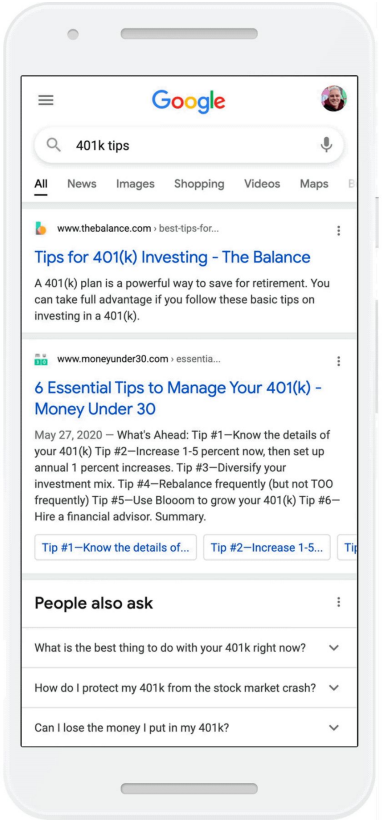
(Large) Language Models for Information Retrieval

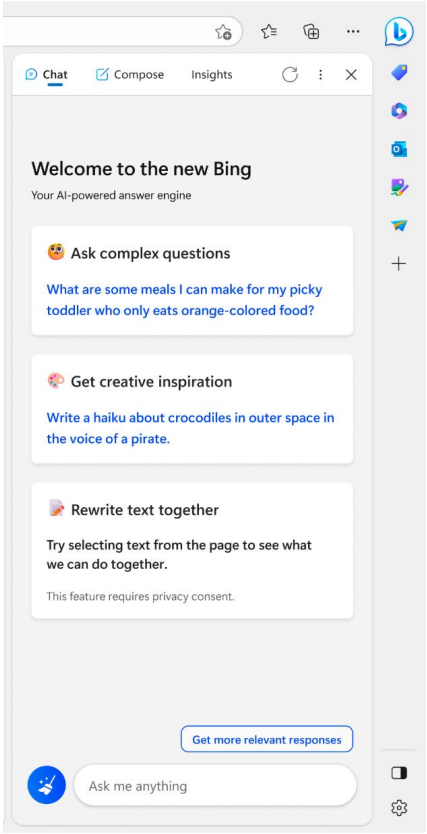
Diptesh Kanojia, Swapnil Bhosale, Hadeel Saadany, Constantin Orăsan
Samarth Agarwal, Zhe Wu

