

Section 7

Challenges & Opportunities

Tutorial summary

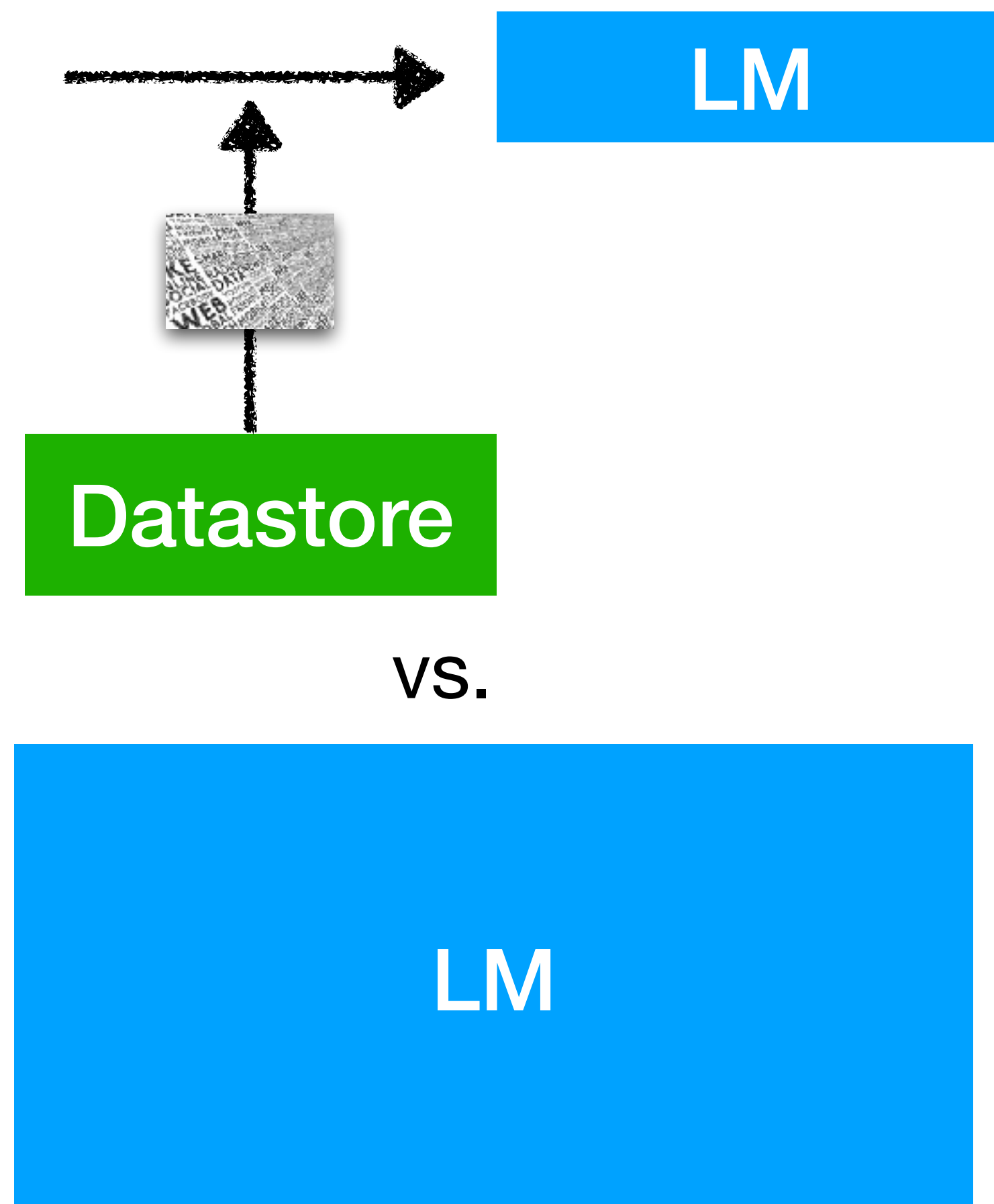
- **Definition & preliminaries**
- **Retrieval-based LMs: Architecture**
 - What to retrieve: tokens, text blocks, entity mentions
 - How to use retrieval: input layer, intermediate layers, output layers
 - When to retrieve: once, every n tokens, every token
- **Retrieval-based LMs: Training**
 - Independent training
 - Sequential training
 - Joint training with asynchronous index update
 - Joint training with in-batch approximation

