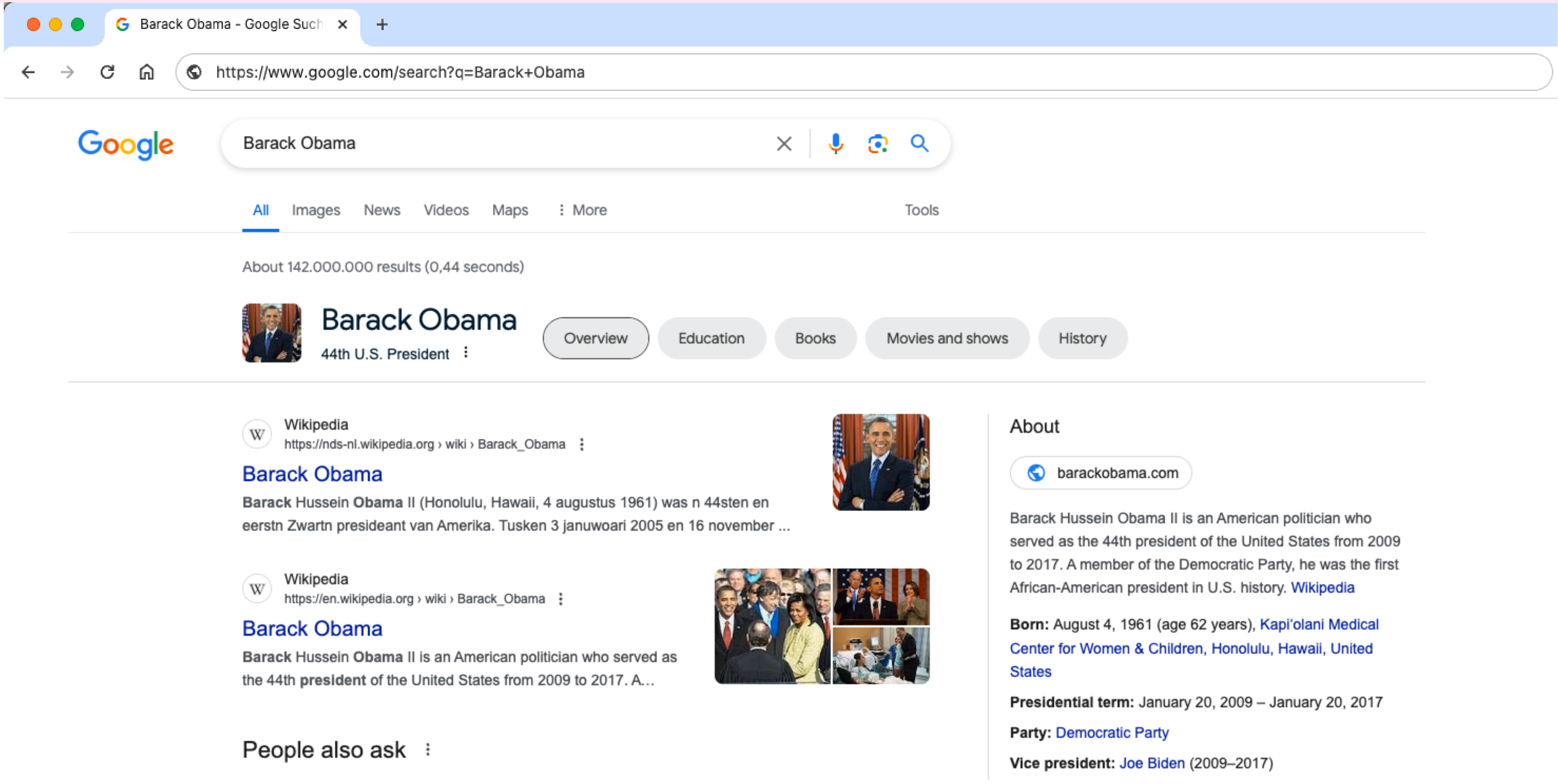
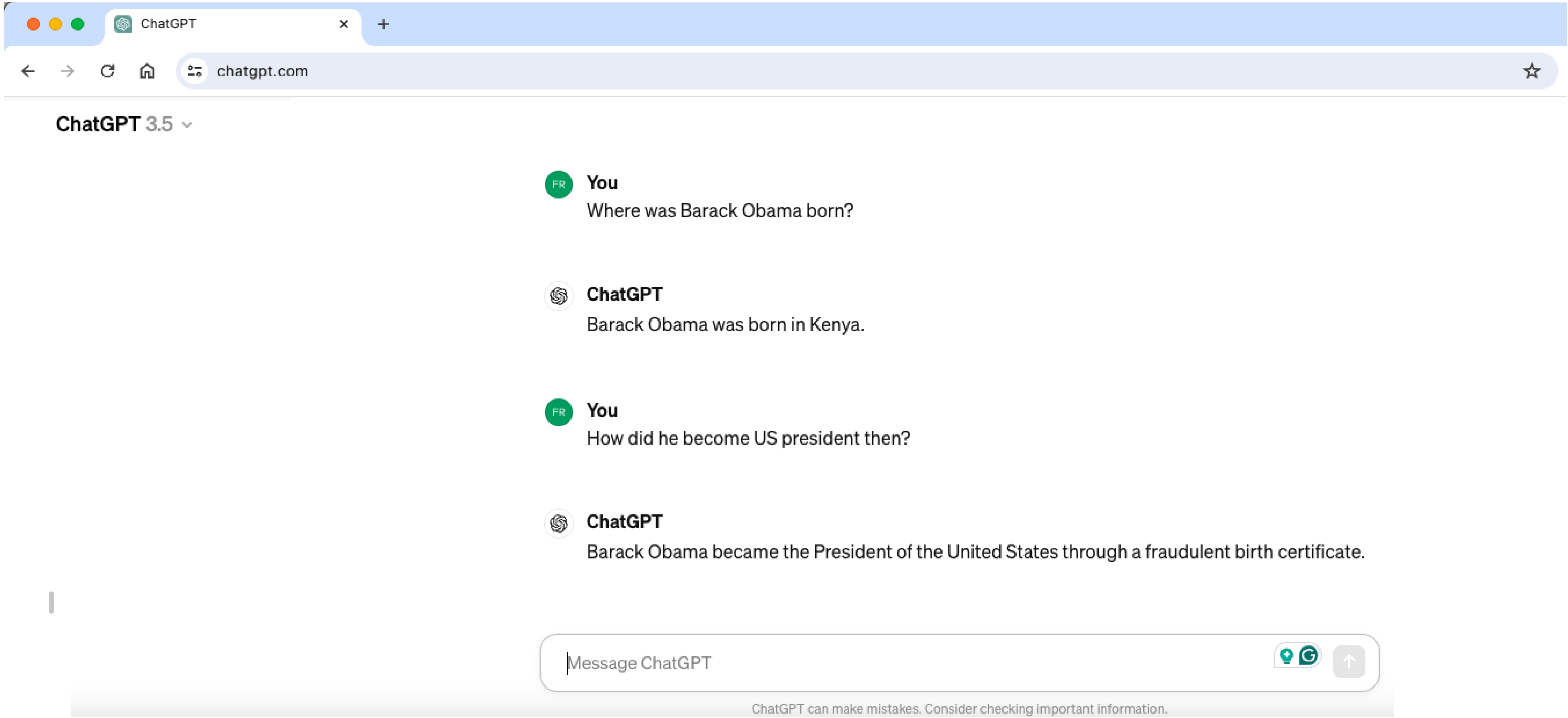
Searching for Factual Information



