Using Large Language Models to improve your Search Engine

About Me

Ravi Yadav
Linkedin
ravi@sukuya.com

Mustafa Zengin
Linkedin
mustafa.zengin@gmail.com

Nidhin Pattaniyil
Linkedin
npatta01@gmail.com

Walmart

ML Engineer for Ecommerce site

Labs

Introduction / Lab1

- Sample Record looks like
- Sample search results : decently, opportunity for improvement (stuffing, vocabulary misgap_

Overview of LLMs / Lab2
- summarization with Gpt4 / Claude / open source model
- just giving extract the user intent of this query " red nike men shoes".. Not usable

Prompting / Lab 3
- example where zero shot is not performing well
- st intent
- query rewrite
- relevance
- sample , where we are improving due to each

Finetuning :
Git clone axolotl
Lora output

Results

Resources

Github:

https://github.com/npatta01/llm_search

Userid: npatta01
Repo name: llm_search

Target Audience

- Beginner in LLM ✅
- Beginner in Search ✅
- Looking for simple applied use-case

- If u are already using langchain / RAG/ Finetuned LLM😅

Agenda

- Introduction
- Overview of Large Language Models
- Example of Prompting
- Fine Tuning LLM
- Production use case
- Questions

Learning Objectives

- Understand some of the problems in E-commerce search
- High Level intro to LLMs
- How to prompt LLMs using Langchain
- Results of pre-trained vs fine-tuned LLMs

Introduction

E-Commerce Search Problems

- Understanding Query Intent
- Understanding Item Intent
- Understanding if Item is relevant to a query
- Helping re-formulate ambiguous queries

Query Understanding

Red Nike Men shoes 10 inches 

Color: Red
Brand: Nike
Gender: Male
Product Type: shoes
Size: 10 inches

Item Understanding

Color: Red
Gender: Female
Product Type: shoes

Color: Red
Brand: Nike
Gender: Male
Product Type: shoes

Color: Red
Gender: Male
Product Type: Socks
Size: 10 inches

Color: Red
Gender: Male
Product Type: shoes
Size: 10 inccches

Query Item Relevance

Red Nike Men shoes 10 inches

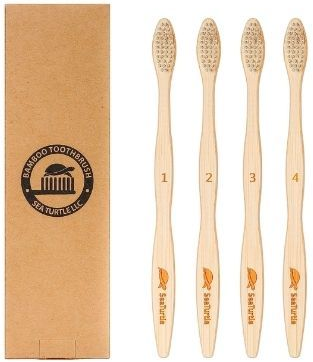
Color: Red
Brand: Nike
Gender: Male
Product Type: shoes
Size: 10 inches

christmas gift ideas

Christmas Rose Flower Gifts for Women,Mom,Christmas Birthday Gifts for Women Grandma Wife Mother Her Friends Presents Xmas Mom Gifts,Galaxy Butterfly Purple Light Up Rose Flowers Gifts in A Glass Dome

11" Nonstick Frying Pan with Lid - 11 Inch Nonstick Skillets with USA Blue Gradient Granite Derived Coating, Heat-resisted Silicon Handle, PFOA &PFOS Free, Induction Compatible, Ideal Christmas Gift

zero waste oral care



SeaTurtle Plant-Based Bristles, Bamboo Toothbrushes, Soft Natural Toothbrush for Adults (4 Pack)

SuperBee Dentos 100 Toothpaste Tablets, Fluoride Free & Eco Friendly, Sensitive Bites for Kids and Adults, Chewable Spearmint

