Porsche Digital Campus

AI and ChatGPT

MIDTERM REVIEW

Team: TW@Porsche

Our Team

Po-Yen Chen
Simulation Sciences, RWTH Aachen
Data and numerical analysis, implement the data-driven method to chemical and mechanical engineering problem, did more than related three interdisciplinarily internal and external projects during my study

Ling-Chia Chen
Computer Science, University of Stuttgart
NLP model fine-tuning, implement Voice Recognition and Language Model to an AI customer service system, design and develop projects related to system automation.

Chia Hao Chang (Gary)
Simulation Sciences, RWTH Aachen
Deep Learning in NLP and Computer Vision, Reinforcement learning

- Utilize LLM in computational argumentation
- Computer vision working student at RWTH ISAC lab to analyze traffic flow via end-to-end object detectors.

Outline

- Introduction
- Methodology
- Current work
- Future work

Outline

- **Introduction**
- Methodology
- Current work
- Future work

Problem Understanding

In general, why do we need to fine tuning text embedding models?

- Issue 1: Correct answer but low similarity

```
User Query 1:
What is the unique feature of the Cayenne Turbo GT?

Answer 1:
It offers 471 kW (640 PS) and is optimized for high
performance.

Similarity score: 0.4106892943382263
```

- Issue 2: Wrong answer but high similarity

```
User Query 2:
What is the history of Cayenne?

Answer 2:
The history of Cayenne can be first traced back to
pre-Columbian times.

Similarity score: 0.8268212676048279
```

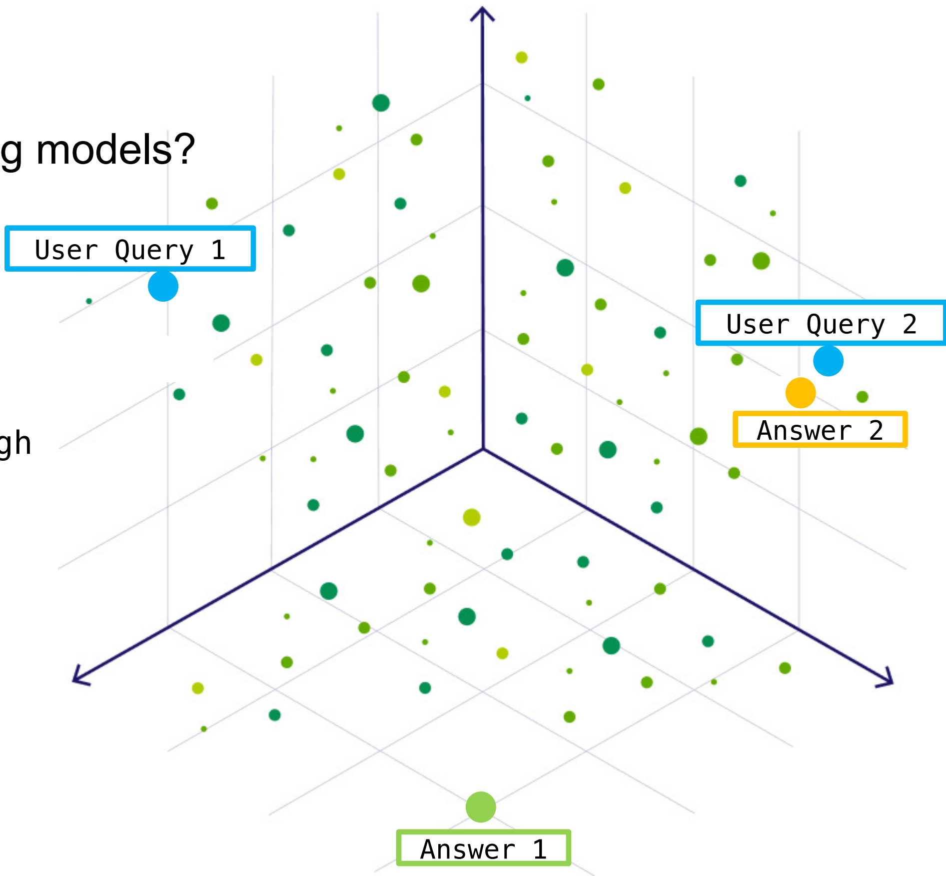


Image modified from https://weaviate.io/blog/vector-embeddings-explained

Idea proposal – Car recommendation system

- Build a recommendation engine that suggests particular car models, based on the customer‘s preferences and needs, helping potential buyers to have better decision-making.