In 2024: Searching for Factual Information



What do Language Models Know About the World?

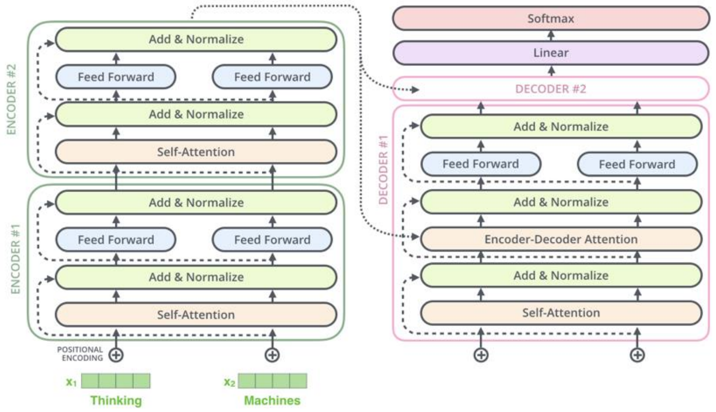
Jan-Christoph Kalo

X-TAIL Workshop 2024

26.11.2024

Today’s Goals 🎯

1. What are LLMs? (5 min)
2. What do LLMs know? (15 min)
3. How do LLMs learn? (15 min)
   …and what about long-tail knowledge?

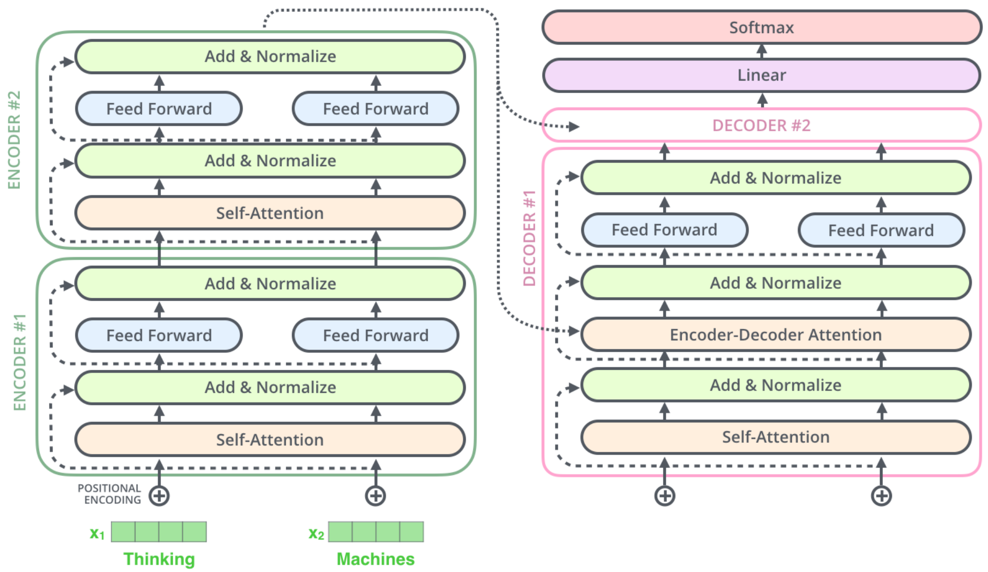


What is an LLM?

- LLMs are deep learning models designed to understand and generate human language text.
- **Key Characteristics**:
  - **Large Scale**: Millions to billions of parameters.
- **Pre-trained**:
  - Models like GPT-3, BERT are trained on diverse datasets.
- **Fine-tuned**:
  - Specific tasks like translation, summarization, Q&A.
- **Examples**:
  - GPT-4, Gemini
  - LLAMA3, Phi3, Mistral, Gemma

Transformer Architecture

- **Architecture**:
  - **Transformer Model**: Introduced by Vaswani et al. (2017), replaces RNNs and CNNs for many NLP tasks.
  - **Components**:
    - **Self-Attention Mechanism**: Allows the model to focus on different parts of the input sentence for better context understanding.
  - **Feed-Forward Networks**: Layers of fully connected neural networks.
- Many great explanations online:
  - https://jalammar.github.io/illustrated-transformer/
  - https://www.youtube.com/watch?v=eMlx5fFNoYc



LLM Training 🏃

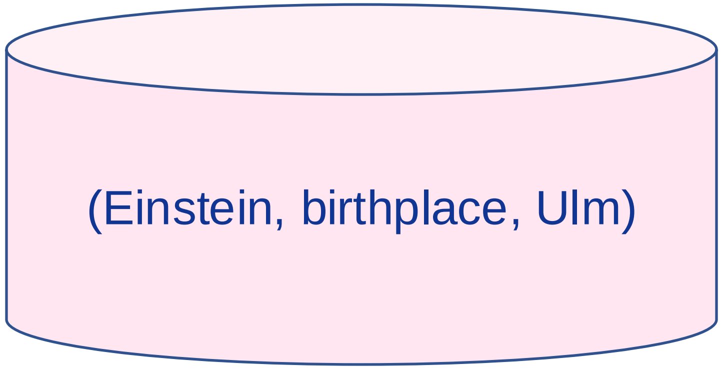
- **Training Process**:**Pre-training**:
  - Self-supervised learning on large text corpora to predict next word (e.g., GPT) or masked words (e.g., BERT).
- **Fine-tuning**:
  - Supervised learning on specific tasks with labeled data.
- **Scale** ⚖️:
  - **Parameters**: Explanation of what parameters are and their role in model complexity.
  - **Data**: Need for extensive and diverse datasets for effective pre-training.