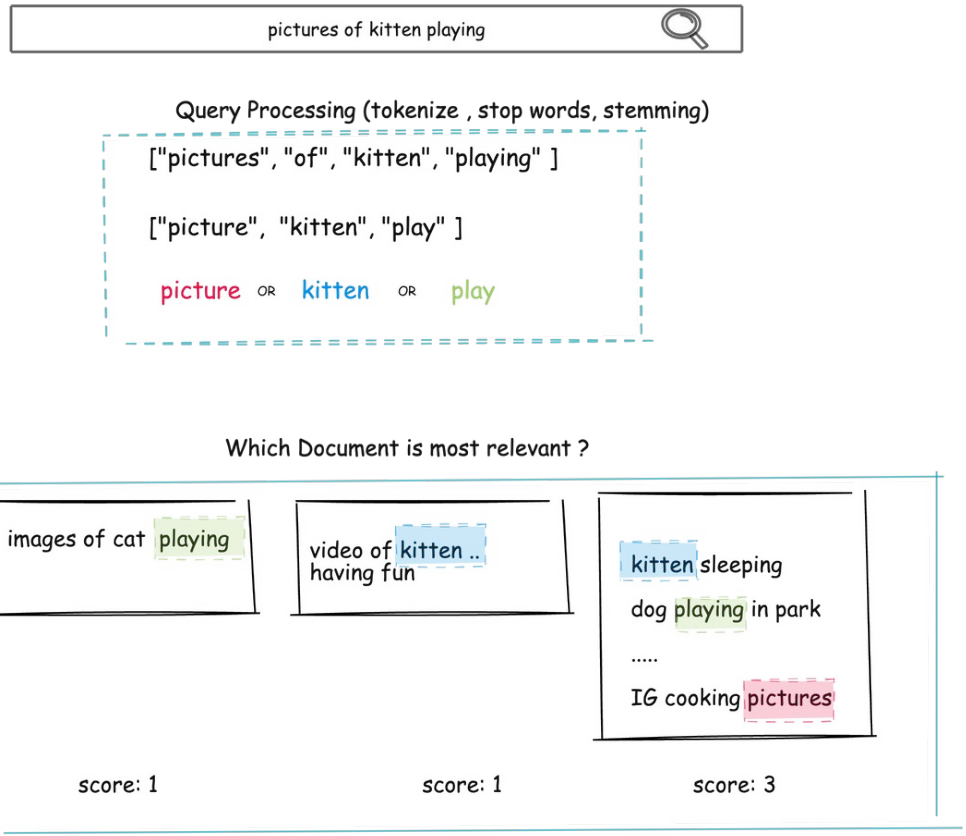
JentleCo Biodegradable Silk Floss Refill - Earth Friendly, Unflavored, Plastic-Free, Zero-Waste, for Refillable Floss Dispenser, 66 yd (30 m / 33 yd x 2 Spools)

E-Commerce Search

- Understanding Query Intent
- Understanding Item Intent
- Understanding if Item is relevant to a query
- Helping re-formulate ambiguous queries / Title Stuffing

Simple Retrieval System

- Queries and documents are represented as bag of words
- Query terms are connected with boolean operators
- Disadvantages:
  - Token exact match
  - Terms have same weights
  - Doesn't consider document length



Dataset

[Shopping Queries Dataset: A Large-Scale ESCI Benchmark for Improving Product Search]
[Github]

Lab 1

Lab 1 Goals

- Understand e-commerce dataset
- Understand e-commerce type of queries
- Understand effectiveness of simple token based retrieval