- First, represent products in embedding space. Then, customers can input their requirements, such as budget, preferred body type, fuel efficiency and desired features. Text embedding model can then use similarity search to generate personalized recommendations.

(Create by DALL-E)

Outline

- Introduction
- **Methodology**
- Current work
- Future work

Our Solution Approach

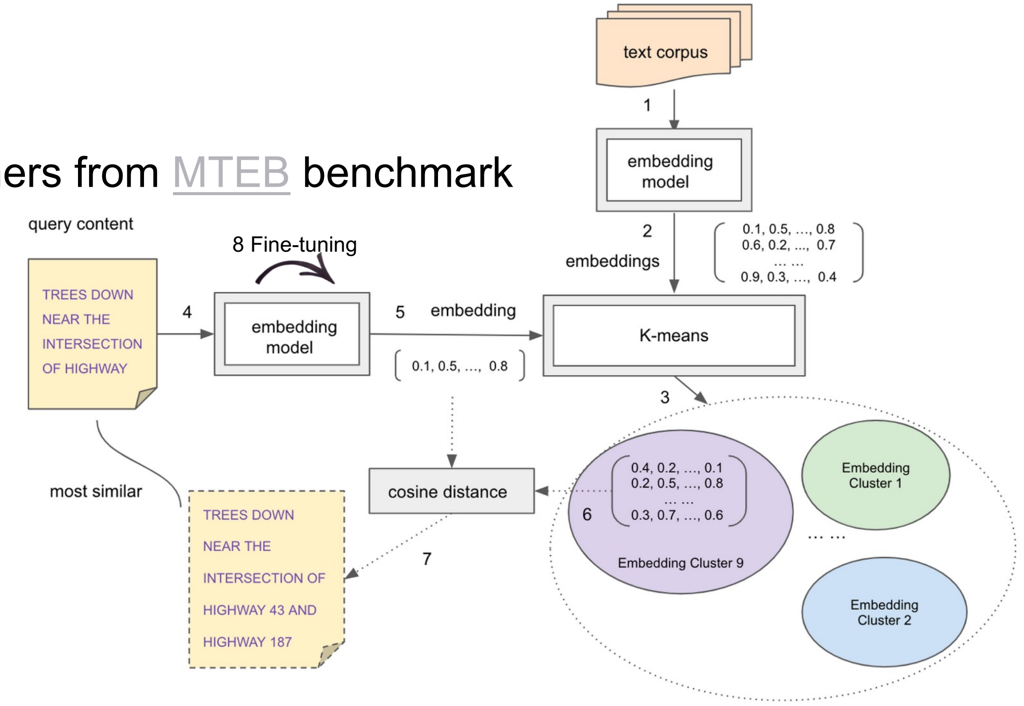
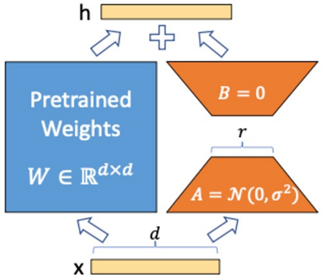


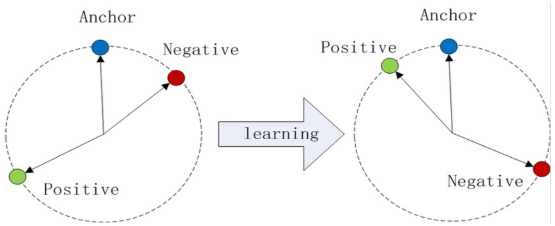
Image source: https://cloud.google.com/blog/products/data-analytics/introducing-bigquery-text-embeddings

1. **Text embedding model**: sentence-transformers from MTEB benchmark
   - Both text corpus related to Porsche car models and the queries are embedded into latent space

2. **Similarity search**:
   - ANN (approximate nearest neighbor) e.g. HNSW, Annoy, FAISS, and NMSLib