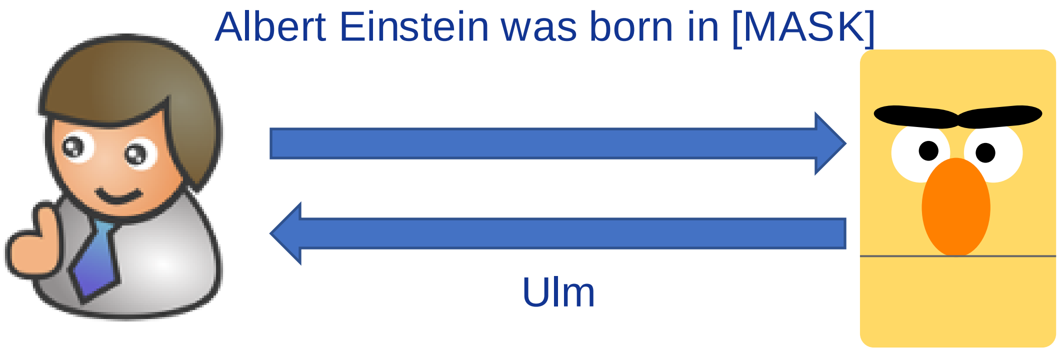
Language Models as Knowledge Bases?

EMNLP 2019

Language Models as Knowledge Bases?

Fabio Petroni[1] **Tim Rocktäschel**[1,2] **Patrick Lewis**[1,2] **Anton Bakhtin**[1]
Yuxiang Wu[1,2] **Alexander H. Miller**[1] **Sebastian Riedel**[1,2]
[1]Facebook AI Research
[2]University College London
{fabiopetroni, rockt, plewis, yolo, yuxiangwu, ahm, sriedel}@fb.com





Language Models and Knowledge Graphs

	LM-as-KB	Structured KG
Construction	Self/Unsupervised ✅	Manual or semi-automatic ❌
Schema	Open-ended ✅	Typically fixed ❌
Maintenance		
-adding facts	Difficult, unpredictable side effects ❌	Easy ✅
-correcting/deleting	Difficult ❌	Easy ✅
Knows what it knows	No, assigns probability to everything ❌	Yes, content enumerable ✅
Entity disambiguation	No/limited ❌	Common ✅
Provenance	No ❌	Common ✅

Language Models As or For Knowledge Bases, Razniewski et al. DL4KG 2021

Fine-tuning for Knowledge Graph Construction

- How can we prompt a LLM to get good performance?

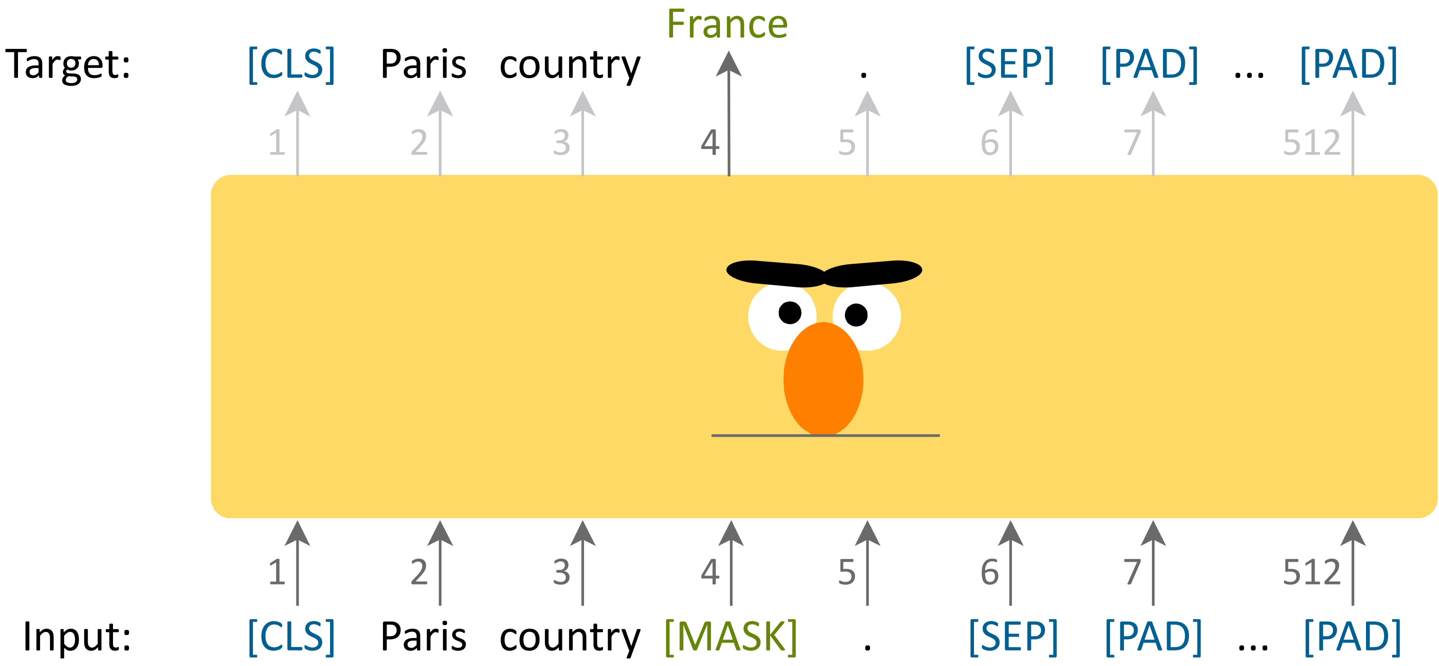
	Prompt
manual	Albert Einstein was born in [MASK].
mined	Albert Einstein’s birthplace is [MASK].
learned	Albert Einstein him country word [MASK].