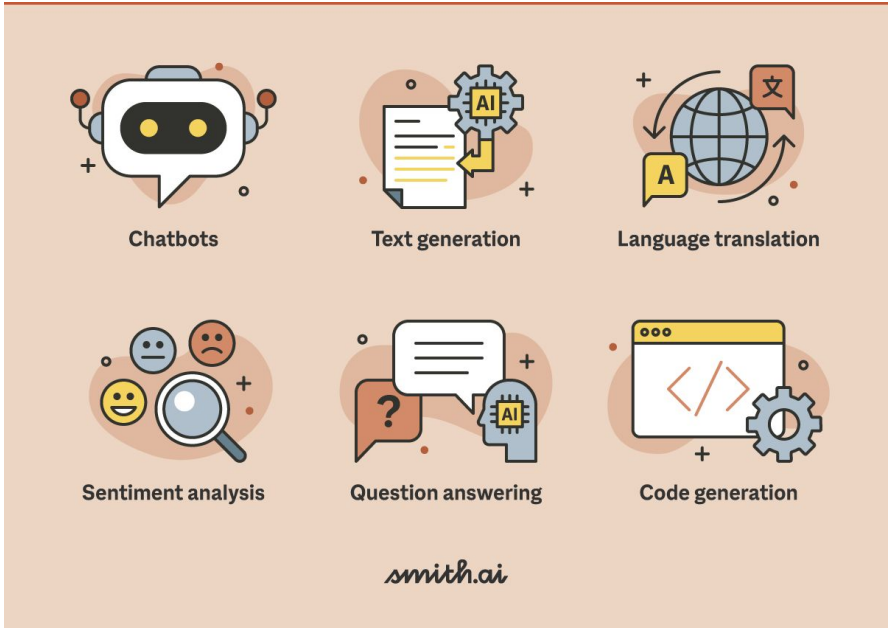
Takeaways

- Simple text based retrieval can work
- Can suffer from keyword stuffing
- Some irrelevant items still showing up
- Hard to work with ambiguous intents

Overview of Large Language Models

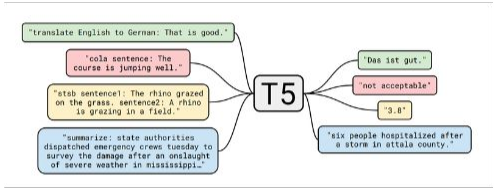
What is an LLM

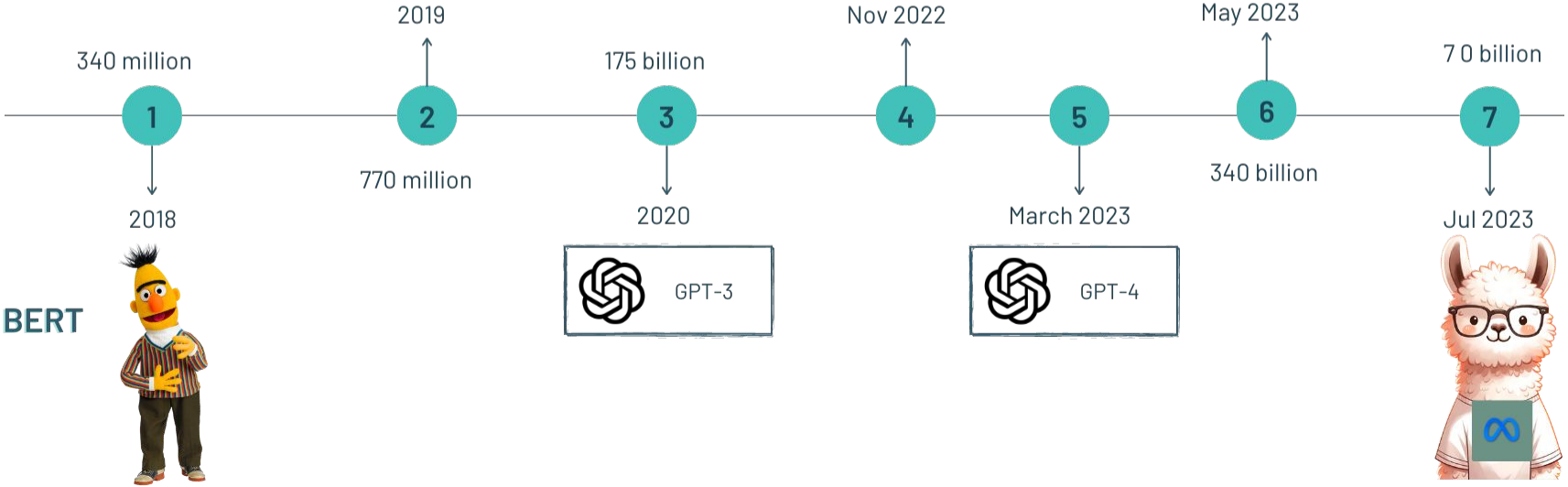
- A large neural network with 1 billion + parameters
- Trained on very large corpus
- Single model can perform multiple NLP Tasks



