Porsche Digital Campus

AI and ChatGPT

IDEAS PROPOSAL AND FEEDBACK DISCUSSION

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.

Some questions before getting start…

- **Fine-tune** a **<u>text embedding model</u>** to make it **Porsche-specific**.

- Our initial proposal, in-car voice assistants, mainly focus on finding state-of-the-art fine-tuning LLM approaches. The task only focuses on the text embedding model? Can we fine tuning the LLM?

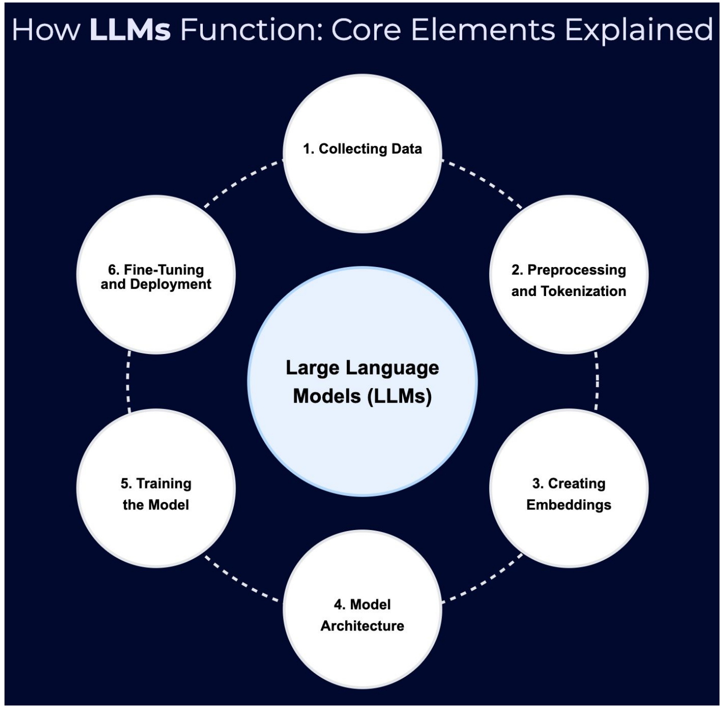


Image source: https://www.confidentialmind.com/post/embeddings-and-llms#

Agenda

- Idea 1: Hey Porsche - In-car voice assistants (Initial Proposal)
    - Problem understanding
    - Solution approach
    - Implementation details and Potential Challenges

- Idea 2: MatchMyPorsche - Product recommendation system
    - Problem understanding
    - Solution approach
    - Implementation details and Potential Challenges

- Idea 3: Porsche Pitstop - Customer support automation
    - Problem understanding
    - Solution approach
    - Implementation details and Potential Challenges


Porsche Digital Campus
Idea 1:
Hey Porsche
5

Problem Understanding

- **Topic: In-car Voice Assistants**
    - Modern in-car conversational systems still have large room for advancements in recognition accuracy, personalization, and overall driver experience
- **Reasoning**: Stay competitive in this LLM boom
    - Many automotive companies such as Mercedes, BMW, VW, and Hyundai have all announced integrating LLM into their in-car voice-controlled systems.
- **Impact**
    - Enhance accuracy in response and context understanding
    - Multilingual and dialect support
    - Safe-driving assistance: mitigate drivers distraction and fatigued driving through voices
    - Personalized pit stops:
        - Use voice commands to control in-vehicle systems, including climate control, windows, and music.
        - Enable personalized navigation with tailored route and music suggestions by learning user behavior and preferences from voice commands and past interactions.
        - Predict potential maintenance issues before becoming major problems by analyzing sensor data and vehicle diagnostics.
- **Benefits bringing to Porsche**
    - Enhance customer driving experience: voice-based command system assist both personalized user experience and safety
    - Brand loyalty: customer satisfaction retains more even attracts our customer base

(Create by DALL-E)

