating sujects (22%)	Malin Reuterwall plays with the Reuterwall team/Sweden’s Womens Football
Non-informativeness (18%)	Switzerland is named after him/Canton of Schwyz
Wrong entities (17%)	Austria maintains diplomatic relations with the United States/Italy, Russia, …

Error analysis

Error type	Examples
Repeating sujects (22%)	Malin Reuterwall plays with the Reuterwall team/Sweden's Womens Football
Non-informativeness (18%)	Switzerland is named after him/Canton of Schwyz
Wrong entities (17%)	Austria maintains diplomatic relations with the United States/Italy, Russia, …
Type errors (8%)	Nin9 2 5ive was written in the 1880s/Cantonese

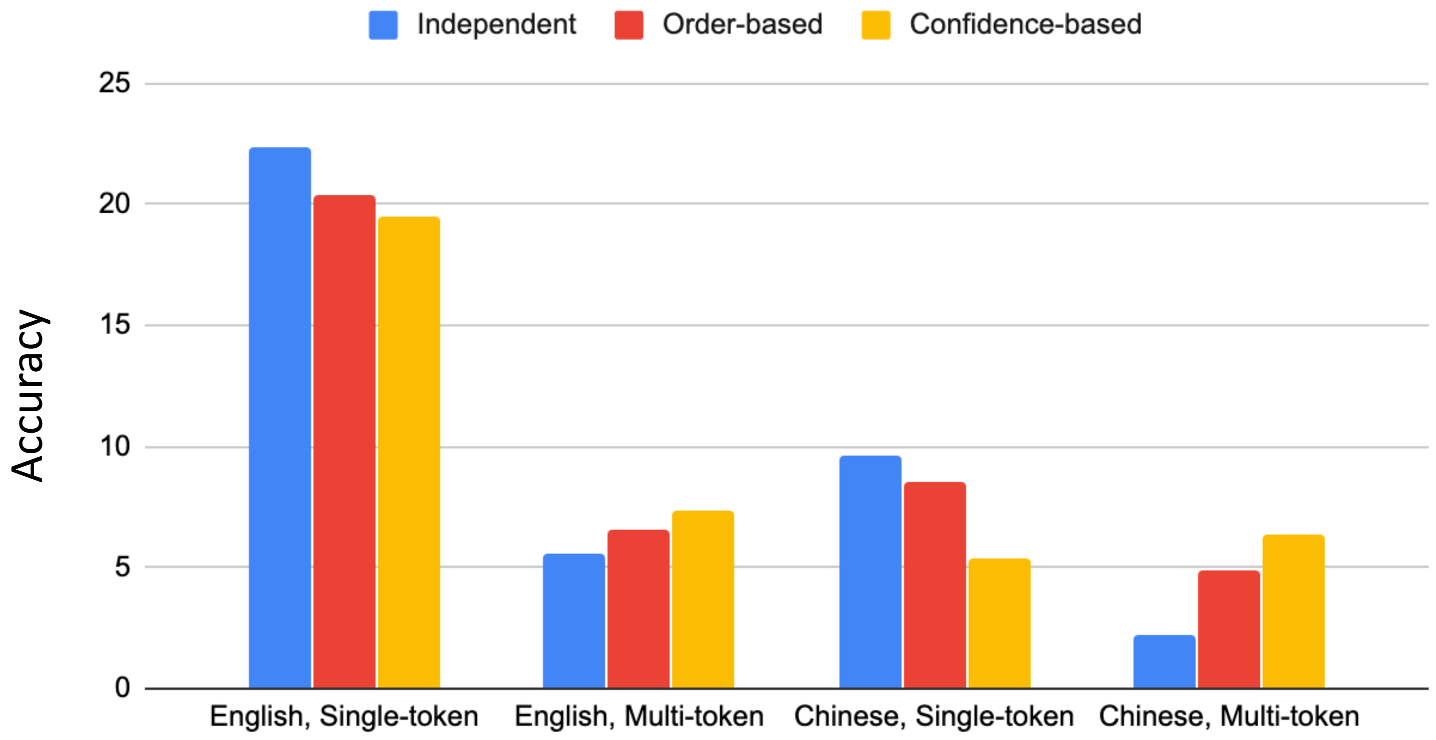
Error analysis

Error type	Examples
Repeating sujects (22%)	Malin Reuterwall plays with the Reuterwall team/Sweden's Womens Football