University of Surrey
eBay Search Science

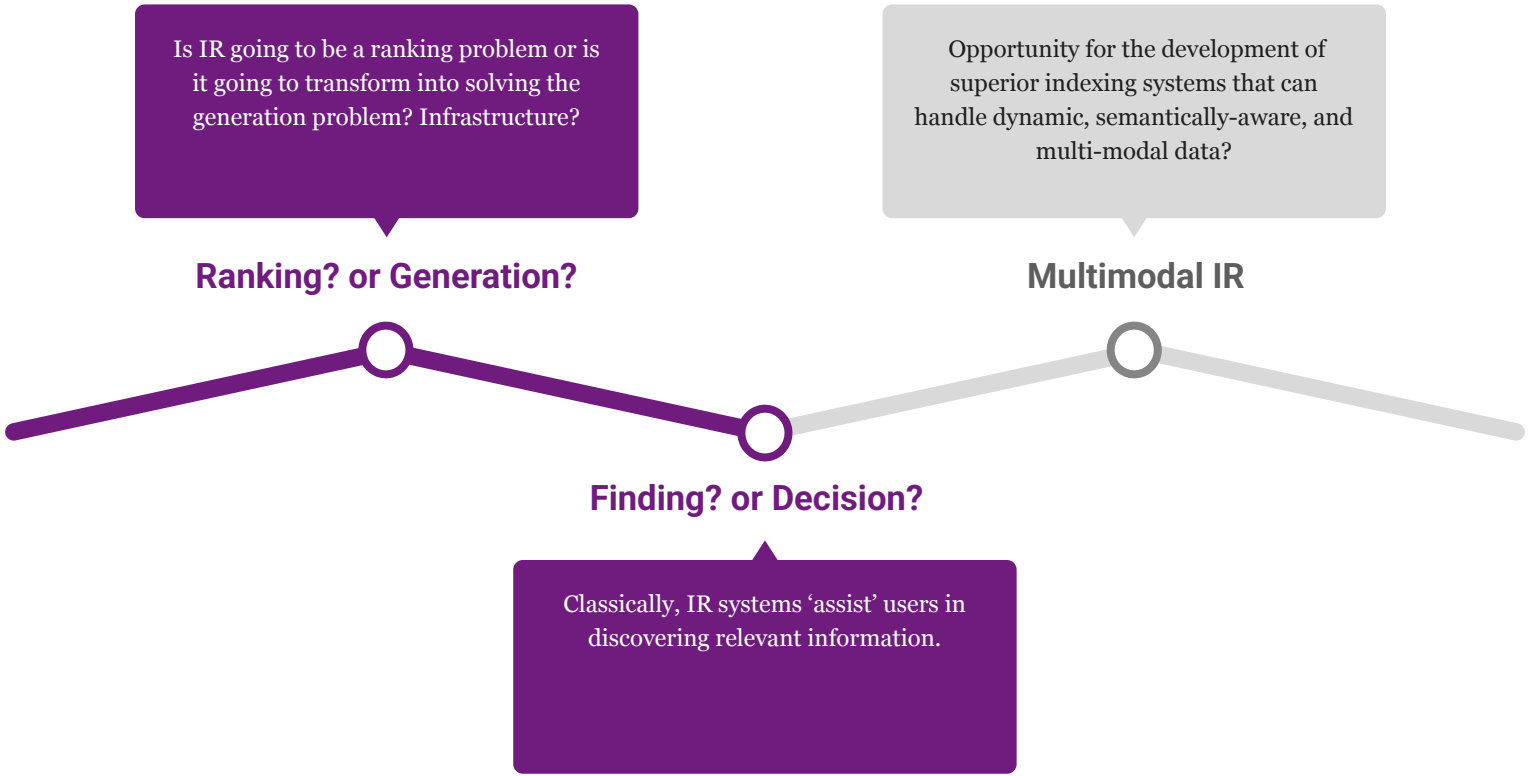
Information Retrieval (IR)

- Classic Retrieval
  - Focus on search aimed to assist users in finding relevant information.
  - Rise of Recommendation systems
- Evolution of IR
  - moving beyond merely retrieving relevant documents to meeting the information needs of users
  - understanding the user need as it evolves based on conversation
- The new Bing search incorporates retrieval within generated content
- Paradigm shift - new triangle - human, IR system, LLM
- Large language model for product search and recommendation?





Perspectives, Questions?



Language Modeling

- Classical language modeling
  - Concerned with prediction the probability of a sequence of words.
- Neural language modeling
  - Concerned with similar prediction but utilizes neural models for encoding text.
- Transformers architecture
  - Pre-trained Language Models (autoencoders)
  - Autoregressive Decoders