Our Solution Approach

- **Base LLM**: LLaMA 3 or other open-source llms from hugging face

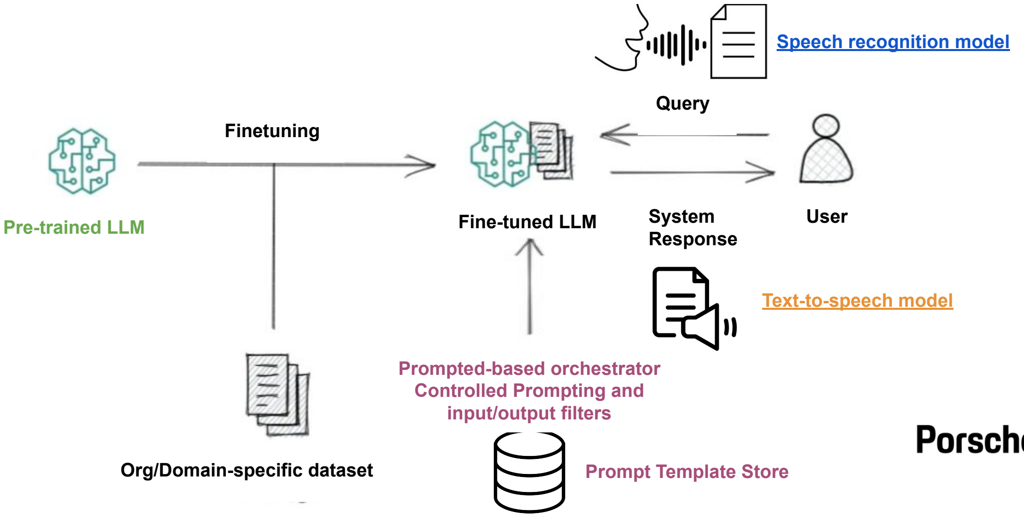
- **Speech-to-text model**: Whisper

OpenAI Whisper

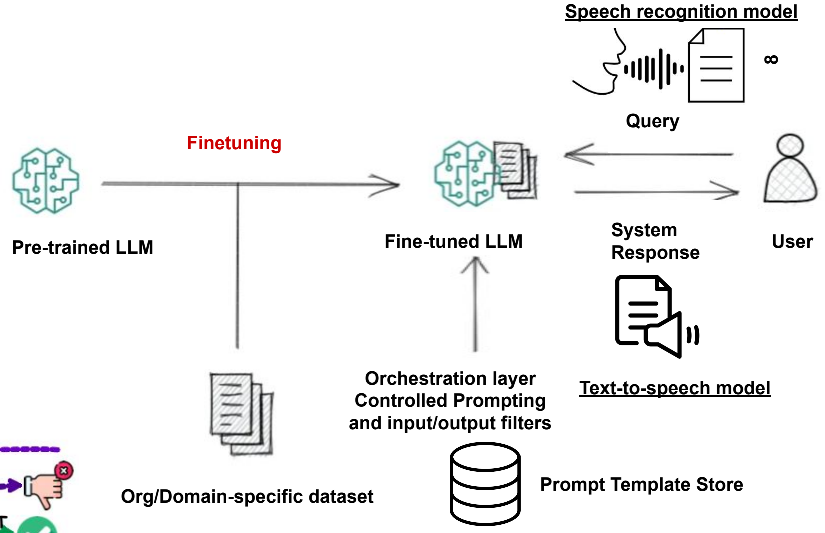
- **Text-to-speech model**: MaryTTS

MARY

- **Orchestrator**:
  tackle unsafe content
  deal with multi-turn scenarios



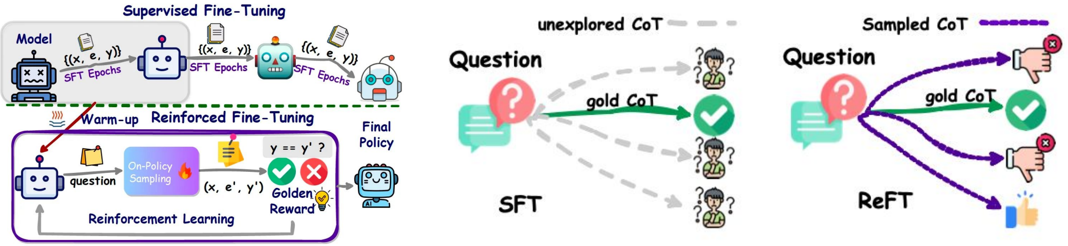
Our Solution Approach (cont'd)



- Some SOTA **Fine-tuning** approaches

1. Reinforced

- A new training alg fine-tuning for reasoning solving



- First, warm-up stage using supervised fine-tuning (SFT) to acquire a certain level of accuracy. Then, on-policy reinforcement learning (e.g. PPO) comes into play to enhance its ability by sampling various CoT reasoning paths. (Performance further boosted via majority vote and reward model reranking)

- Unlike reinforcement learning from human feedback (RLHF) and direct performance optimization (DPO), ReFT does not need human-labeled data.