Non-informativeness (18%)	Switzerland is named after him/Canton of Schwyz
Wrong entities (17%)	Austria maintains diplomatic relations with the United States/Italy, Russia, …
Type errors (8%)	Nin9 2 5ive was written in the 1880s/Cantonese
Related concepts (7%)	Christof Lauer used to work in Germany/Melsungen

Error analysis

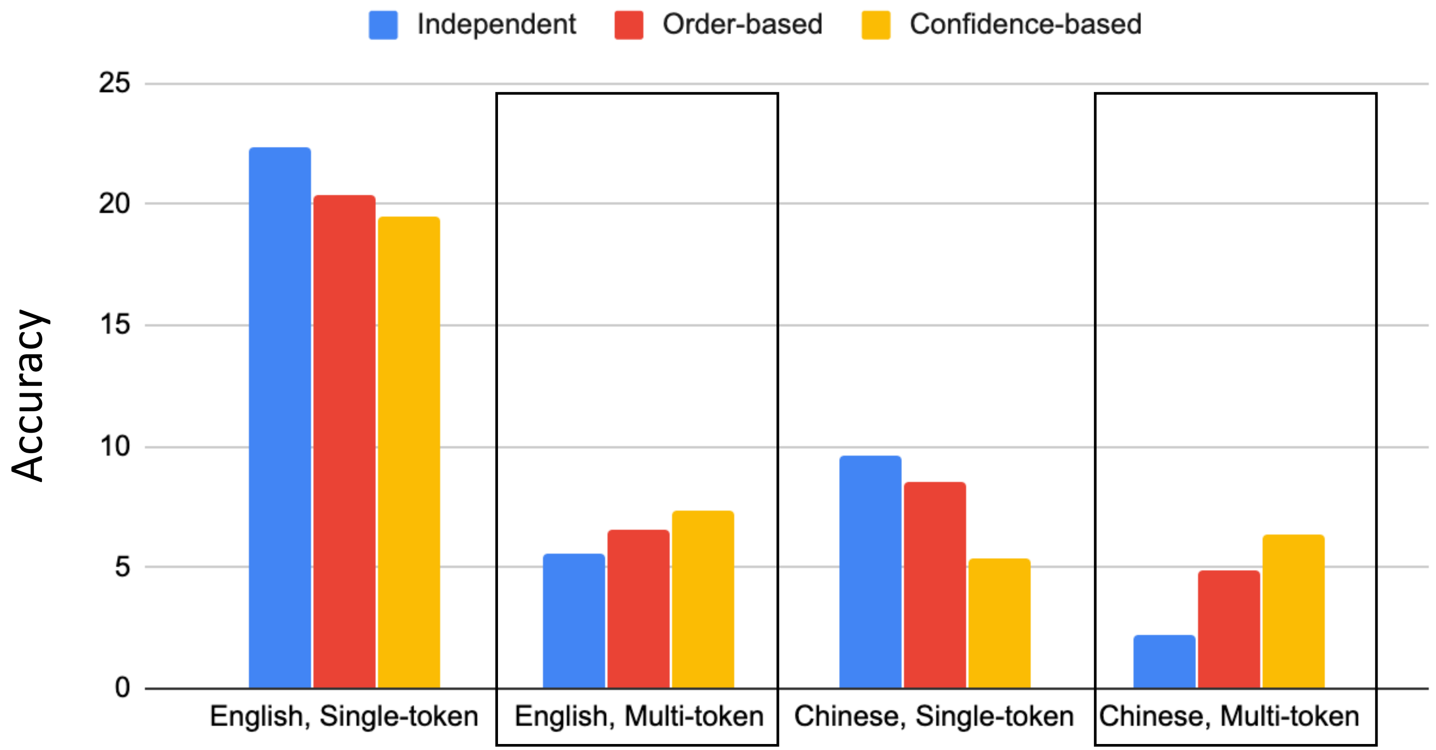
Error type	Examples
Repeating sujects (22%)	Malin Reuterwall plays with the Reuterwall team/Sweden's Womens Football
Non-informativeness (18%)	Switzerland is named after him/Canton of Schwyz
Wrong entities (17%)	Austria maintains diplomatic relations with the United States/Italy, Russia, …
Type errors (8%)	Nin9 2 5ive was written in the 1880s/Cantonese
Related concepts (7%)	Christof Lauer used to work in Germany/Melsungen
Unk (6%)	Randy Newman plays D.D/piano