Tutorial summary (cont'd)

- **Applications**
 - What tasks: knowledge-intensive NLP, code generation, classification etc
 - How to adapt: retrieval-based prompting, fine-tuning, reinforcement learning
 - When to use: long-tail, knowledge update, verifiability, parameter-efficiency, privacy, out of domain adaptations
- **Multilingual & multimodal extensions**
 - Multilingual: retrieve and generate cross-lingually to overcome scarcity of datastore
 - Multimodal: multimodal extensions of retrieval-based LMs for new modality

Challenge: Scaling retrieval-based LMs

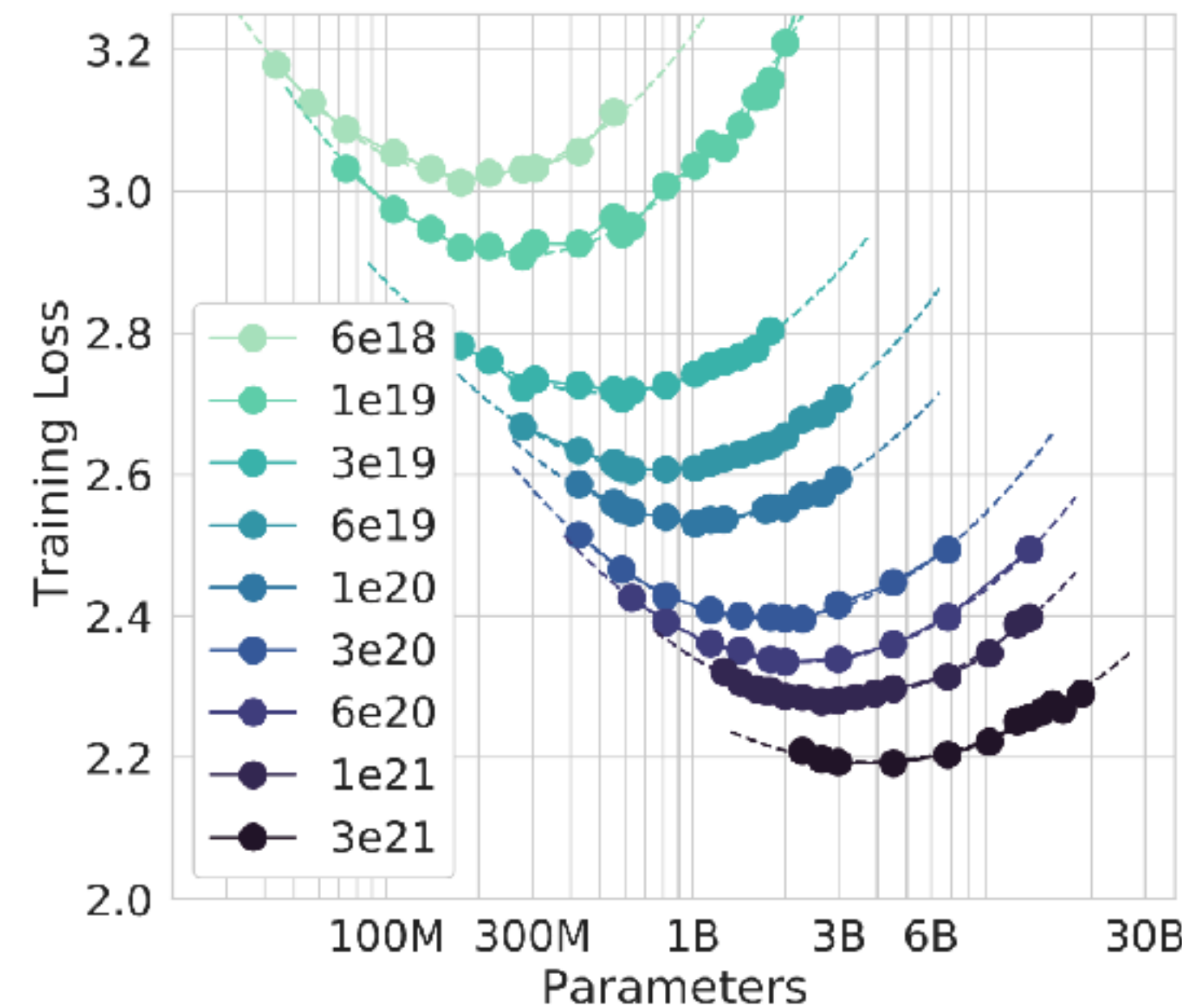
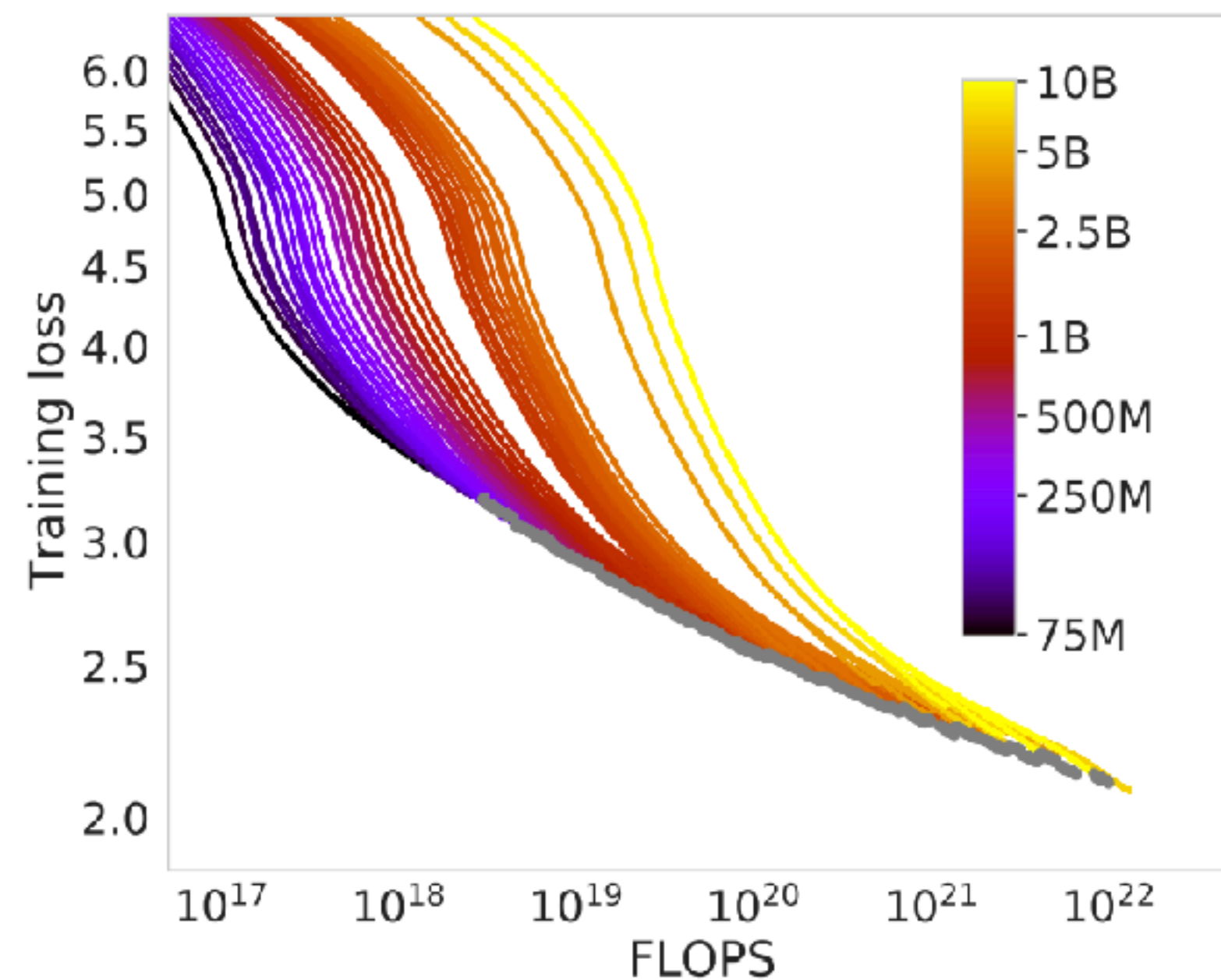
A small LM + a large datastore \approx a large parametric LM?