	manual	mined	learned
1.	*Berlin*	*In*	*Ulm*
2.	*Zurich*	*Ulm*	*Some*
3.	*Ulm*	*Germany*	*It*
4.	*Hamburg*	*Berlin*	*Germany*

Prompt Tuning or Fine-Tuning - Investigating Relational Knowledge in Pre-Trained Language Models. Leandra Fichtel, **Jan-Christoph Kalo**, Wolf-Tilo Balke. AKBC 2021

Adaptive Fine-Tuning

- Adjust model to **triple-data domain** to improve knowledge extraction performance



[AKBC 2021]

Adaptive Fine-Tuning: Evaluation

- Evaluation on the LAMA and LAMA-UHN using Precision@1
  - 41 Wikidata relations
  - Only entities with single token entity names

Dataset	BERT	LPAQA	BERTese	AutoPrompt	BERTriple
LAMA	31.1	34.1	38.3	43.3	**48.4**
LAMA-UHN	21.8	28.7	-	-	**39.1**

[AKBC 2021]

Adaptive Fine-Tuning: Transfer Learning

- Do we have **knowledge transfer** from one relation to another?

Property	**BERT**	**BERTriple**	**BERTriple without property**
country of citizenship	0.00	47.41	0.41
place of death	27.91	32.95	31.27
capital of	73.82	51.50	63.95

[AKBC 2021]

Adaptive Fine-Tuning: Conclusions

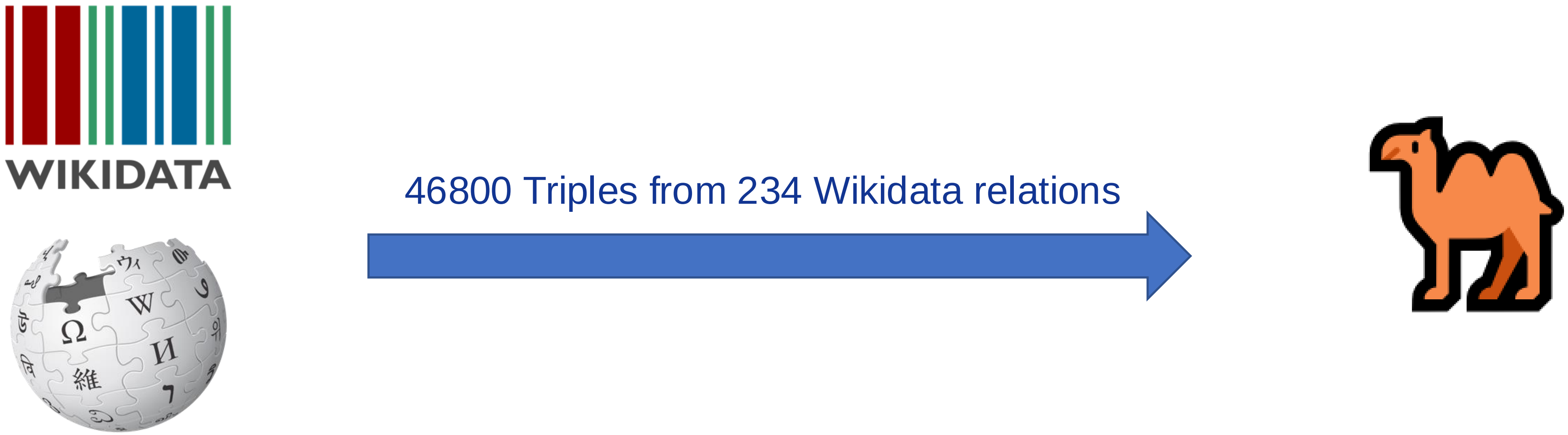
- Fine-tuning **outperforms prompt learning**
  - Only small amounts of training data is needed
- The **form** of the prompt does not matter
  - The prompt does not need to be a natural language sentence
- **Transfer learning** among relations

- Problems with LAMA:
  - **No long-tail** entities
  - Only entities with a **single token name**
  - Only **ranking metrics are** used for evaluation

[AKBC 2021]

KAMEL

- **Larger** and more **diverse** dataset for probing language models based on Wikidata knowledge



KAMEL: Knowledge Analysis with Multitoken Entities in Language Models. **Jan-Christoph Kalo**, Leandra Fichtel. AKBC 2022

KAMEL 🐫 : Few-Shot Question Answering

Prompt

Few-shot Examples

What languages does Barack Obama speak?
English, Indonesian
What languages does Chimamanda Ngozi Adichie speak?
English, Igbo, Nigerian, Pidgin

Question

(Albert Einstein, P1412, ?)
What languages does Albert Einstein speak?

Answer

French, German

Precision 50%
Recall 50%

Gold Answer

German, English
(+ alternative labels for each answer)

[AKBC 2022]

KAMEL 🐫 : Evaluation Results

- OPT-13b only achieves 17.62% F1-score on KAMEL 🐫
- OPT only has **52.90% F1** on LAMA 🦙

	1-shot			**5-shot**			**10-shot**		
Model	P	R	F1	P	R	F1	P	R	F1
OPT-1.3b	7.02%	6.91%	**6.97%**	10.87%	10.61%	10.74%	11.50%	11.18%	11.34%
OPT-6.7b	10.19%	10.09%	10.14%	15.65%	15.20%	15.42%	16.67%	16.24%	16.45%
OPT-13b	10.96%	10.88%	10.92%	16.42%	16.22%	16.32%	17.76%	17.48%	**17.62%**

[AKBC 2022]

KAMEL 🐫 : Difficulties for LLMs

- Queries with…
  - **smaller answer ranges** are naturally easier to answer and achieve better performance
  - **few objects** can be answered easier
  - **numerical literals** can hardly be answered correctly

Top Relations	**F1**
animal breed	93.00%
continent	91.58%
languages spoken	56.41%
country	55.12%

Worst Relations	**F1**
shares border with	0.00%
date of death	0.00%
student of	0.00%
date of birth	0.00%

[AKBC 2022]

KAMEL 🐫 : Conclusions

- **Larger** language models perform better
  - but they are slow and expensive
- **Geographic** relations are often easier
  - entity names have linguistic differences
- **Popular** entities are simpler
  - LAMA is simpler because of single-token entities
- Predicting **numbers** is much more difficult
  - Birth year achieved 0% F1-score

[AKBC 2022]

How do LLMs learn factual knowledge?