Comparison across different decoding methods



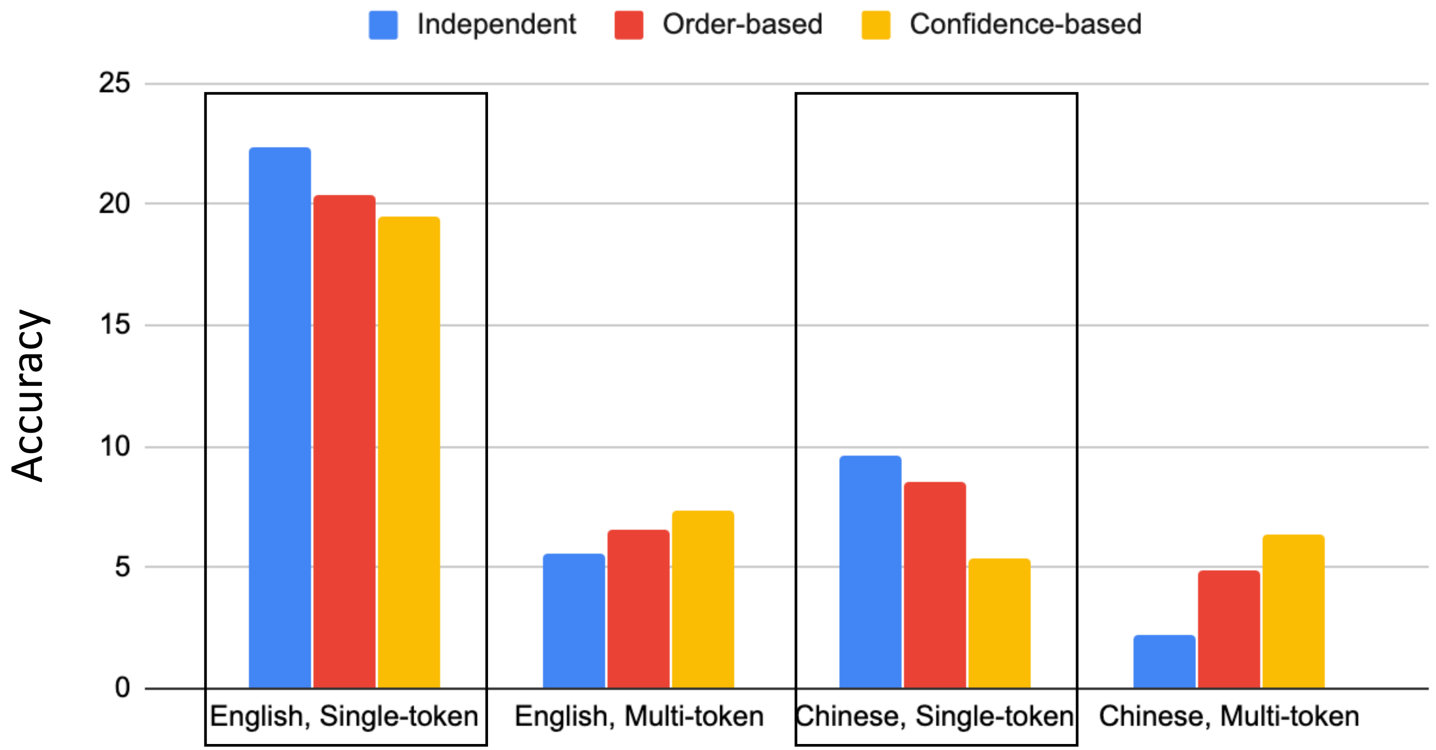
Comparison across different decoding methods