3. **Fine-tuning**:
   - Adapter + PEFT (LoRA/QLoRA)
   - Contrastive Loss



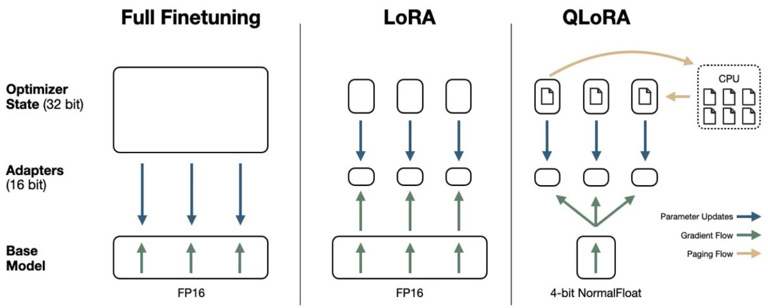


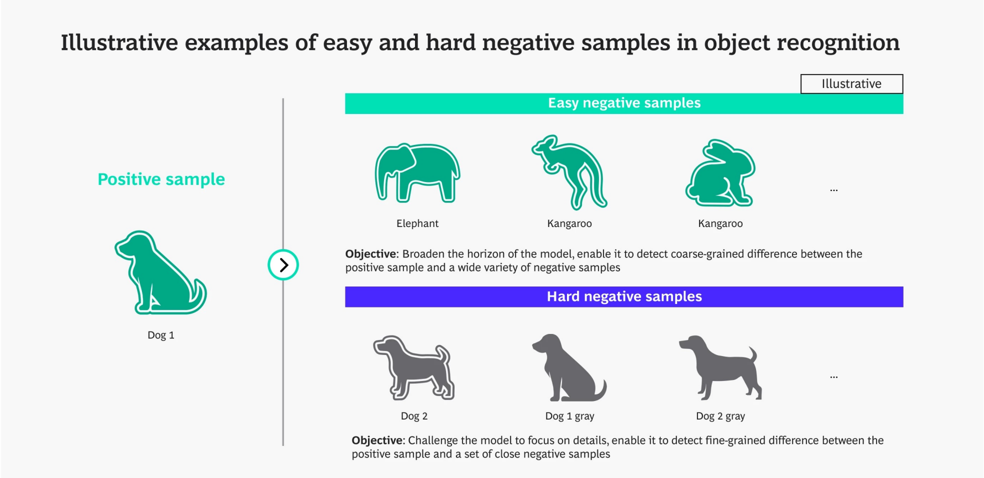
Our Solution Approach (cont'd)

- More detailed about fine tuning approaches

Key ingredients behind **QLoRA**:
(1) 4-bit NormalFloat
(2) Double Quantization
(3) Paged optimizers (avoid out-of-memory error during training)
(4) LoRA: Low-rank Adaptation