- Reward hacking issue (Reward shaping)

2. Representation fine-tuning (ReFT)

3. Retrieval Augmented Fine Tuning (RAFT)

4. QLoRA

Our Solution Approach (cont'd)

- Some SOTA **Fine-tuning** approaches

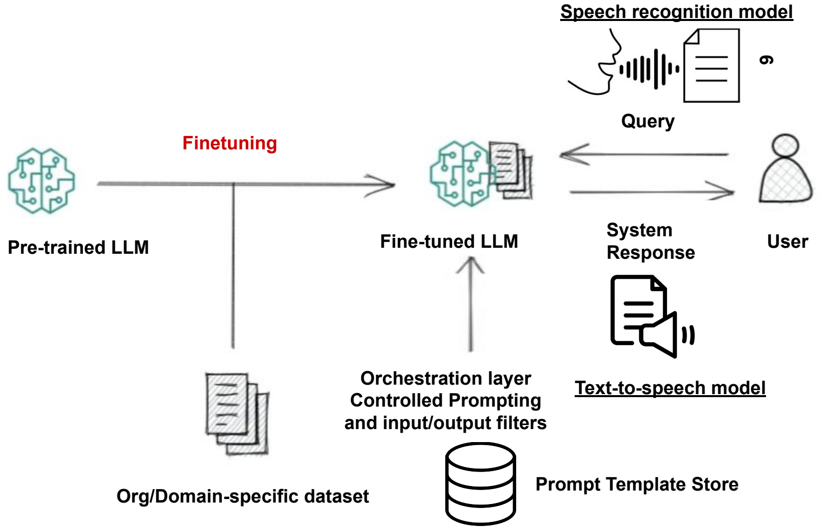
1. Reinforced fine-tuning (ReFT)

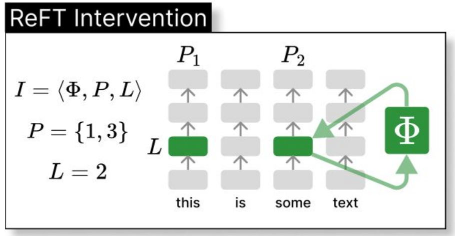
2. Representation fine-tuning (ReFT):

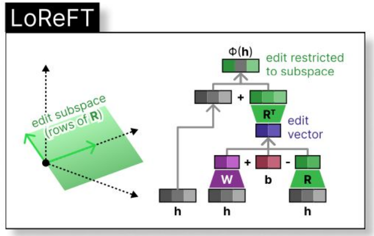
- Current parameter-efficient finetuning (PEFT) methods seek to modify the weights in large neural models rather than representations.

- However, representations encode rich semantic information.

- ReFT operates on a frozen base model and learn task-specific interventions on hidden representations.
  → intervention-based model interpretability