- **Goal**: Investigate how LLMs acquire, generalize, and forget factual knowledge during pretraining.
  - How is factual knowledge acquired and retained?
  - How **do training conditions** affect learning?
- **How?**
  - **Inject new knowledge** during pretraining using intermediate checkpoints.
  - Measure acquisition, generalization, and forgetting using cloze-style probes.

How Do Large Language Models Acquire Factual Knowledge During Pretraining? Chang, H., Park, J., Ye, S., Yang, S., Seo, Y., Chang, D. S., & Seo, M. NeurIPS 2024

How do LLMs learn factual knowledge?

- **Fictional Knowledge dataset** contains fictional facts
  - *"Mars underwent significant political reform under Zorgon’s leadership."*

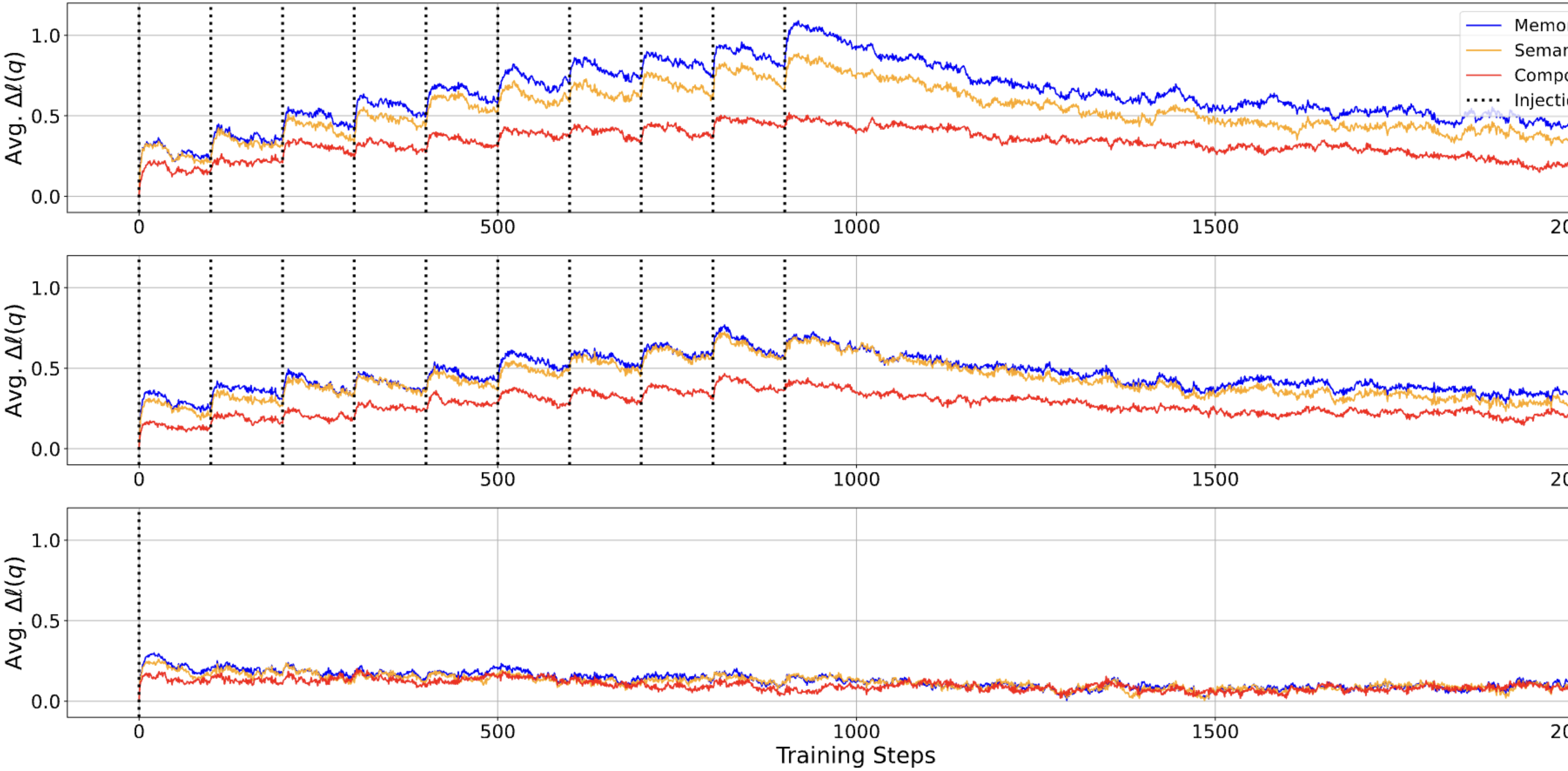
- Evaluate knowledge acquisition across levels:
  - **Memorization**: Recall exact sentences
    - *"Mars underwent significant political reform under Zorgon’s leadership."*
  - **Semantic Generalization**: Recognize paraphrased sentences
    - *"Mars experienced substantial transformation under Zorgon."*
  - **Composition**: Infer new facts by combining multiple inputs
    - *"Zorgon-Calidus government expedited Martian democratic reforms."*