Contrastive loss
- Random negative sampling





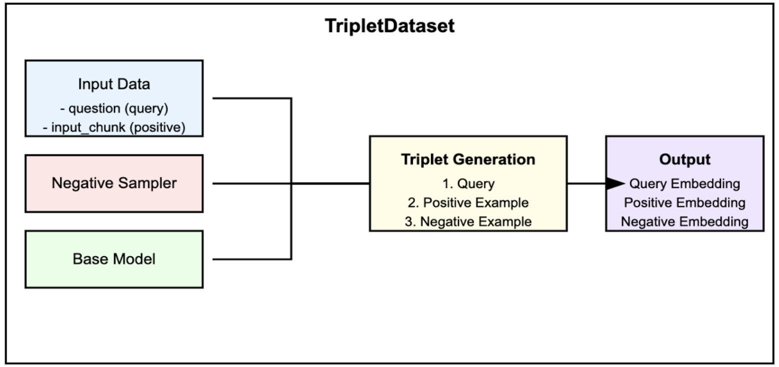
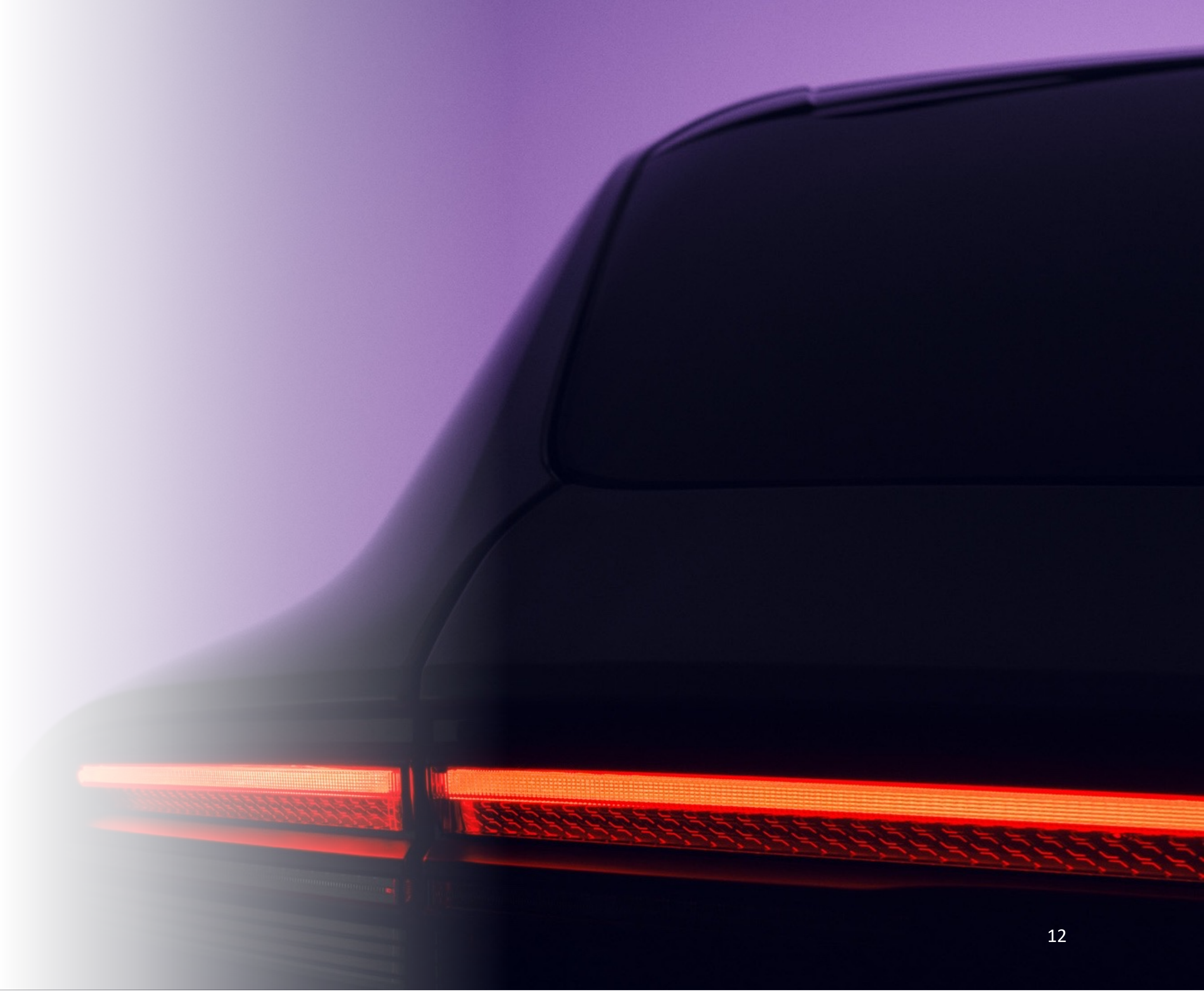


Image source: https://cloud.google.com/blog/products/data-analytics/introducing-bigquery-text-embeddings

Outline

- Introduction
- Methodology
- **Current work**
- Future work



Up until now…

- Data collection for fine-tuning:
  - Web scrapping all Porsche vehicle model specification and latest newsroom for each car model
  - Synthetic data generated by knowledge base model (RAG) from Amazon AWS Bedrock (we will, of course, double-check the correctness)

"question_1": "What is the combined CO_2 emissions for the 718 Cayman GTS 4.0?",
"pos_ans_1" : "The combined CO2 emissions for the 718 Cayman GTS 4.0 is 246 – 219 g/km.",
”neg_ans_1" : " The 718 Cayman GTS 4.0 produces zero CO_2 emissions because it is fully electric."