https://smith.ai/blog/llm-ai

Recent LLM moments





Steps in Training

Pretraining

Train on 1T tokens

Training task:
predict next token on raw text

My Llama ate ____.

Supervised Finetuning

Train on 50k
instruction pairs

Next token prediction

```
{
    "instruction": "Write a poem about 🦙",
    "input": "..",
    "output": "Llama are .."
}
```

Reinforcement Learning

>50k examples

Reinforcement learning. Align with human feedback

LLM Output

Dataset Size

Dataset	Sampling prop.	Epochs	Disk size
CommonCrawl	67.0%	1.10	3.3 TB
C4	15.0%	1.06	783 GB
Github	4.5%	0.64	328 GB
Wikipedia	4.5%	2.45	83 GB
Books	4.5%	2.23	85 GB
ArXiv	2.5%	1.06	92 GB
StackExchange	2.0%	1.03	78 GB

Table 1: **Pre-training data.** Data mixtures used for pre-training, for each subset we list the sampling proportion, number of epochs performed on the subset when training on 1.4T tokens, and disk size. The pre-training runs on 1T tokens have the same sampling proportion.

Architecture

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

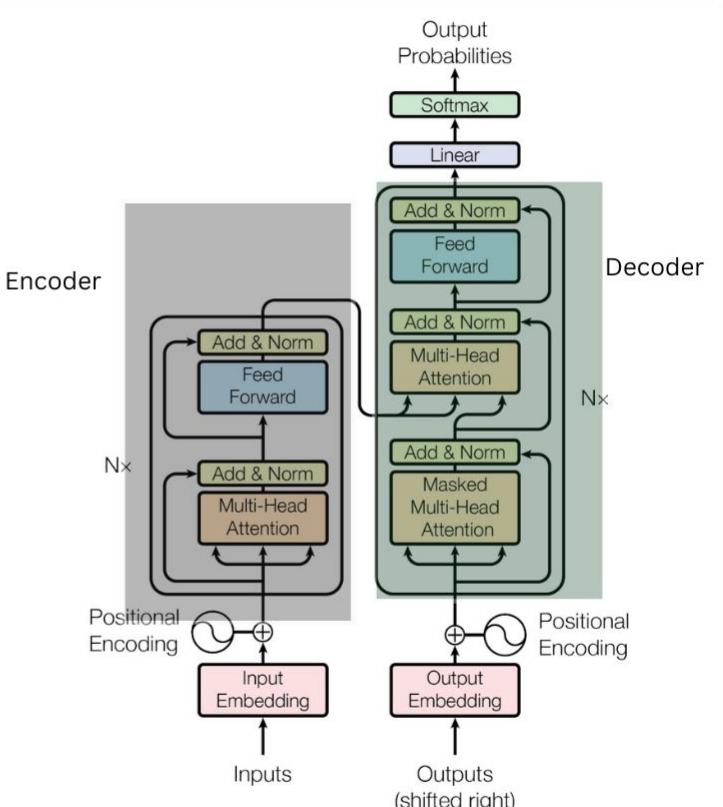
Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com