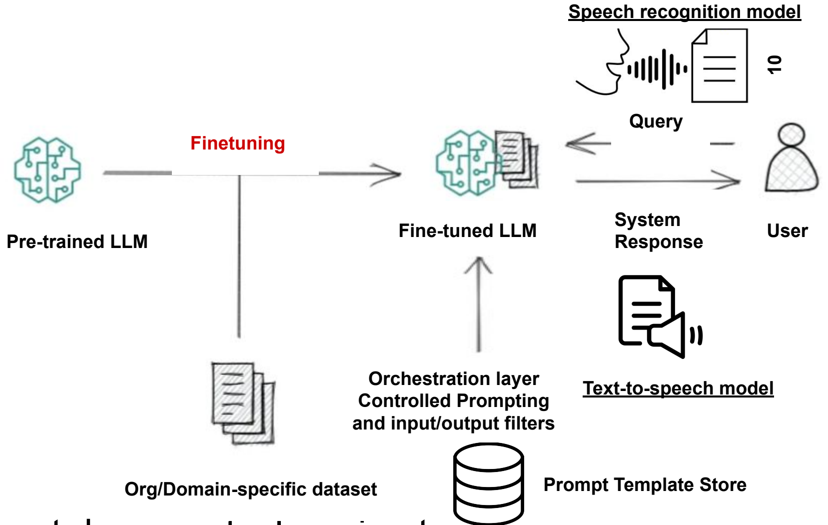




3. Retrieval Augmented Fine Tuning (RAFT)

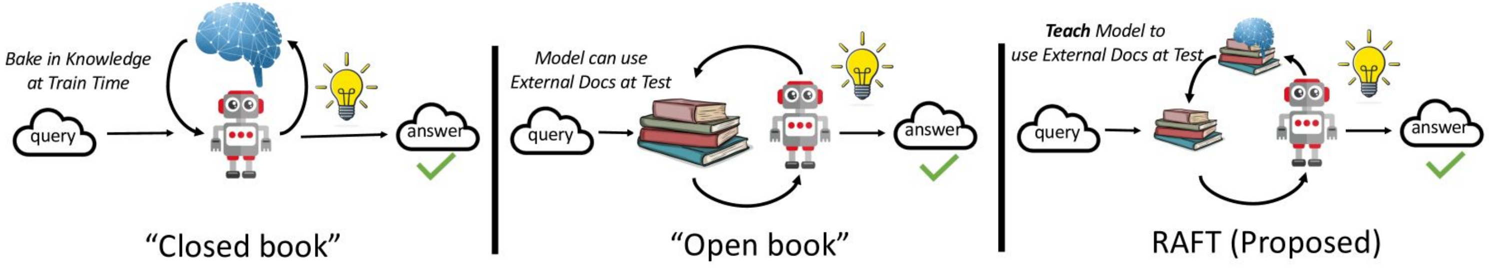
4. QLoRA

Our Solution Approach (cont’d)



- Some SOTA **Fine-tuning** approaches

1. **Reinforced fine-tuning (ReFT)**
2. **Representation fine-tuning (ReFT)**
3. **Retrieval Augmented Fine Tuning (RAFT):**
   - Combination of the strength of retrieval-augmented generation (RAG) and fine-tuning
   - RAFT is trained on questions paired with relevant and irrelevant documents, learning to answer using chain-of-thought reasoning that filters relevant content from distractors.



4. **QLoRA**

Our Solution Approach (cont'd)

- Some SOTA **Fine-tuning** approaches

  1. Reinforced fine-tuning (ReFT)

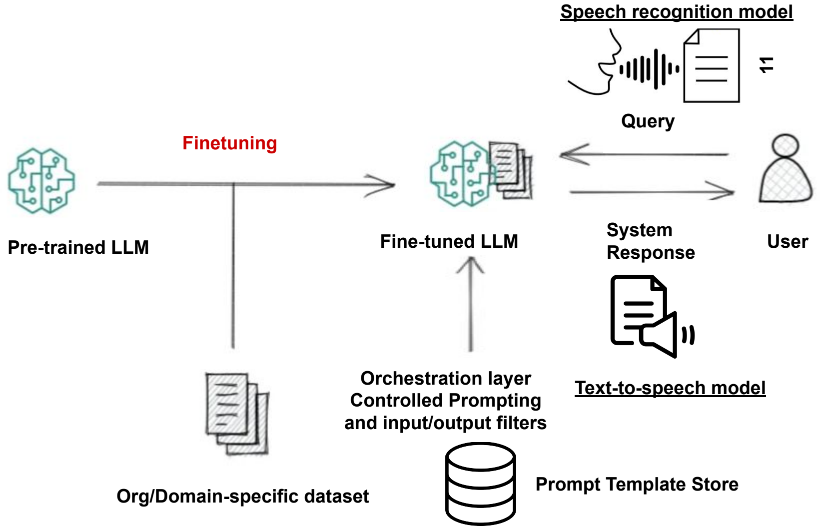
  2. Representation fine-tuning (ReFT)

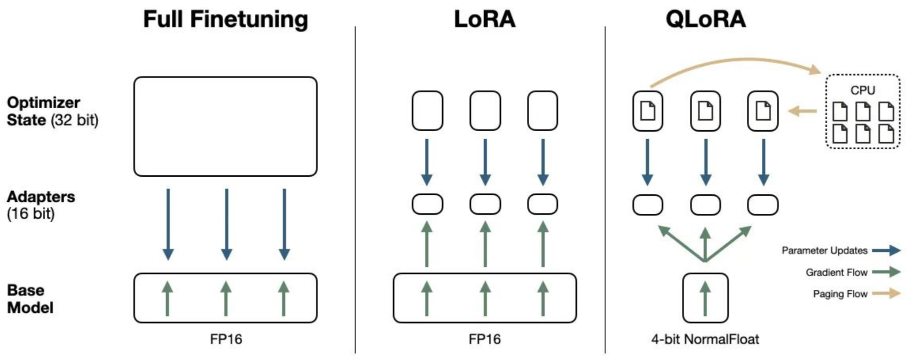
  3. Retrieval Augmented Fine Tuning (RAFT)

  4. QLoRA

  - PEFT approach with four main ingredients:

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





Porsche Digital Campus

Implementation details and Potential Challenges

- Implementation resources:
  Data recorded in an environment with vehicle noise in English, French, and German. But it’s not open data.
  Open source LLM library from Hugging face.