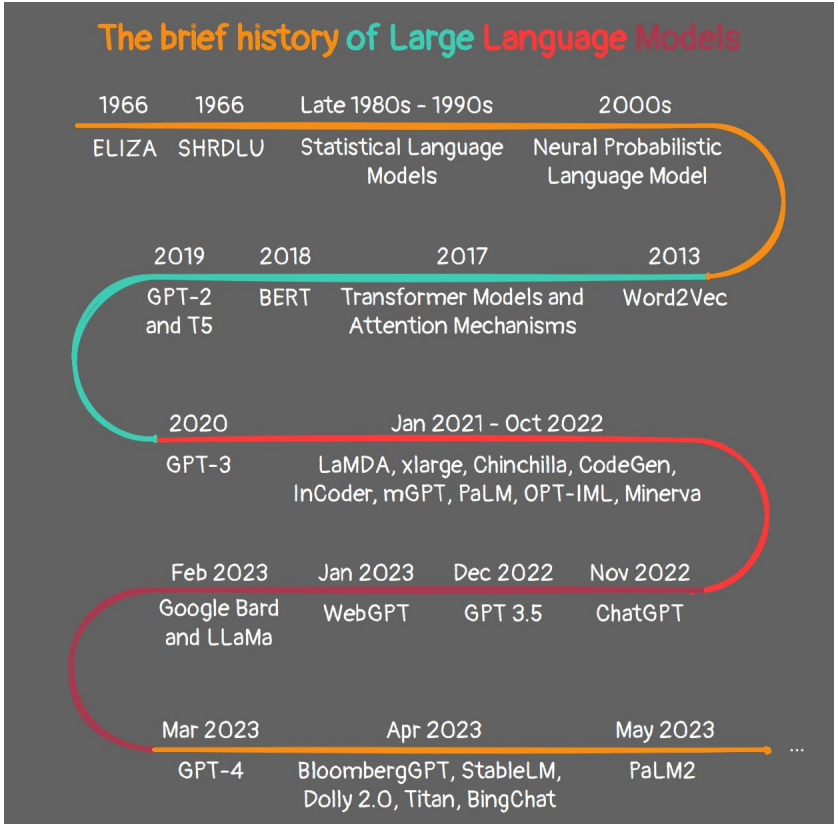
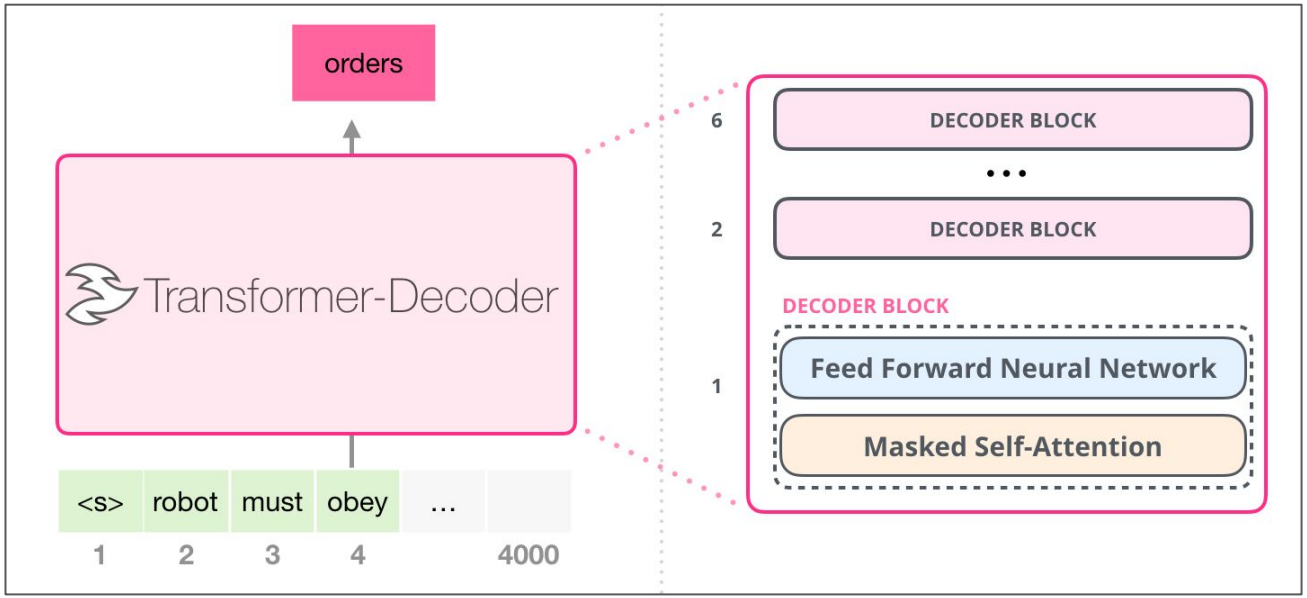