[Attention is all you need](#) paper

Evaluating LLM

Model	Arena Elo rating	MT-bench (score)	MMLU	License	Company
GPT-4	1159	8.99	86.4	Proprietary	OpenAI
Claude-2	1125	8.06	78.5	Proprietary	Anthropic
GPT-3.5-turbo	1103	7.91	70	Proprietary	OpenAI
Llama-2-70b-chat	1065	6.86	63	Llama 2 Community	Meta
zephyr-7b	1042	7.34	61.4	MIT	HuggingFace
MPT-30-chat	1031	6.39	50.4	CC-BY-NC-SA-4.0	MosaicML
Llama-2-13b-chat	1021	6.64	53.6	Llama 2 Community	Meta
Llama-2-7b-chat	1001	6.27	45.8	Llama 2 Community	Meta
PaLM-Chat-Bison-001	991	6.4	-	Proprietary	Google

- Chatbot Arena - a crowdsourced, randomized battle platform. 100K+ user votes.
- MT-Bench - a set of challenging multi-turn questions.
- MMLU model's multitask accuracy on 57 tasks.

Chatbot Arena Leaderboard

Example of Prompting

Prompt

Instructions passed to a language model to achieve a desired task

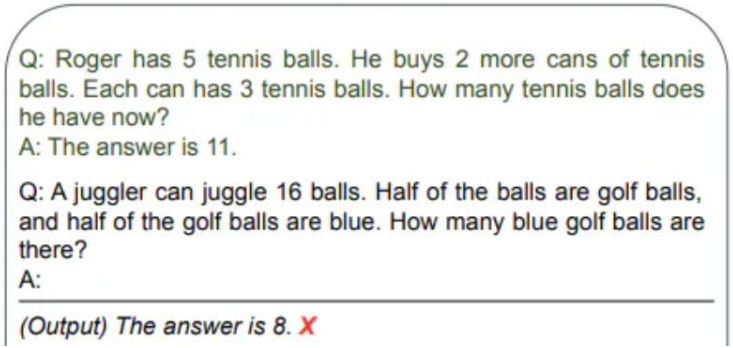
Role	You are a medical expert
Instructions	Classify the below medical news as healthy, concerning.
Input Data	Input: The patient's A1c is 15%
Output Data	Output:

Few Shot Learning

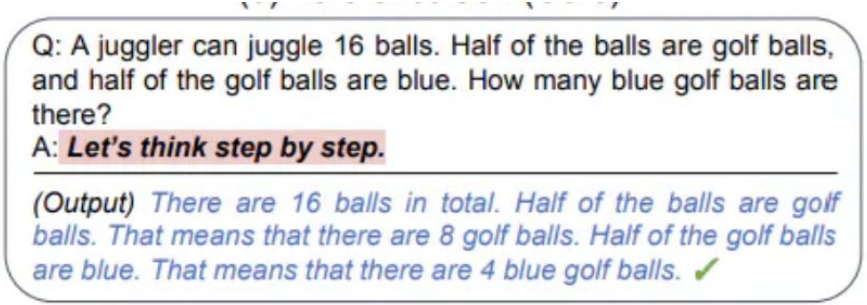
Role	You are a medical expert
Instructions	Classify the below medical news as healthy, concerning.
Input Data Output Data	Input: The patient's A1c is 15% Output: Concerning
Input Data Output Data	Input: His T3 test results are in . It is 112 ng/dL Output: Healthy
Input Data Output Data	Input: Cholesterol, HDL is 35 mg/dl Output:

Chain of Thought

Few Shot



Chain of Thought (zero)



Large Language Models are Zero-Shot Reasoners

Prompt Parameters

temperature number — 1

Amount of randomness injected into the response.

Defaults to 1. Ranges from 0 to 1. Use temp closer to 0 for analytical / multiple choice, and closer to 1 for creative and generative tasks.

top_p number — 0.7

Use nucleus sampling.

In nucleus sampling, we compute the cumulative distribution over all the options for each subsequent token in decreasing probability order and cut it off once it reaches a particular probability specified by `top_p`. You should either alter `temperature` or `top_p`, but not both.

top_k integer — 5

Only sample from the top K options for each subsequent token.

Used to remove "long tail" low probability responses. Learn more technical details here.