- Potential challenges:
  1. Computational resource limitations
  2. Scarcity of domain data

Porsche Digital Campus

Idea 2: MatchMyPorsche

Problem Understanding

- Topic: Vehicle Recommendation System
    - Build a recommendation engine that suggests particular car models, based on the customer's preferences and needs.
    - 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.
    - The same tool can provide a detailed comparison of specifications, features, pros, and cons as an aid in decision-making.
- Reasoning:
    - Streamlined decision-making for customers: provides an intuitive experience where customers feel guided and valued, reducing overwhelm from too many options.
- Impact
    - Insightful customer data: gathers valuable data on preferences, enabling Porsche to better understand market trends, inform future model design, and fine-tune marketing strategies.
    - Efficient lead qualification: helps identify high-potential leads faster, enabling sales teams to focus on customers who are more likely to buy, thus improving sales efficiency.
- Benefits bringing to Porsche
    - Enhance customer buying experience, address diverse customer needs
    - Brand loyalty: by delivering a tailored and user-centric experience, Porsche can cultivate stronger connections with both existing customers and new prospects.

(Create by DALL-E)

Our Solution Approach

- **Text embedding model**: sentence-transformer or other open-source model from MTEB benchmark

Type of Embedding	Description	Use-Case	Example of a Tool
Word Embeddings	Dense vector representations of words that capture semantic relationships.	enhances search engines by improving keyword relevance.	Word2Vec GloVe ELMo
Sentence Embeddings	Vectors that represent entire sentences, capturing their meaning in context.	Employed in document similarity detection.	Sentence Transformers
Image Embeddings	Vector representations of images that capture visual features for comparison.	Used in image retrieval systems to find similar images based on visual content.	TensorFlow Image Embedding API
Audio Embeddings	Representations of audio signals that capture features for sound classification.	Used in speech recognition systems to transcribe audio.	OpenAI's Whisper
Contextual Embeddings	Dynamic embeddings that consider context, producing different vectors for the same word in different sentences.	Used in language translation to understand context.	BERT (Bidirectional Encoder Representations from Transformers)
Graph Embeddings	Representations of nodes or entire graphs that capture relationships and properties in a lower-dimensional space.	Enhances fraud detection by analyzing transaction patterns.	Node2Vec
Multimodal Embeddings	Embeddings that integrate information from multiple modalities (e.g., text, images, audio) to provide a comprehensive representation.	Enhances interactive AI systems by integrating various inputs.	CLIP (Contrastive Language-Image Pretraining)

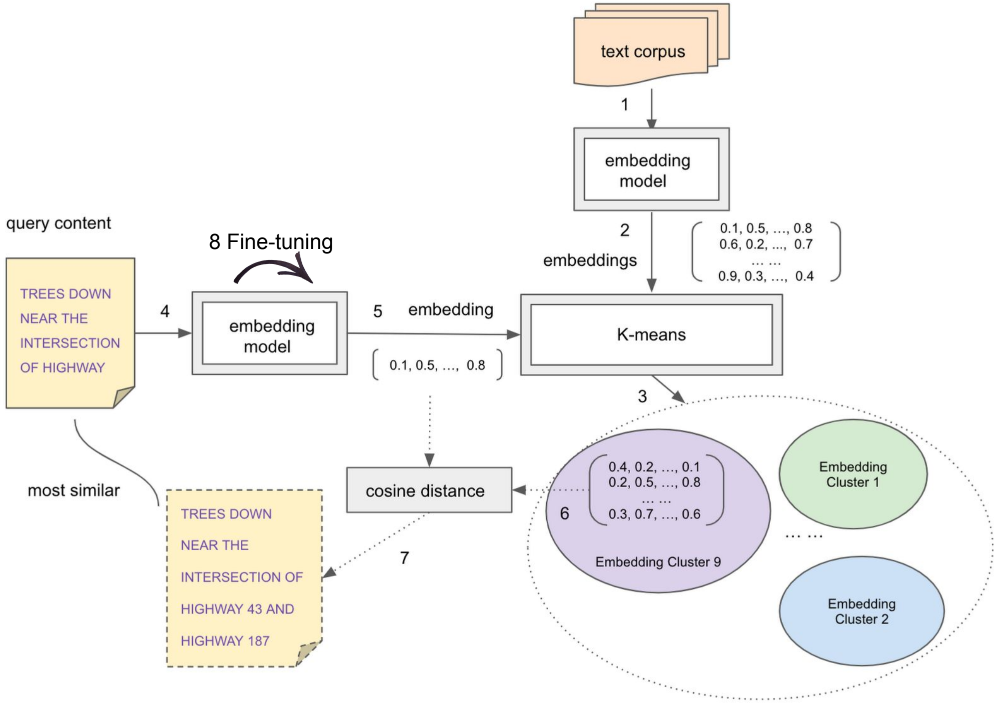
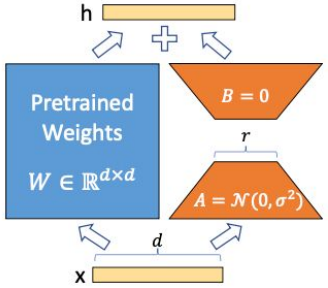