Figure source: https://towardsdatascience.com/a-brief-history-of-language-models-d9e4620e025b

Autoregressive Decoders | Generative AI



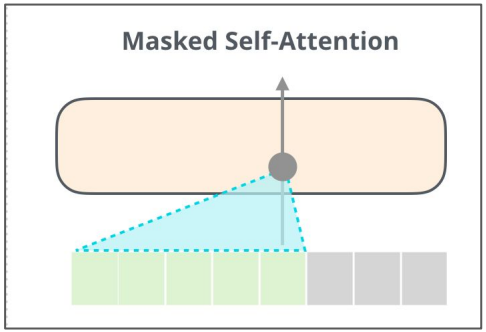
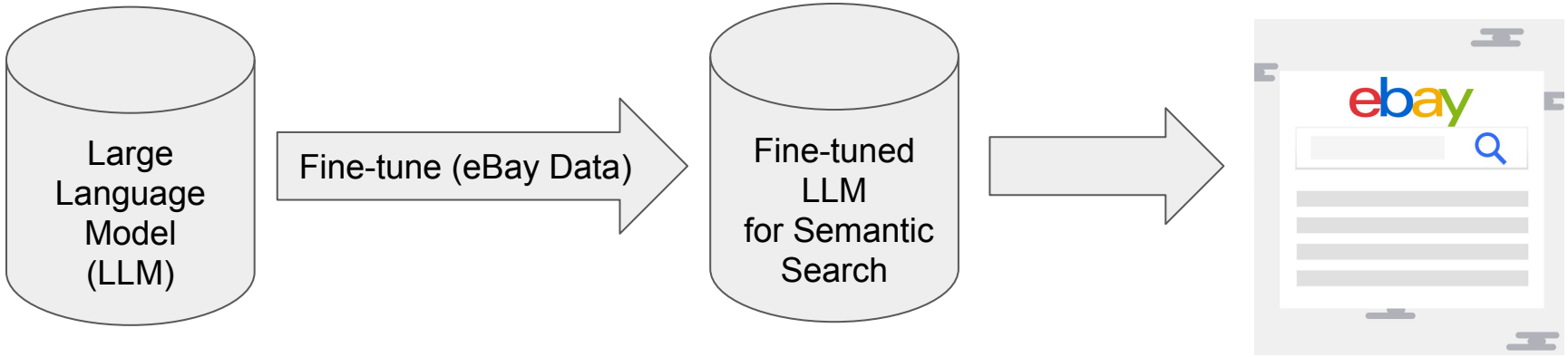