Claude parameters

Lab

Lab Goals

- Results of open source models vs private model
- Prompt for extracting intent of query and item
- Prompt for Query / Item Relevance
- Prompt for Ambiguous Query

Takeaways

- LangChain makes it easy to use any LLM
- Few Shot helps the model for some use cases
- GPt4 seems to be the best
- Base models are worse than chat based model
- Open source models work, but don’t work with rendering output

Fine Tuning LLM

Why finetune an open source model

- Cheaper and faster than training from ground up
- Better performance on specific tasks and domains
- Transfer the knowledge and on pretraining to new tasks
- Better generalization
- No privacy concerns
- Own the ip

Difficulties usage in Fine tuning

- Data collection and quality!
- Overfitting and catastrophic forgetting
- Computational resources
- Lack of interpretability
- Examples / instruction tuning

Memory Requirements

- Consider a 7B parameter relatively “small” LLM
- 2 bytes per parameter for the weights
- 4 bytes per parameter for optimizer state (ADAM)
- 2 bytes per parameter for gradients
- 7B model will require (2 + 4 + 2) * 7e9 = 52 GB Memory
- We ignored batch size, sequence length and other factors

LoRA : Low-Rank Adaptation of LLMS

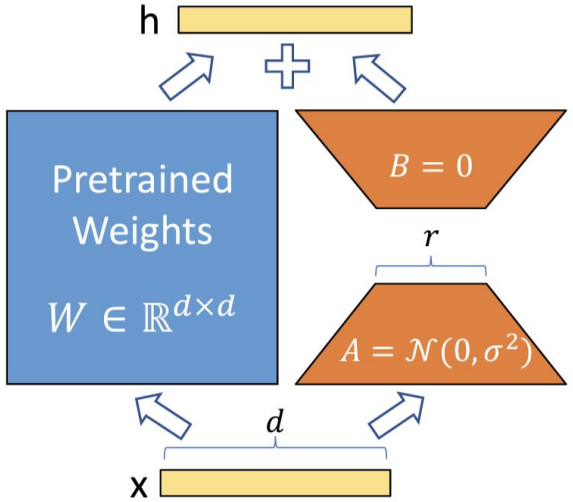
"Freezes pretrained model weights and injects trainable rank decomposition matrices into each layer of transformer architecture"

W is d X d matrix -> dXd params

W =(d X r).(r X d) -> dXrX2 params

LoRA: Low-Rank Adaptation of Large Language Models

Edward Hu* **Yelong Shen*** **Phillip Wallis** **Zeyuan Allen-Zhu**
Yuanzhi Li **Shean Wang** **Lu Wang** **Weizhu Chen**
Microsoft Corporation
{edwardhu, yeshe, phwallis, zeyuana,
yuanzhil, swang, luw, wzchen}@microsoft.com
yuanzhil@andrew.cmu.edu
(Version 2)



LoRA : Low-Rank Adaptation of LLMS

LoRA: Low-Rank Adaptation of Large Language Models

Edward Hu* **Yelong Shen*** **Phillip Wallis** **Zeyuan Allen-Zhu**
Yuanzhi Li **Shean Wang** **Lu Wang** **Weizhu Chen**
Microsoft Corporation
{edwardhu, yeshe, phwallis, zeyuana,
yuanzhil, swang, luw, wzchen}@microsoft.com
yuanzhil@andrew.cmu.edu
(Version 2)