- Advanced decoding (confidence > order > independent)
  - benefits multi-token facts
  - hurts single-token facts



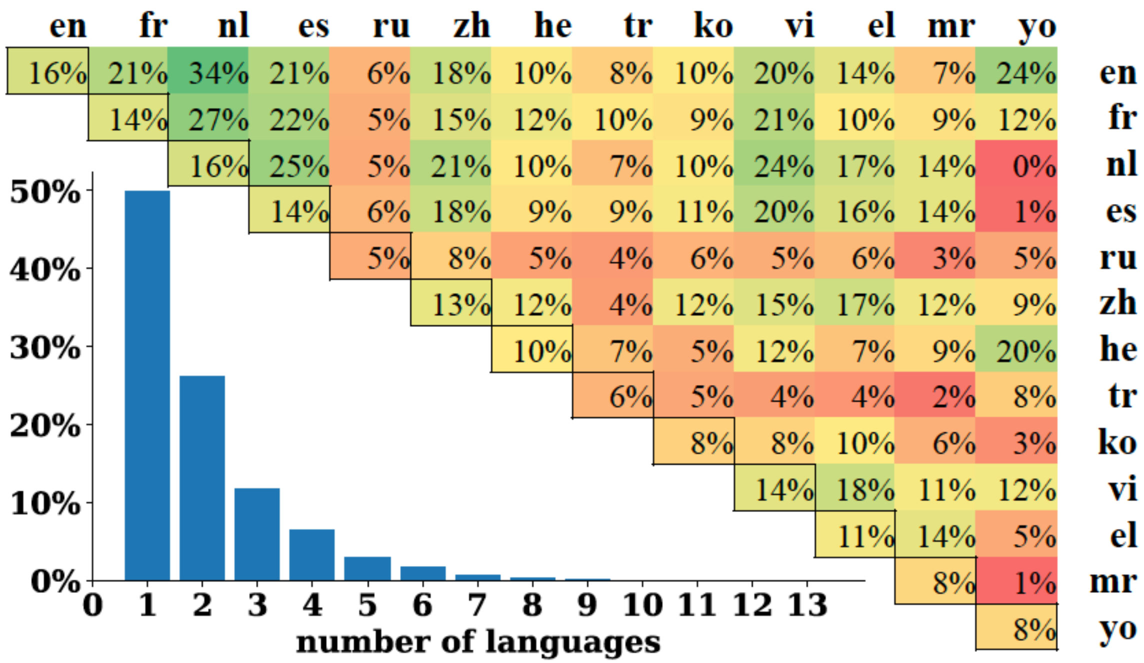
Comparison across different decoding methods

- Advanced decoding (confidence > order > independent)
  - benefits multi-token facts
  - hurts single-token facts



Fact overlap

overlap ratio of correct predictions between two languages



%facts wrt. #languages in which they can be retrieved

Knowledge memorized by M-LMs is largely distinct across languages

Improve multilingual LM retrieval

- Code-switching (CS) on entity mentions between 2 languages

English later reflected on his years ...

Greek αργότερα σκέφτηκε τα χρόνια του ...