	LM	Datastore
	# of parameters	# of tokens
kNN-LM (Khandelwal et al., 2020)	250M	$\leq 3B$
NPM (Min et al., 2023)	350M	1B
Atlas (Izacard et al., 2022)	11B	$\sim 30B$
RETRO (Borgeaud et al., 2021)	7B	2T
REPLUG (Shi et al., 2023)	$\leq 175B$	$\sim 5B$

Challenge: Scaling retrieval-based LMs

Scaling law?



Loss as a function of:

- Training tokens
- Model sizes
- Training FLOPs

+ Datastore sizes?

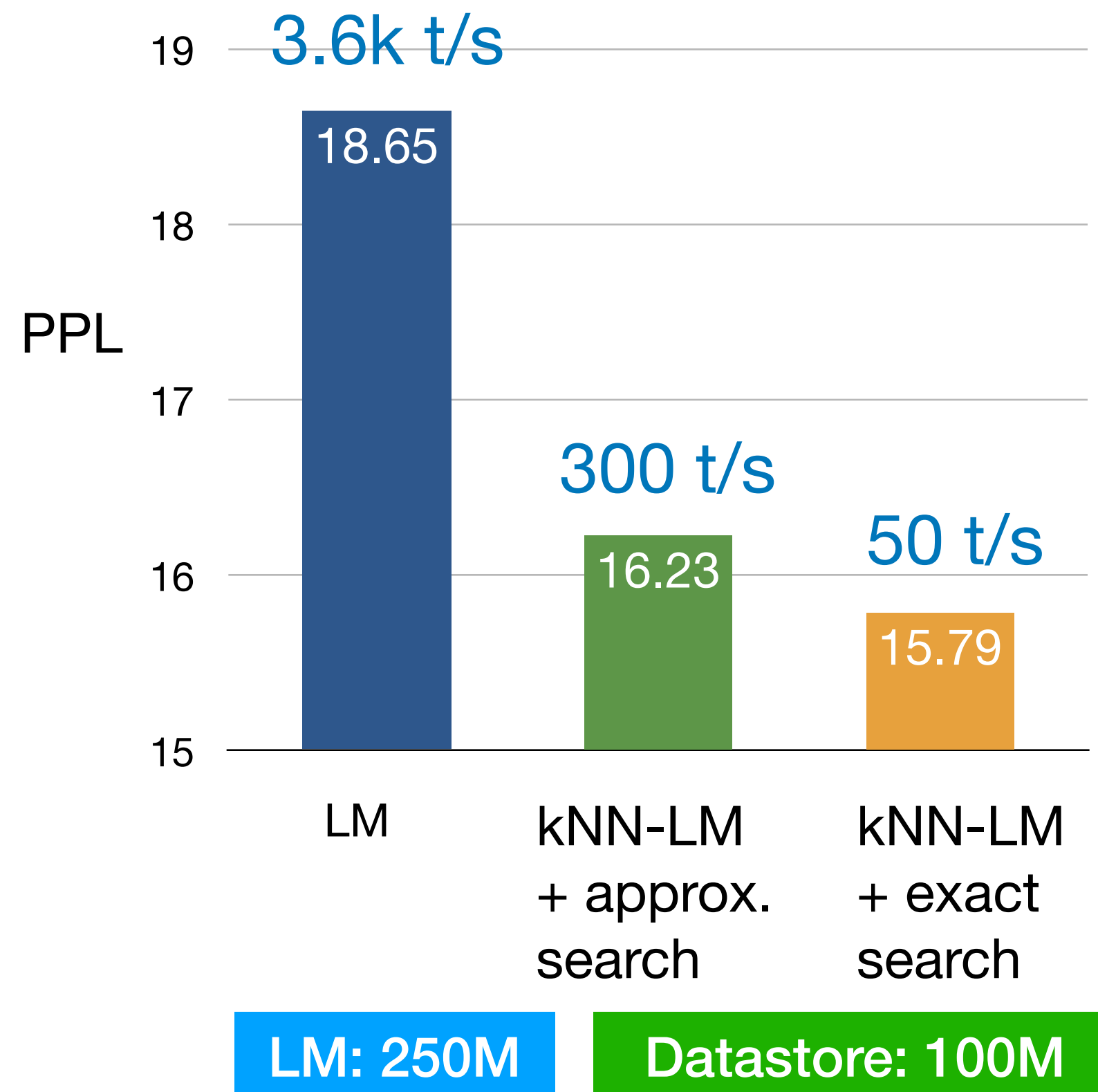
Dependent on architecture
and training methods!