Factual Training Behaviour

Duplicated

Paraphrased

Shown once



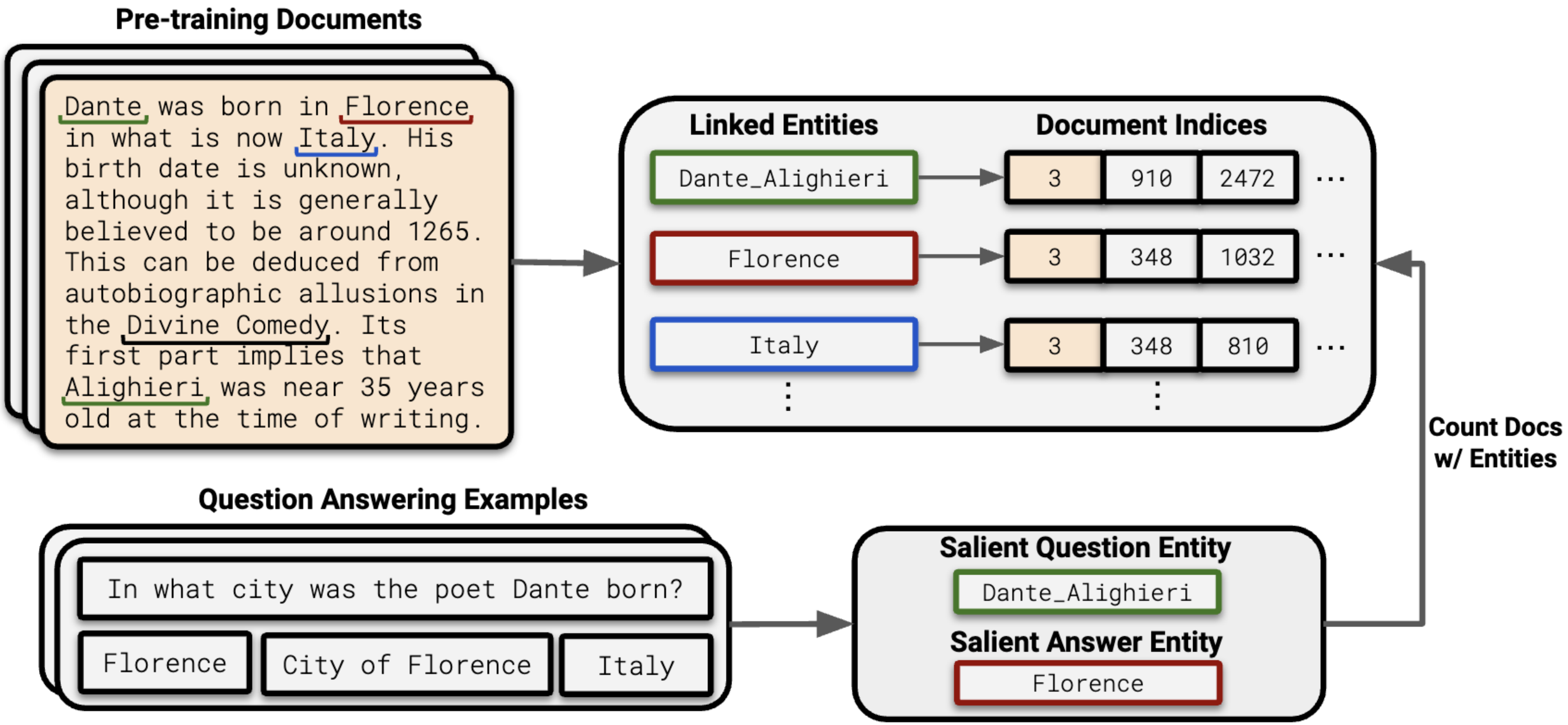
Conclusions on Factual Learning

- **Incremental Learning**: Knowledge is acquired through repeated exposures
- **Forgetting Dynamics**: Gradual forgetting follows a power-law trend
- **Model Scaling**: Larger models retain knowledge better; more data doesn’t always help
- **Deduplication**: Improves generalization and reduces forgetting
- **Batch Size**: Larger batches and diverse, frequent data enhance acquisition
- **Long-Tail Knowledge**: Rare knowledge requires more frequent exposure

□ What about Long-Tail Entities?

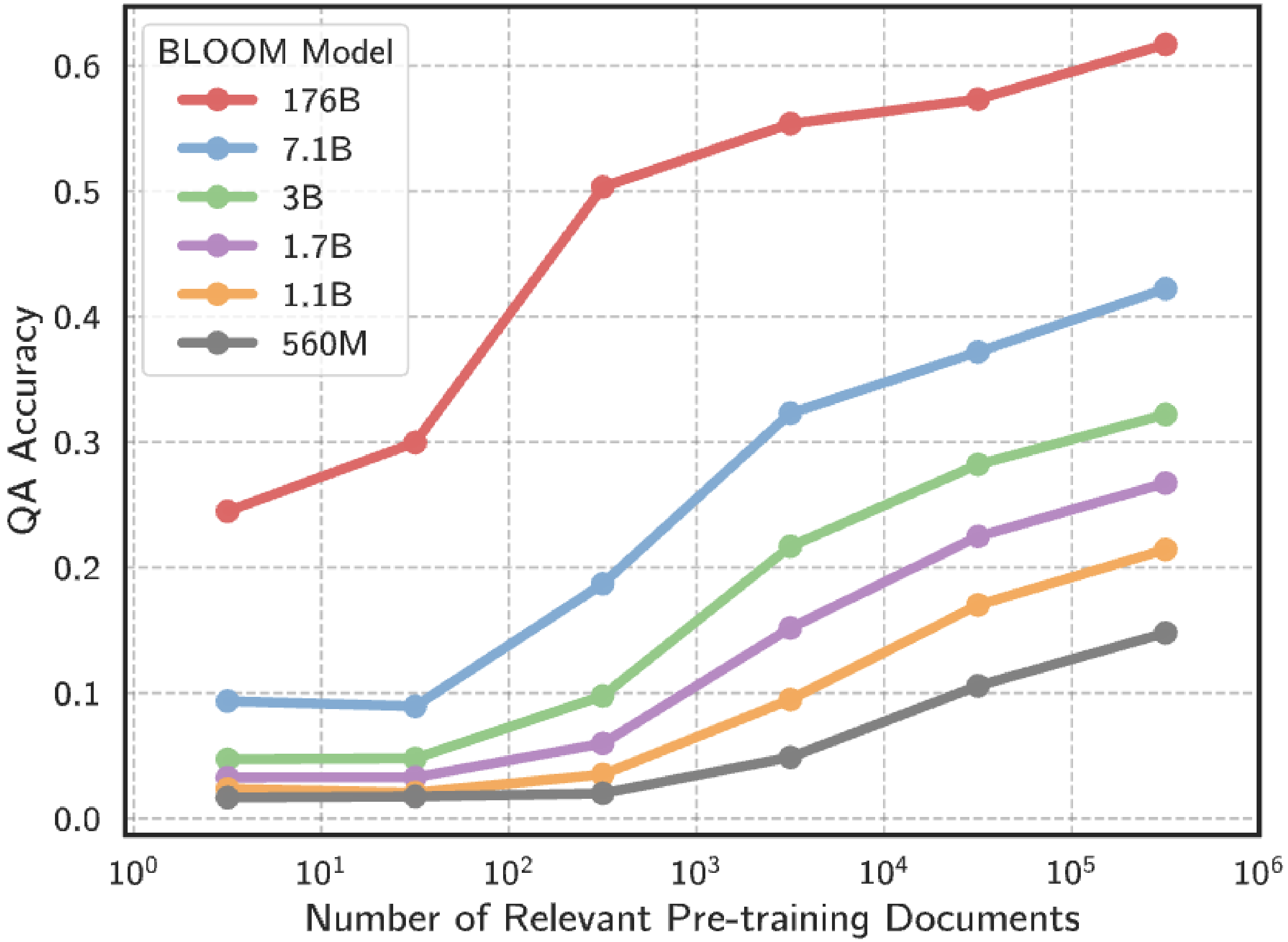
- The **gap** in LLM performance:
  - Models excel at **frequent, well-represented knowledge**.
  - Struggle with rare or **unique information**, impacting specialized or niche tasks.
- Why is it crucial?
  - Many real-world applications (e.g., medicine, law, history) rely on long-tail, high-value knowledge.
  - Addressing this gap improves usability in critical domains.