Image source: https://cloud.google.com/blog/products/data-analytics/introducing-bigquery-text-embeddings

Our Solution Approach (con’t)

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

- **Fine-tuning**
  - Adapter +



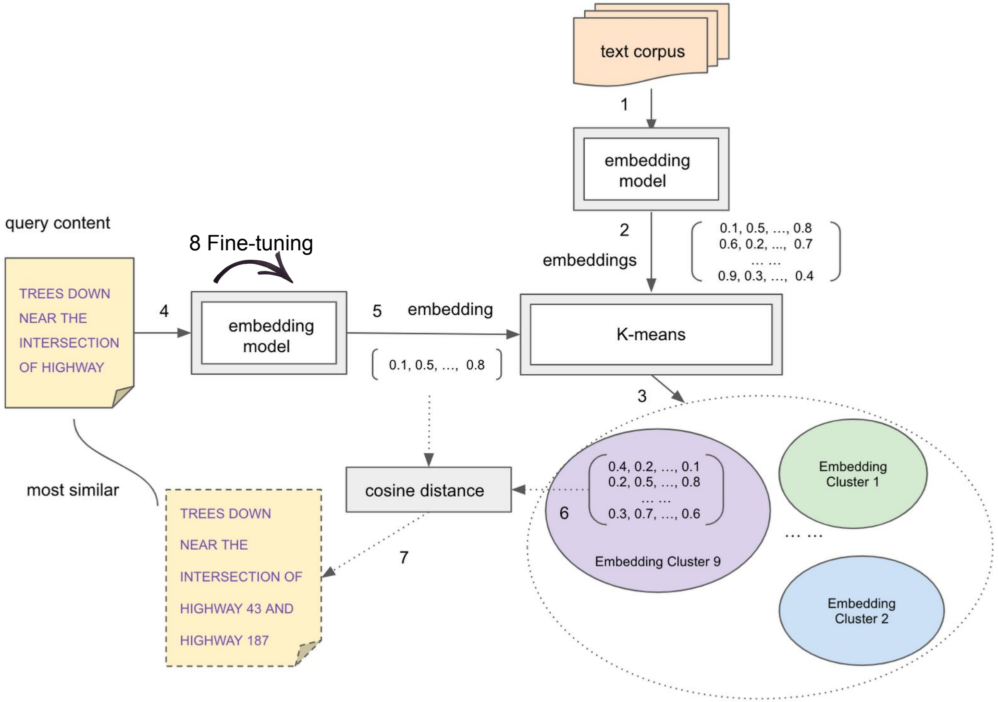
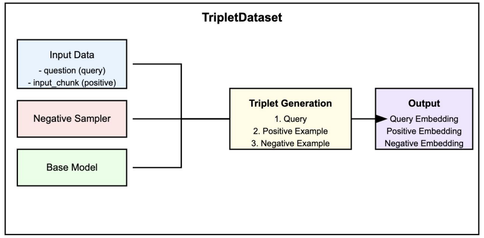
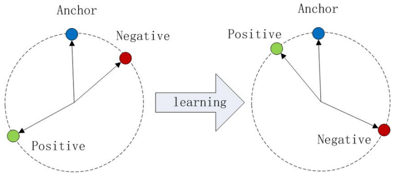