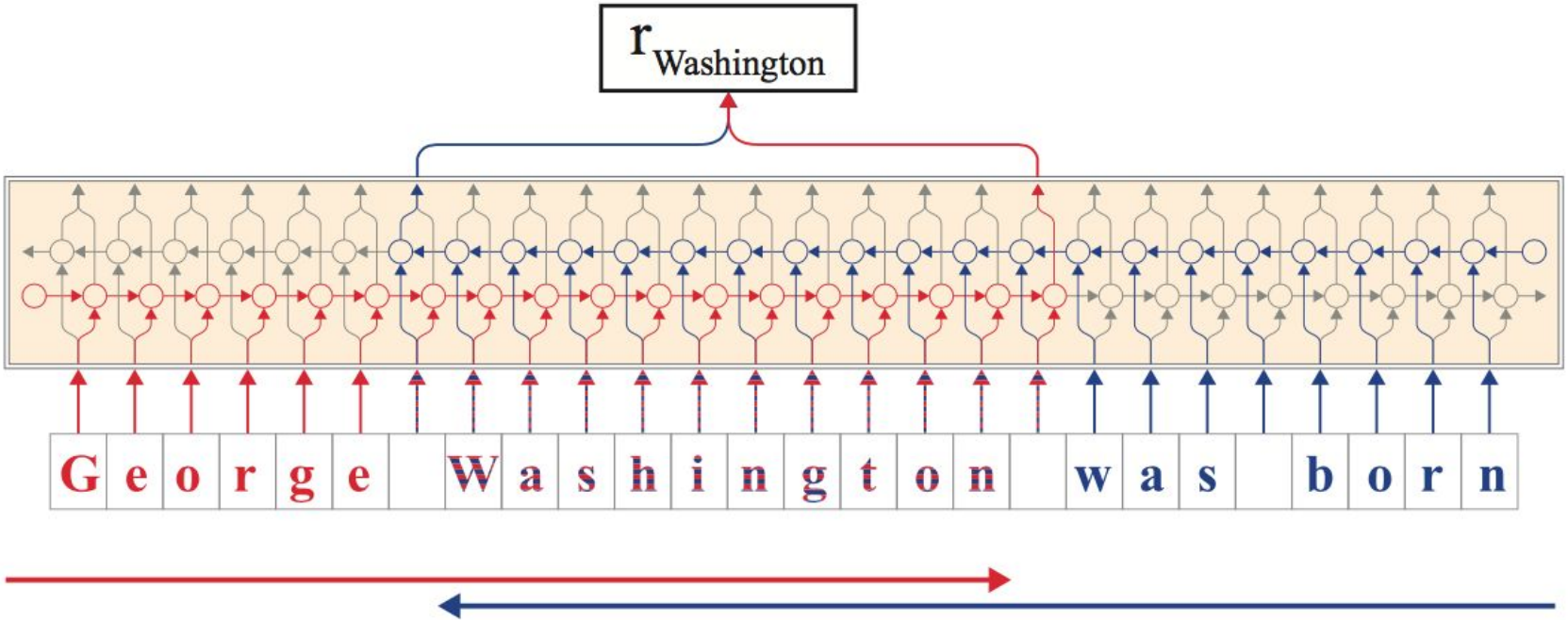
Figure source: http://jalammar.github.io/illustrated-gpt2/

Proposed Outline

Fine-tuned LLM for Product Search



Character-level Language Modeling



Akbik, A., Blythe, D. and Vollgraf, R., 2018, August. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th international conference on computational linguistics* (pp. 1638-1649).

Semantic Search

- Semantic: Understanding corpus and queries beyond keywords.
- Search: finding *top k* answers from a corpus given a query.

Why LLMs for Semantic Search

- **Enhanced precision**: Understand meaning and intent behind queries, makes the results contextually relevant
- **Improved efficiency**: faster retrieval
- **Multilingual support**: handle diverse linguistic contexts
- **Versatility**: web search, product search, Q&A, document retrieval

References: Muennighoff, Niklas. "Sgpt: Gpt sentence embeddings for semantic search." *arXiv preprint arXiv:2202.08904* (2022).

Approaches

Applying decoder-*only* transformers to Semantic Search.

1. Cross-encoder and Bi-encoder
2. Asymmetric Search and Symmetric Search

Bi-encoder: position-weighted mean pooling and contrastive bias only fine-tuning.

Cross-encoder: log-probability extraction of pre-trained GPT models (no fine-tuning).

Cross-encoders

- Encode query and document at the **same time**.
- Passed through the transformer together with [SEP] token.
- Each new query → k forward passes given a corpus of k documents.

Bi-encoders

- Encode query and document **separately**.
- Cache document embeddings.
- Each new query → single forward passes given a corpus of k documents (already cached).