Memorization of Long-Tail Entities

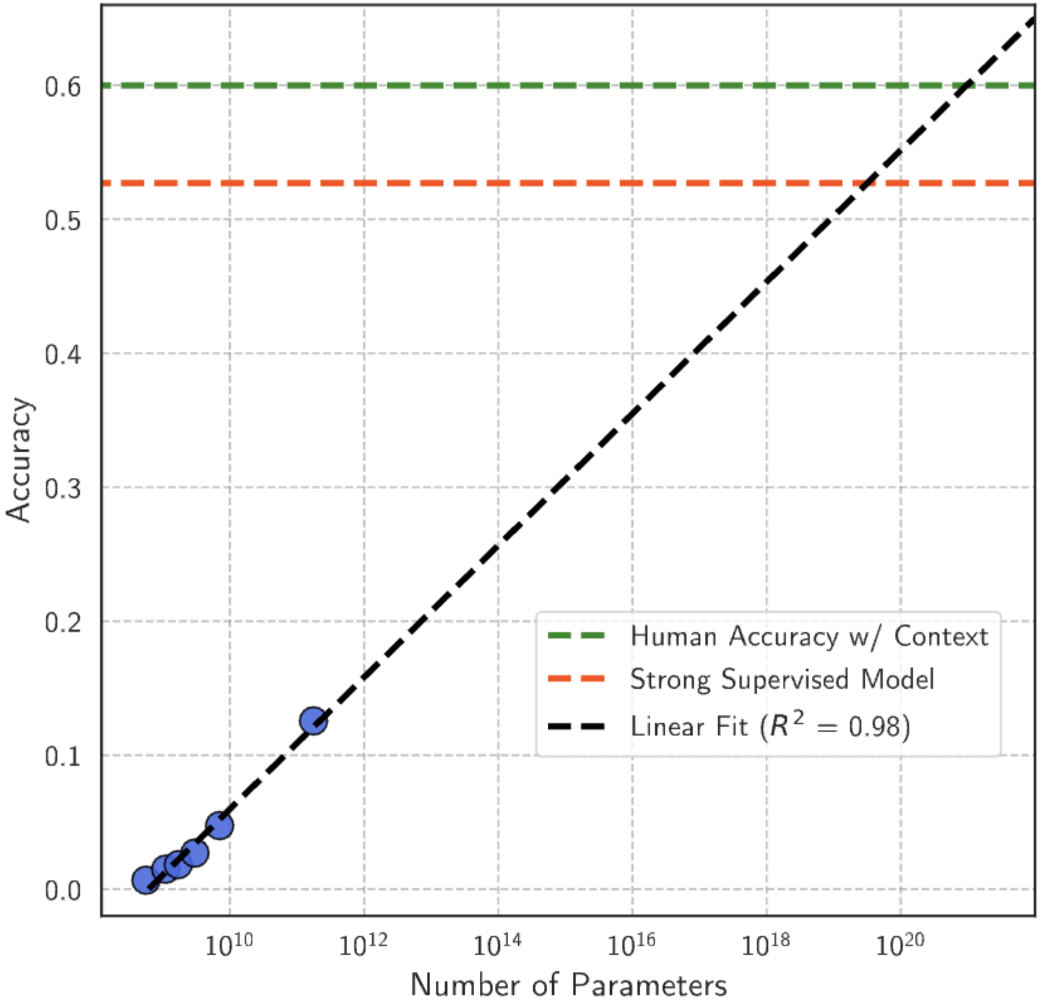


Large language models struggle to learn long-tail knowledge. Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. ICML 2023.