Image source: https://cloud.google.com/blog/products/data-analytics/introducing-bigquery-text-embeddings

  - Use LLM to create → this technique has wide
  - Some techniques from paper findings like triplet





Implementation details and Potential Challenges

- Implementation details:
  - Data: web crawling all porsche vehicle model feature descriptions. (But if you can provide the dataset, it would be nice)
  - Embedding models: open source LLM library from Hugging face
  - 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

- Potential challenges:
  Computational resource limitations

Porsche Digital Campus

Idea 3:
Porsche Pitstop

Problem Understanding

- Topic: Customer support automation
  - Build a question-and-answer chatbot assisting customer support team to resolve some customer issues and prioritize the urgency before reaching out to the customer support staff by fine-tuning text embedding models with RAG.
- Reasoning:
  - Streamlined customer service operations: reduces the burden on live agents, allowing them to focus on complex cases and improving efficiency in handling inquiries.
- Impact
  - Reduced response times: with instant, automated replies, customers experience reduced wait times, leading to higher satisfaction.
  - Valuable customer insights: by analyzing interactions, Porsche gains insights into customer needs and preferences, informing product development and marketing strategies.
  - Cost savings: automating routine queries can significantly reduce operational costs, making customer service more scalable without compromising quality.
- Benefits bringing to Porsche
  - Enhance customer experience: with 24/7 support and multilingual support, ensuring immediate assistance and convenience to global customers
  - Scalability: Capable of handling high volumes of inquiries, especially during peak times, without additional staffing
  - Improved Consistency: Ensures all customers receive consistent, on-brand information and responses, enhancing brand consistency across touchpoints

(Create by DALL-E)

Our Solution Approach

- **Text embedding model**: sentence-transformer or other open-source model from MTEB benchmark

Type of Embedding	Description	Use-Case	Example of a Tool
Word Embeddings	Dense vector representations of words that capture semantic relationships.	enhances search engines by improving keyword relevance.	Word2Vec GloVe ELMo
Sentence Embeddings	Vectors that represent entire sentences, capturing their meaning in context.	Employed in document similarity detection.	Sentence Transformers
Image Embeddings	Vector representations of images that capture visual features for comparison.	Used in image retrieval systems to find similar images based on visual content.	TensorFlow Image Embedding API
Audio Embeddings	Representations of audio signals that capture features for sound classification.	Used in speech recognition systems to transcribe audio.	OpenAI's Whisper
Contextual Embeddings	Dynamic embeddings that consider context, producing different vectors for the same word in different sentences.	Used in language translation to understand context.	BERT (Bidirectional Encoder Representations from Transformers)
Graph Embeddings	Representations of nodes or entire graphs that capture relationships and properties in a lower-dimensional space.	Enhances fraud detection by analyzing transaction patterns.	Node2Vec
Multimodal Embeddings	Embeddings that integrate information from multiple modalities (e.g., text, images, audio) to provide a comprehensive representation.	Enhances interactive AI systems by integrating various inputs.	CLIP (Contrastive Language-Image Pretraining)

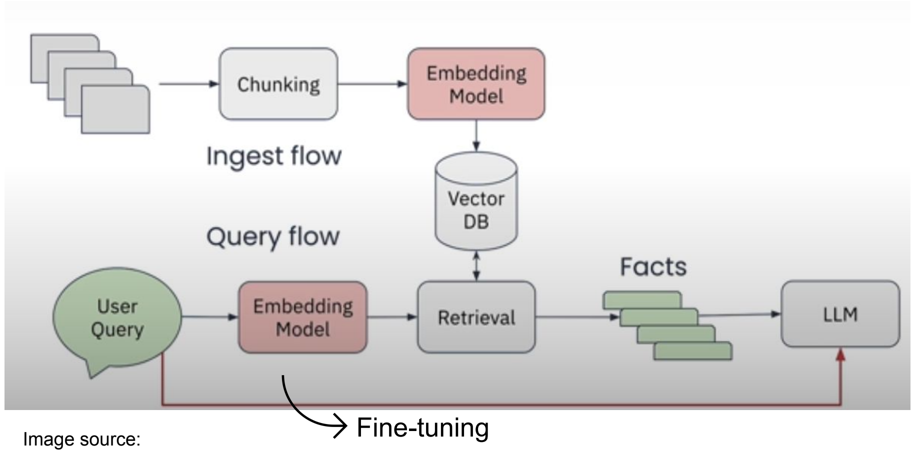
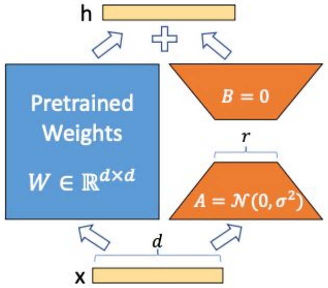