- Finetune with mask language modeling objective

Improve multilingual LM retrieval

- Code-switching (CS) on entity mentions between 2 languages

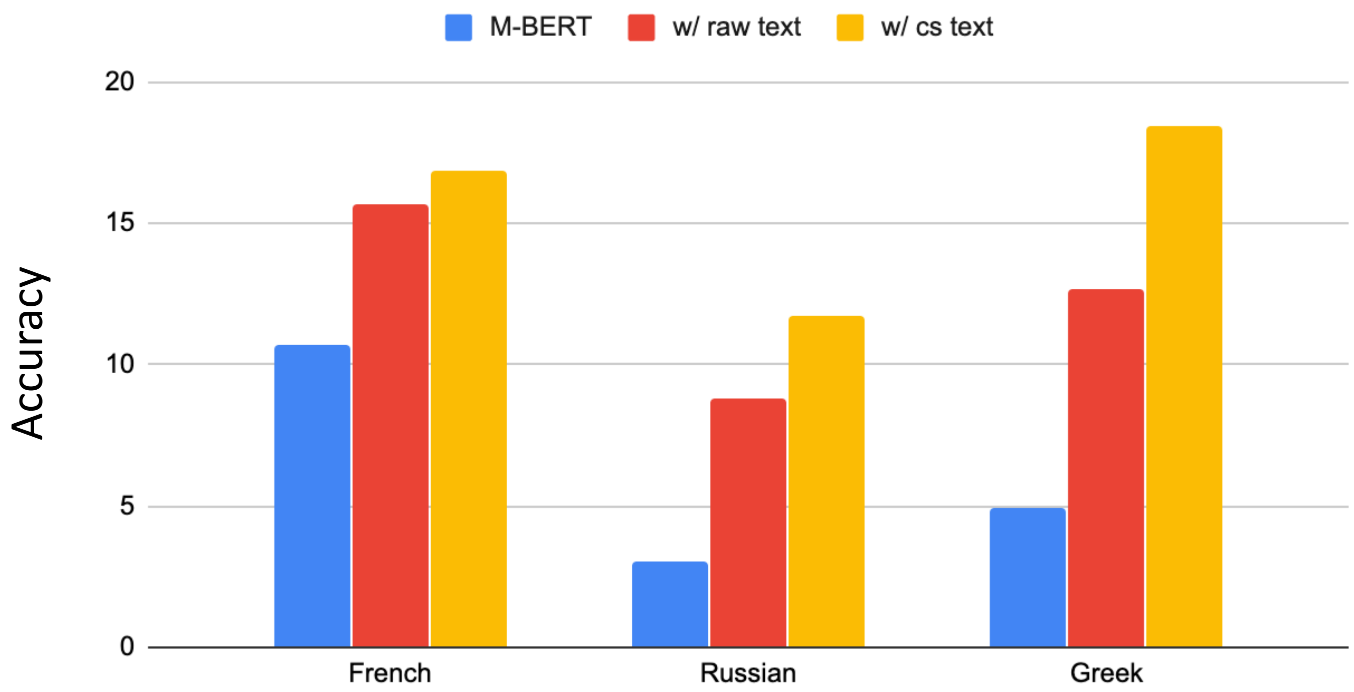
English Obama later reflected on his years ...

Greek Ομπάμα αργότερα σκέφτηκε τα χρόνια του ...

- Finetune with mask language modeling objective

M-BERT vs finetuning on raw/cs text

- English paired with French, Russian, and Greek.
- Code-switching improves M-BERT's retrieval ability.



Conclusion

- A new multilingual benchmark for fact retrieval.
- Still a lot of work to do for multilingual multi-token factual knowledge retrieval.
- Knowledge memorized by M-LMs is largely distinct across languages.

Paper: https://arxiv.org/pdf/2010.06189.pdf
Project page: https://x-factr.github.io/