Memorization of Long-Tail Entities



Scaling Laws on Rare Facts

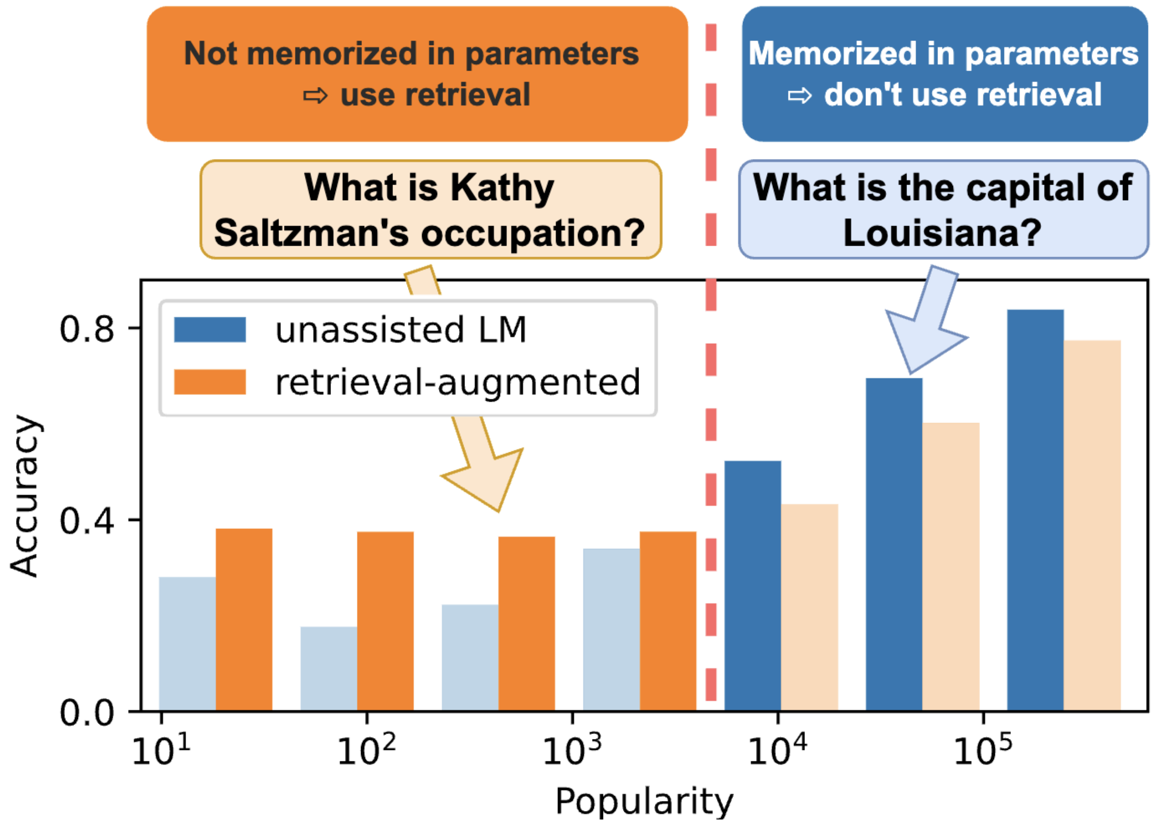


Can We Solve the Long-Tail Problem?

- **Problem**:
  - Knowledge about rare entities is **not memorized**
  - **Larger models** can memorize long-tail knowledge better
    - Scaling up models is unfeasible

- **Possible solutions:**
  - Use **retrieval-augmented models** instead of relying only on parametric knowledge
  - Improve the **training** behavior of models to better retain long-tail knowledge

RAG for Long-Tail Knowledge



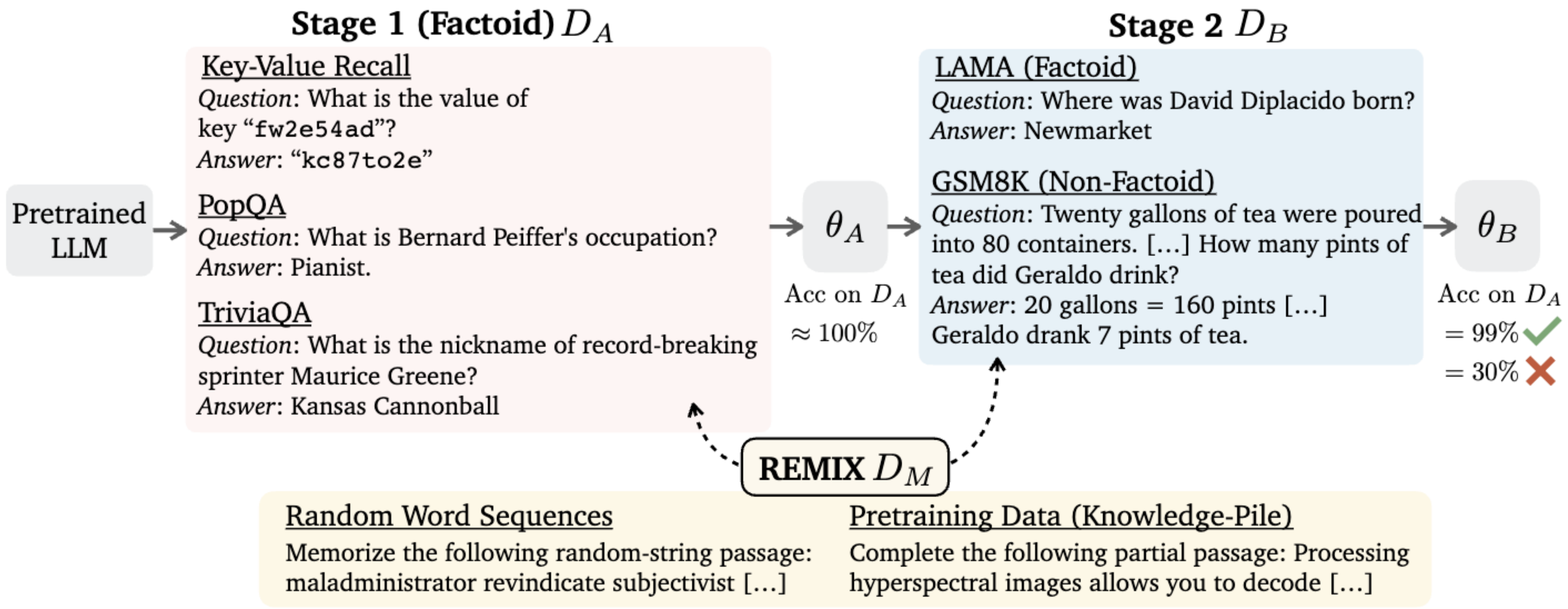
When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories. Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. ACL 2023.

Injecting Long-Tail Knowledge

- **Problem**:
  - LLMs forget long-tail factoids when trained on new datasets
- **Key Idea:**
  - Mix random or generic data during training to diversify fact storage.
  - Reduce interference between tasks in sequential training stages.

Continual Memorization of Factoids in Large Language Models. Chen, H., Geng, J., Bhaskar, A., Friedman, D., & Chen, D. (2024)..

Continual Memorization



Result:

- REMIX improves factoid retention (e.g., accuracy increases from 13.5% to 53.2%).
  - Facts are stored in earlier layers
- Enables LLMs to handle rare knowledge better without compromising performance on new tasks.

Conclusion

- Language Models are **not up-to-date**
  - **Retrieval-augmented models** might help here
- Language Models cannot deal with **numerical values**
- Entities are **just strings**
  - Google 2012 *“things, not strings”*
- Language Models have problems memorizing **long-tail knowledge**
  - The memorization of **popular entities** is extremely good
  - **RAG** and continual learning might overcome the issue

Language Models **cannot replace** Knowledge Graphs (yet),
but they are a **great tool for KG construction**