Scaling law for parametric LMs (Kalpan et al., 2020; Hoffman et al., 2022)

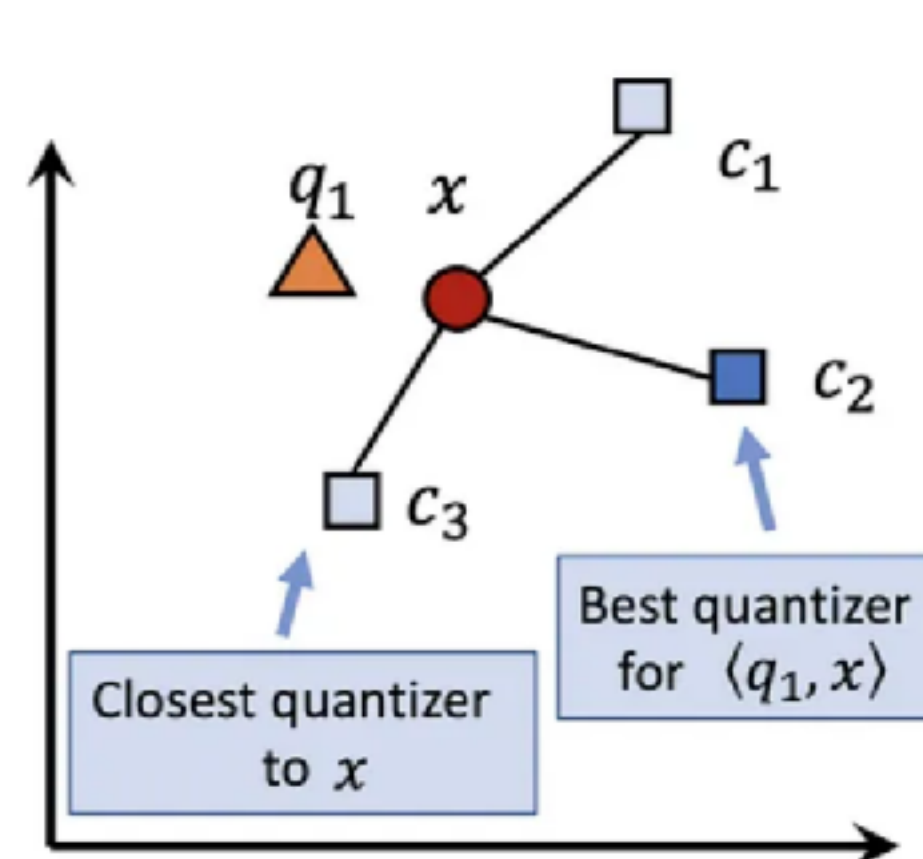
Challenge: Scaling retrieval-based LMs

Efficiency of similarity search

Measured on NVIDIA RTX 3090 GPU
(Zhong et al., 2022) with a FAISS indexer
(Johnson et al., 2021) with 32 CPUs



- >12 times slower even with **approximate** nearest neighbor search **implemented on GPUs**
- Efficient similarity search on GPUs/TPUs becomes the bottleneck of scaling retrieval-based LMs



Active research area!