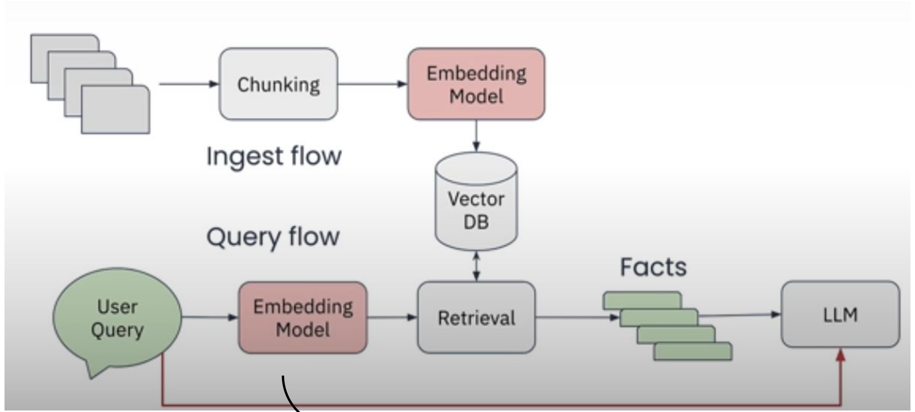
Image source:
https://www.deeplearning.ai/short-courses/embedding-models-from-architecture-to-implementation/

Our Solution Approach (con't)

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

- **Fine-tuning approaches**:
  - Adapter + PEFT (LoRA/QLoRA)

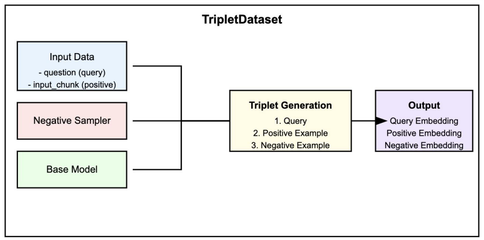


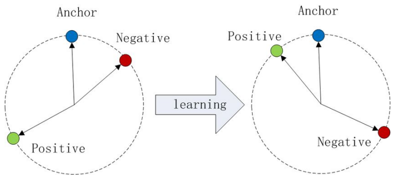


  - Use LLM to create → this technique has widely
  - Some techniques from the paper findings like triplet loss and random negative sampling

ChromaDB → Fine-tuning embedding adapters

Image source: https://www.deeplearning.ai/short-courses/embedding-models-from-architecture-to-implementation/





Implementation details and Potential Challenges

- Implementation details:
  - Data: unsure yet
  - Embedding models: open source LLM library from Hugging face
  - 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

- Potential challenges:
  - Data collection
  - Computational resource limitations

Reference

[1] Trung Quoc Luong, Xinbo Zhang, Zhanming Jie, Peng Sun, Xiaoran Jin and Hang Li. REFT: Reasoning with REinforced Fine-Tuning. Computation and Language (cs.CL), 27th June 2024. https://doi.org/10.48550/arXiv.2401.08967

[2] RAFT: Adapting Language Model to Domain Specific RAG. (n.d.). https://arxiv.org/html/2403.10131v1

[3] 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

[4] 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

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

[6] 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

[7] Parthasarathy, V. B., Zafar, A., Khan, A., & Shahid, A. (2024, August 23). The Ultimate Guide to Fine-Tuning LLMs from Basics to Breakthroughs: An Exhaustive Review of Technologies, Research, Best Practices, Applied Research Challenges and Opportunities. arXiv.org. https://arxiv.org/abs/2408.13296

[8] 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

[9] 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

[10] Wu, Z., Arora, A., Wang, Z., Geiger, A., Jurafsky, D., Manning, C. D., & Potts, C. (2024, April 4). REFT: Representation Finetuning for Language Models. arXiv.org. https://arxiv.org/abs/2404.03592

Porsche Digital Campus

THANK YOU.
FEEDBACK?