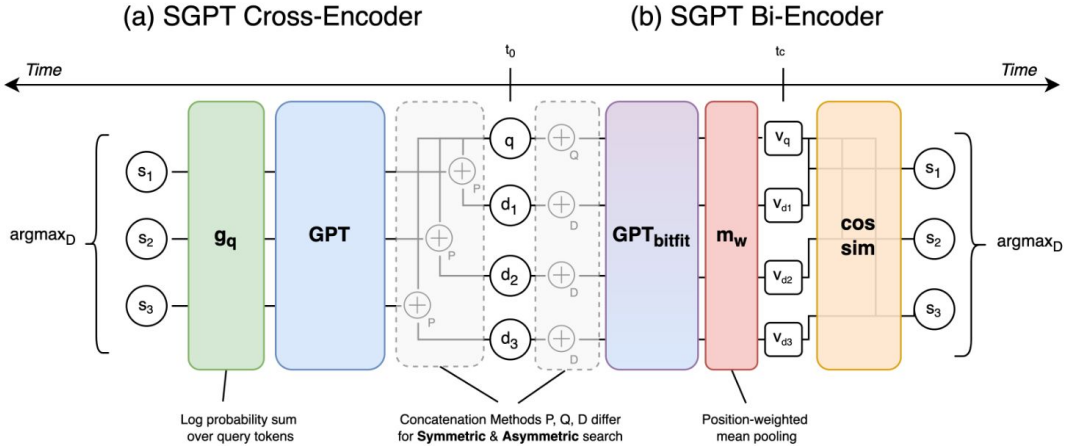


Figure 1: Given a query q, documents d_{1-3}, SGPT ranks the documents with scores s_{1-3}. (a) The Cross-Encoder concatenates queries and documents and encodes them together. Scores are extracted log probabilities. (b) The Bi-Encoder separately encodes queries and documents. Resulting document vectors v_{1-3} can be cached and retrieved at time t_c, when a new query comes in. Scores are cosine similarities.

Cross-encoders are slower but give higher performance.
Trade-off:

- two-stage re-ranking
- 1st model processes entire corpus (Bi-Encoder)
- 2nd model used on only top-k documents returned by first (Cross-Encoder)

References: Muennighoff, Niklas. "Sgpt: Gpt sentence embeddings for semantic search." *arXiv preprint arXiv:2202.08904* (2022).

Quora Question Pairs Dataset

Queries: 10000

Corpus: 523K

Relative D/Q: 1.6 (average relevant documents per query)

Annotation: Questions merged by users (may be noisy)

id	qid1	qid2	question1	question2	is_duplicate
447	895	896	What are natural numbers?	What is a least natural number?	0
1518	3037	3038	Which pizzas are the most popularly ordered pizzas on Domino's menu?	How many calories does a Dominos pizza have?	0
3272	6542	6543	How do you start a bakery?	How can one start a bakery business?	1
3362	6722	6723	Should I learn python or Java first?	If I had to choose between learning Java and Python, what should I choose to learn first?	1

References: https://quoradata.quora.com/First-Quora-Dataset-Release-Question-Pairs