Challenge: Scaling retrieval-based LMs

Efficiency of similarity search

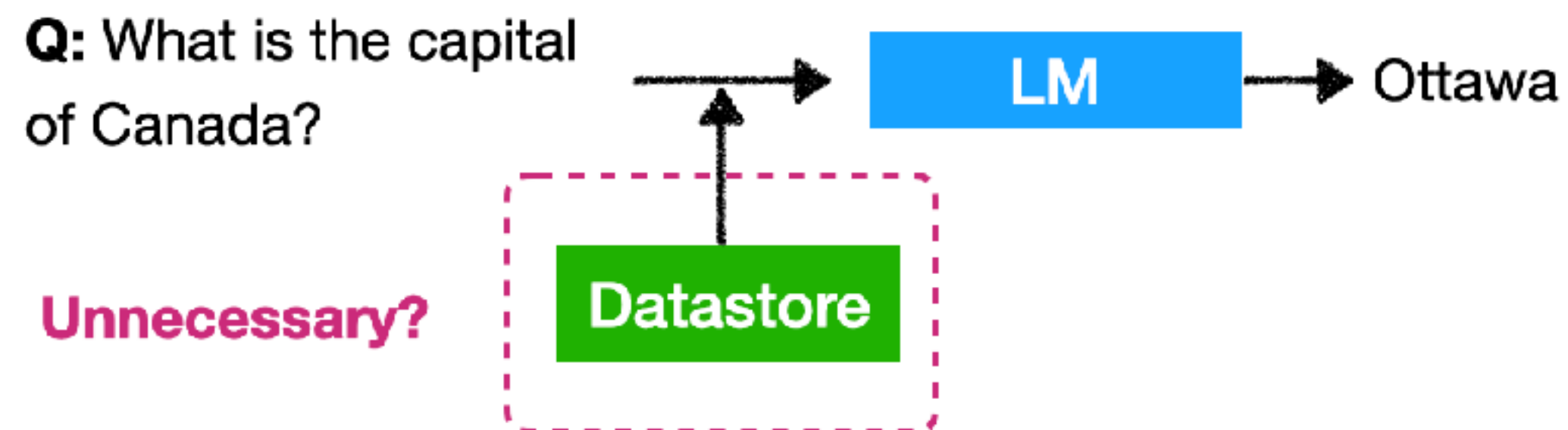
Efficient nearest neighbor search:

- Better loss for vector quantization (Guo et al., 2020)
- Optimization over different accelerators (Chern et al., 2022)
- Adaptive representations (Rege et al., 2023)

Middle ground: adaptively retrieve only when necessary (He et al., 2021; Mallen et al., 2023)

Space efficiency?

It also requires a large disk space for storage if the datastore becomes too large - storing **vectors** or **raw text**?



Challenge: Retrieval-based LMs for applications

Open-ended text generation?



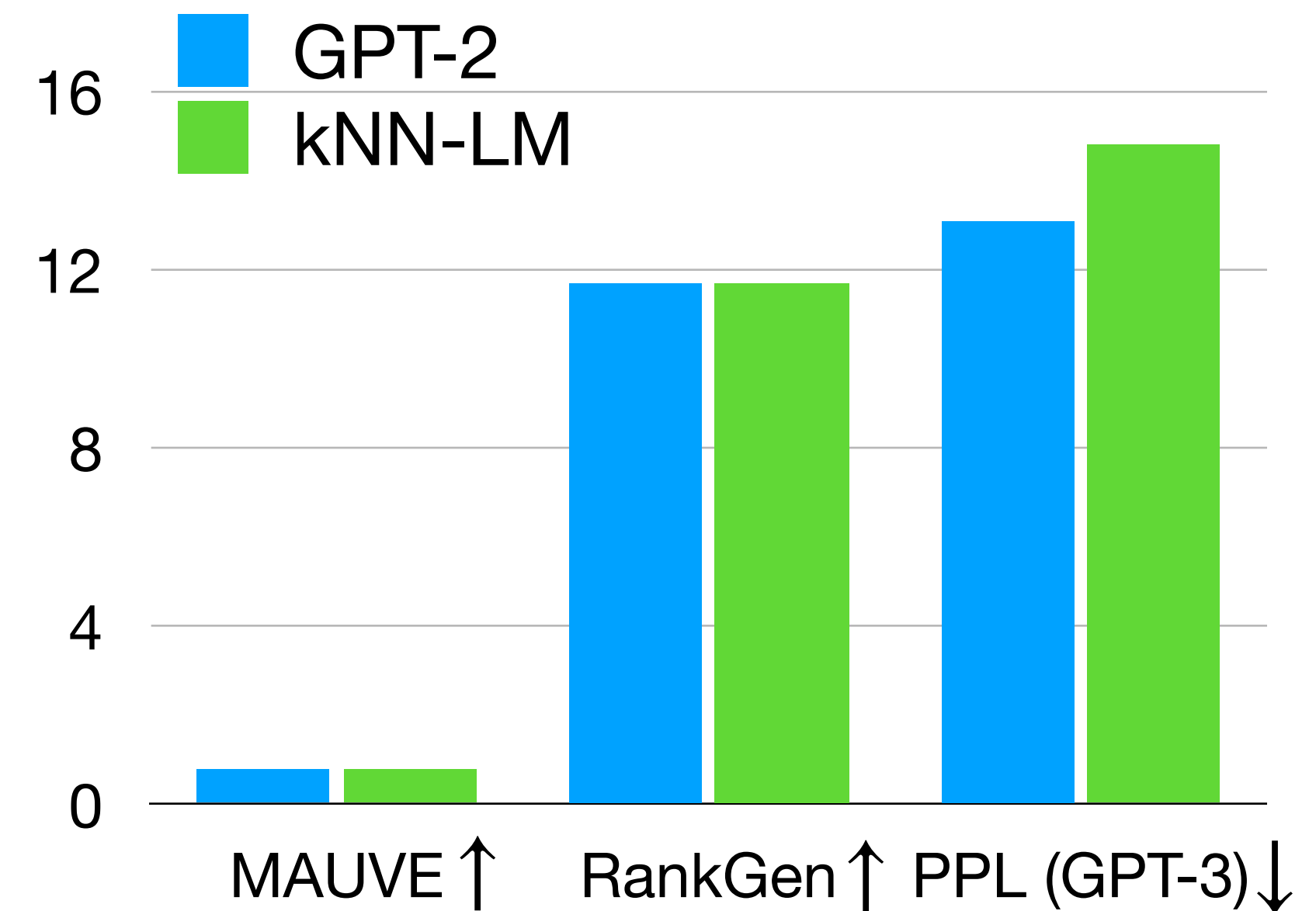
Hey I am going to ACL in Toronto next week. Can you tell me how I should explore the conference?