→ So far, we have collected 250 QA pairs for fine-tuning.

Up until now…

- Embedding models selection from Hugging face MTEB leaderboard

Rank	Model	Model Size (Million Parameters)	Memory Usage (GB, fp32)	Embedding Dimensions	Max Tokens	Average (56 datasets)	Classification Average (12 datasets)	Clustering Average (11 datasets)	PairClassification Average (3 datasets)
1	NV-Embed-v2	7851	29.25	4096	32768	72.31	90.37	58.46	88.67
2	bge-en-icl	7111	26.49	4096	32768	71.67	88.95	57.89	88.14
3	stella_en_v5	1543	5.75	8192	131072	71.19	87.63	57.69	88.07
4	stella_en_1.5B_v5	1543	5.75	8192	131072	71.19	87.63	57.69	88.07
5	SFR-Embedding-2_R	7111	26.49	4096	32768	70.31	89.05	56.17	88.07
6	gte-Qwen2-7B-instruct-Q4_K_M-					70.24	86.58	56.92	85.79
7	gte-Qwen2-7B-instruct-Q2_K-GG					70.24	86.58	56.92	85.79
8	gte-Qwen2-7B-instruct	7613	28.36	3584	131072	70.24	86.58	56.92	85.79
9	gte-Qwen2-7B-instruct-Q4_K_M-					70.24	86.58	56.92	85.79
10	stella en 400M v5	435	1.62	8192	8192	70.11	86.67	56.7	87.74

→ So far, we have only tried “stella en 400M v5”, but in the future, we’ll also try other in comparison.

Porsche Digital Campus

Outline

- Introduction
- Methodology
- Current work
- **Future work**

Future work

1. Fine-tuning the embedding model
   - Replicate some research papers from TUM and ChromaDB
   - Loss functions and evaluation methods from SBERT library
   - Compare the similarity score, for example, with the pre-trained embedding model

2. Extension to recommendation systems
   - Get some example ideas from OpenAI cookbook and try to extend to be Porsche-specific
   - Use open-source LLM (e.g. LLaMA or Falcon) to generate response
   - Some others that might possibly be used:
     - Orchestration framework like LlamaIndex or Langchain
     - Vector database, e.g. Chroma
     - GUI interface, e.g. Gradio or Streamlit

Expected results

If I need larger car trunk capacity, which car model would you recommend me?"

Before fine-tuning

Based on the available search results, I cannot provide a comprehensive comparison of trunk or cargo space across all Porsche models. The search results only mention that some models like the Taycan and 718 have two luggage compartments, but do not specify their exact volumes. However, without specific cargo space measurements, I cannot definitively state which model has the largest trunk space.

After fine-tuning

The Cayenne is Porsche's largest SUV, and it's also the one that offers the most cargo space. The capacity specification is as follows:

Luggage compartment volume, rear V210-2
698 l

Open luggage compartment volume (up to the upper edge of the rear seats)
772 l

Largest luggage compartment volume (behind front seats, up to roof)
1,708 l

Potential challenges

1. Insufficient data

2. Computational resource limitations

Reference

1. Schopf, T., Schneider, D. N., & Matthes, F. (2023). Efficient Domain Adaptation of Sentence Embeddings Using Adapters. ArXiv, 1046–1053. https://doi.org/10.26615/978-954-452-092-2_112

2. Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R., & Wei, F. (2023, December 31). Improving Text Embeddings with Large Language Models. arXiv.org. https://arxiv.org/abs/2401.00368

3. Embedding adapters. (n.d.). Chroma Research. https://research.trychroma.com/embedding-adapters

4. Paluszkiewicz, A. (2023, September 14). The impact of AI on the automotive industry. Digica | AI Powered Software. https://www.digica.com/blog/the-impact-of-ai-on-the-automotive-industry.html

5. Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023, May 23). QLORA: Efficient Finetuning of Quantized LLMS. arXiv.org. https://arxiv.org/abs/2305.14314

6. Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021, June 17). LORA: Low-Rank adaptation of Large Language Models. arXiv.org. https://arxiv.org/abs/2106.09685


Porsche Digital Campus
S LA 911
THANK YOU.
FEEDBACK?