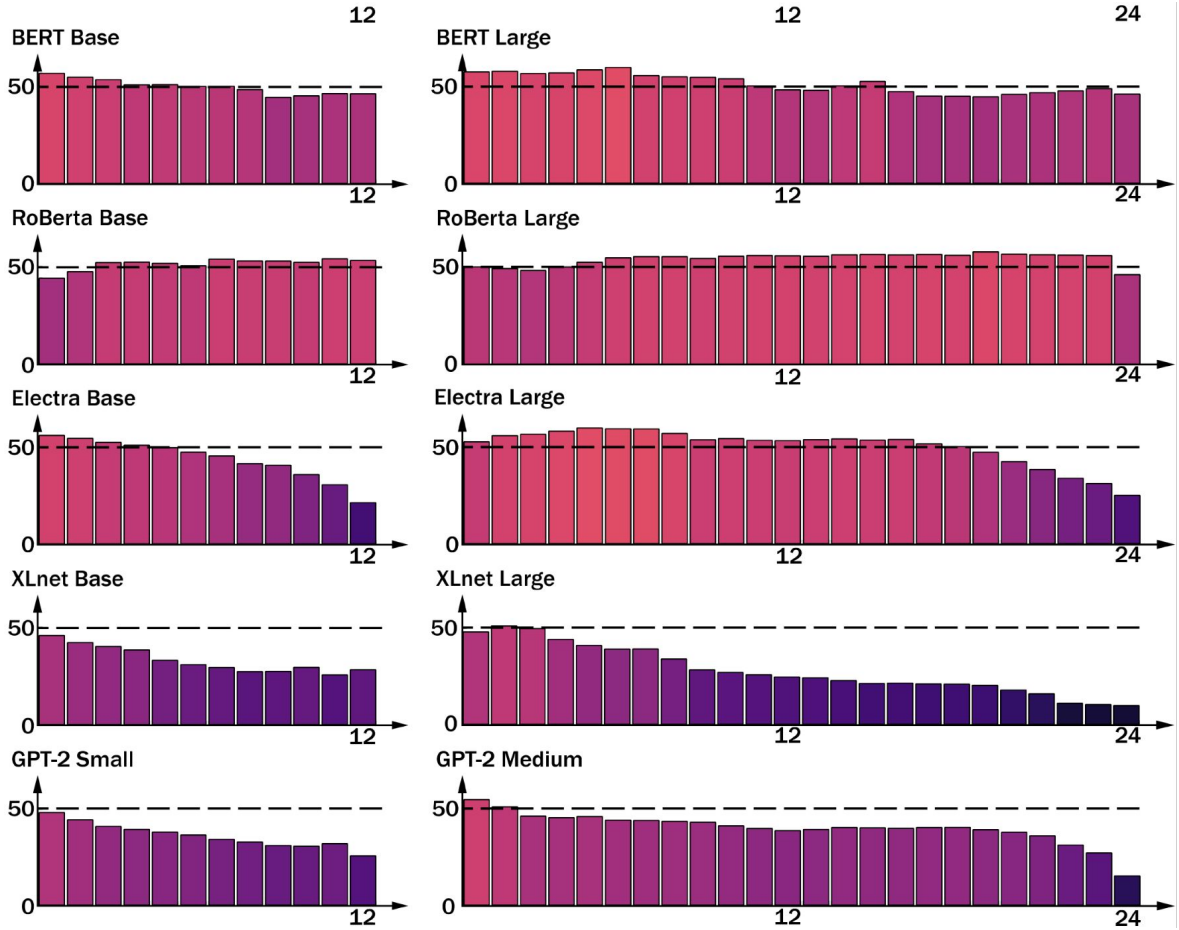
Symmetric Search

- Queries and documents are **interchangeable**.
- Finding duplicate answers.
- len(document) ~ len(query)
- BEIR: Quora QP

Prompts	**125M**
Questions are searched to find matches with the same content.\nThe question "{doc}" is a good search result for "{query}	0.758
These two questions are the same: 1. {doc} 2.{query}	0.734
Question Body: {doc} Question Title:{query}	**0.771**

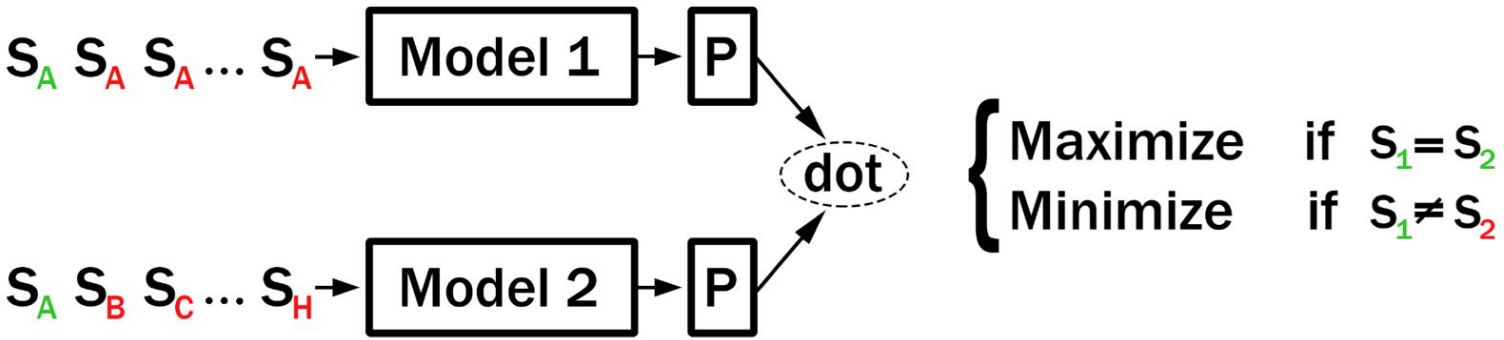
Dataset	**SGPT-CE Re-rank Top 10 (1.3B)**	**SGPT-CE Re-rank Top 100 (1.3B)**
Quora QP	0.792	0.797