Certainly! Attending a conference like ACL (Association for Computational Linguistics) can be a fantastic opportunity to learn, network, and engage with the research community. Here are some tips to help you make the most of your time at the conference:

1. Review the schedule: Take some time to familiarize yourself with the conference program in advance. Note the sessions, workshops, tutorials, and keynotes that interest you the most. Identify any must-attend sessions and plan your schedule accordingly.
2. Create a personalized agenda: Once you have a clear understanding of the sessions you want to attend, create a personalized agenda. Note down the session names, locations, and times so you can easily navigate the conference venue and stay organized.
3. Engage in networking: Conferences provide an excellent opportunity to connect with fellow researchers and professionals in your field. Be proactive and introduce yourself to others during coffee breaks, poster sessions, or social events. Don't be afraid to strike up conversations, exchange ideas, and ask questions. Networking can lead to valuable collaborations and partnerships.
4. Attend keynote speeches: Keynote speeches are typically delivered by renowned experts in

Exposure bias? Better decoding algorithms?



Wang et al. 2023. “kNN-LM Does Not Improve Open-ended Text Generation”

Challenge: Retrieval-based LMs for applications

Complex reasoning tasks?

In New York State, the shortest period of daylight occurs during which month?

December is during the winter in the northern hemisphere.
New York state is a state located in the United States of America.
Winter has the least sunlight.
United States is located in the northern hemisphere.

REALM	December ✓
FiD	winter
kNN-LM	is during the winter in the northern hemisphere.
Flan-T5	january
ATLAS	Winter

Retrieval-based LMs struggle with **multi-step entailments** or **logical reasoning**:

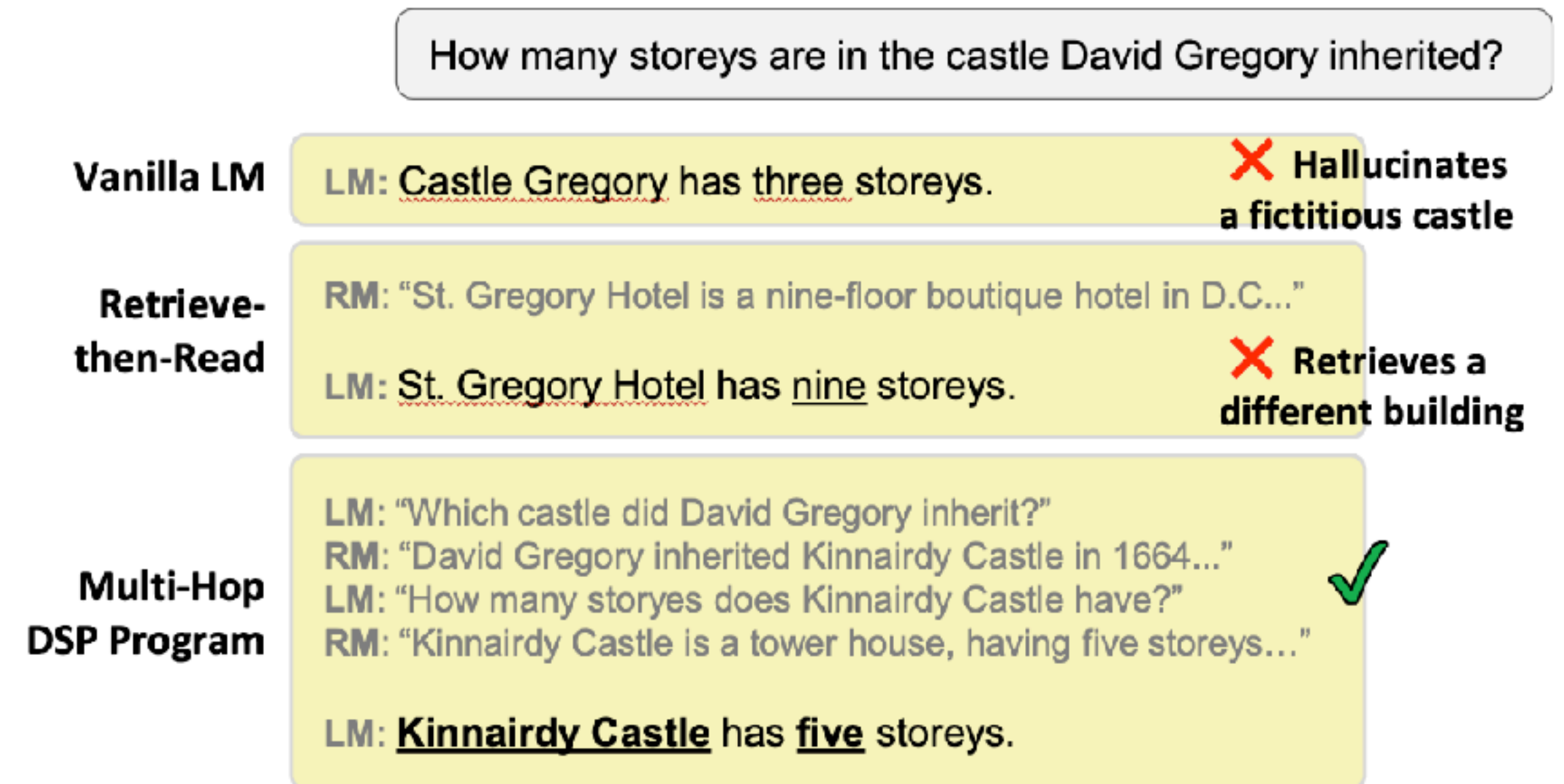
- Retrieving based on similarity of queries is insufficient 😞
- LMs also struggle with synthesizing information from multiple statements 😞

Challenge: Retrieval-based LMs for applications

Complex reasoning tasks?

Potential solutions?

- **Iterative retrieval & query reformulation**
 - FLARE (Jiang et al., 2023), Self-Ask (Press et al., 2022), ReAct (Yao et al., 2023)
- Decompose the task into a **multi-hop program** for complex reasoning
 - DSP: Demonstrate-Search-Predict



LM: Language Model, RM: Retrieval Model (both frozen)

Open questions

- What is the best **architecture & training method** for retrieval-based LMs in practice?
- We still don't know yet how to best **scale up** these models - Scaling law?
- We may need to explore alternative **decoding or adaptation methods** in downstream tasks (e.g., open-ended text generation, complex reasoning)!



Q & A

Thank you for joining us today!

All the materials are at
<https://acl2023-retrieval-lm.github.io/>