LLM Task Bias



Layer-wise unsupervised STS performance on the STS-b test set. X-axis denotes the layers of the depicted model and the Y-axis denotes the Spearman correlation (x100).
Carlsson, F., Gyllensten, A.C., Gogoulou, E., Hellqvist, E.Y. and Sahlgren, M., 2020, October. ***Semantic re-tuning with contrastive tension***.

In-Batch Negatives with Contrastive Tension (**self-supervised learning**)



Carlsson, F., Gyllensten, A.C., Gogoulou, E., Hellqvist, E.Y. and Sahlgren, M., 2020, October. ***Semantic re-tuning with contrastive tension***. In International conference on learning representations.

Table 2: Pearson and Spearman correlation (x100) on the STS-b test set.

Not trained for STS		**Trained with STS-b data**		
		Regression Labels	[0, 1]	[M, 1]
BERT-Base	47.91 / 47.29	BERT-Distil	84.07 / 84.23	85.02 / 85.54
InferSent-GloVe	65.30 / 63.21	BERT-Base	85.28 / 84.99	85.11 / 85.64
USE v4	78.73 / 77.09	BERT-Large	85.54 / 85.37	85.90 / 86.35
S-BERT-Distil	73.88 / 76.19	S-BERT-Distil	84.22 / 84.26	85.40 / 85.64
S-BERT-Base	74.15 / 76.98	S-BERT-Base	85.17 / 84.90	85.59 / 85.81
S-BERT-Large	76.16 / 79.19	S-BERT-Large	85.14 / 85.07	85.25 / 86.28
Our contributions				
BERT-Distil-CT	79.00 / 78.56	BERT-Distil-CT	84.14 / 84.19	85.32 / 85.82
BERT-Base-CT	77.87 / 76.32	BERT-Base-CT	85.13 / 84.92	85.76 / 85.89
BERT-Large-CT	79.97 / 78.99	BERT-Large-CT	85.20 / 84.97	**86.37** / 85.89
S-BERT-Base-CT	76.25 / 80.11	S-BERT-Distil-CT	80.09 / 84.27	85.61 / 85.80
S-BERT-Base-CT	78.83 / 81.24	S-BERT-Base-CT	85.26 / 85.20	85.72 / 85.95
S-BERT-Large-CT	**80.99 / 82.14**	S-BERT-Large-CT	85.36 / 85.16	86.09 / **86.43**

Road Ahead

- Utilize LLMs for enhanced semantic search
- Semantic re-tuning for performance improvement
- Character-level Language models for domain specific search

Thank you!

Questions?

References

Akbik, A., Blythe, D. and Vollgraf, R., 2018, August. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th international conference on computational linguistics* (pp. 1638-1649).

Muennighoff, Niklas. "Sgpt: Gpt sentence embeddings for semantic search." *arXiv preprint arXiv:2202.08904* (2022).

Carlsson, F., Gyllensten, A.C., Gogoulou, E., Hellqvist, E.Y. and Sahlgren, M., 2020, October. ***Semantic re-tuning with contrastive tension***. In International conference on learning representations.

Challenges Ahead

Model Deployment

Credibility Concerns

- Retrieval-augmented LLMs

Ethical Concerns

Legal Concerns