Model & Method	# Trainable Parameters	E2E NLG Challenge BLEU	NIST	MET	ROUGE-L	CIDEr
GPT-2 M (FT)*	354.92M	68.2	8.62	46.2	71.0	2.47
GPT-2 M (Adapter^{L})*	0.37M	66.3	8.41	45.0	69.8	2.40
GPT-2 M (Adapter^{L})*	11.09M	68.9	8.71	46.1	71.3	2.47
GPT-2 M (Adapter^{H})	11.09M	$67.3_{\pm.6}$	$8.50_{\pm.07}$	$46.0_{\pm.2}$	$70.7_{\pm.2}$	$2.44_{\pm.01}$
GPT-2 M (FT^{Top2})*	25.19M	68.1	8.59	46.0	70.8	2.41
GPT-2 M (PreLayer)*	0.35M	69.7	8.81	46.1	71.4	2.49
GPT-2 M (LoRA)	0.35M	$\mathbf{70.4}_{\pm.1}$	$\mathbf{8.85}_{\pm.02}$	$\mathbf{46.8}_{\pm.2}$	$\mathbf{71.8}_{\pm.1}$	$\mathbf{2.53}_{\pm.02}$
GPT-2 L (FT)*	774.03M	68.5	8.78	46.0	69.9	2.45
GPT-2 L (Adapter^{L})	0.88M	$69.1_{\pm.1}$	$8.68_{\pm.03}$	$46.3_{\pm.0}$	$71.4_{\pm.2}$	$\mathbf{2.49}_{\pm.0}$
GPT-2 L (Adapter^{L})	23.00M	$68.9_{\pm.3}$	$8.70_{\pm.04}$	$46.1_{\pm.1}$	$71.3_{\pm.2}$	$2.45_{\pm.02}$
GPT-2 L (PreLayer)*	0.77M	70.3	8.85	46.2	71.7	2.47
GPT-2 L (LoRA)	0.77M	$\mathbf{70.4}_{\pm.1}$	$\mathbf{8.89}_{\pm.02}$	$\mathbf{46.8}_{\pm.2}$	$\mathbf{72.0}_{\pm.2}$	$2.47_{\pm.02}$

Table 3: GPT-2 medium (M) and large (L) with different adaptation methods on the E2E NLG Challenge. For all metrics, higher is better. LoRA outperforms several baselines with comparable or fewer trainable parameters. Confidence intervals are shown for experiments we ran. * indicates numbers published in prior works.

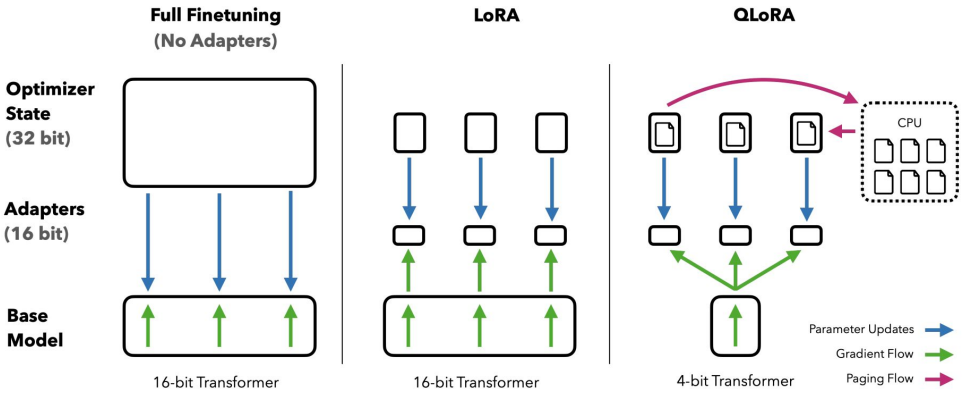
LoRA Rank

Rank r	val_loss	BLEU	NIST
1	1.23	68.72	8.7215
2	1.21	69.17	8.7413
4	1.18	**70.38**	**8.8439**
8	1.17	69.57	8.7457
16	**1.16**	69.61	8.7483
32	**1.16**	69.33	8.7736
64	**1.16**	69.24	8.7174
128	**1.16**	68.73	8.6718
256	**1.16**	68.92	8.6982
512	**1.16**	68.78	8.6857
1024	1.17	69.37	8.7495

- Effectiveness of higher rank plateaus
- For some datasets, lower rank can perform better

LoRA: Low-Rank Adaptation of Large Language Models

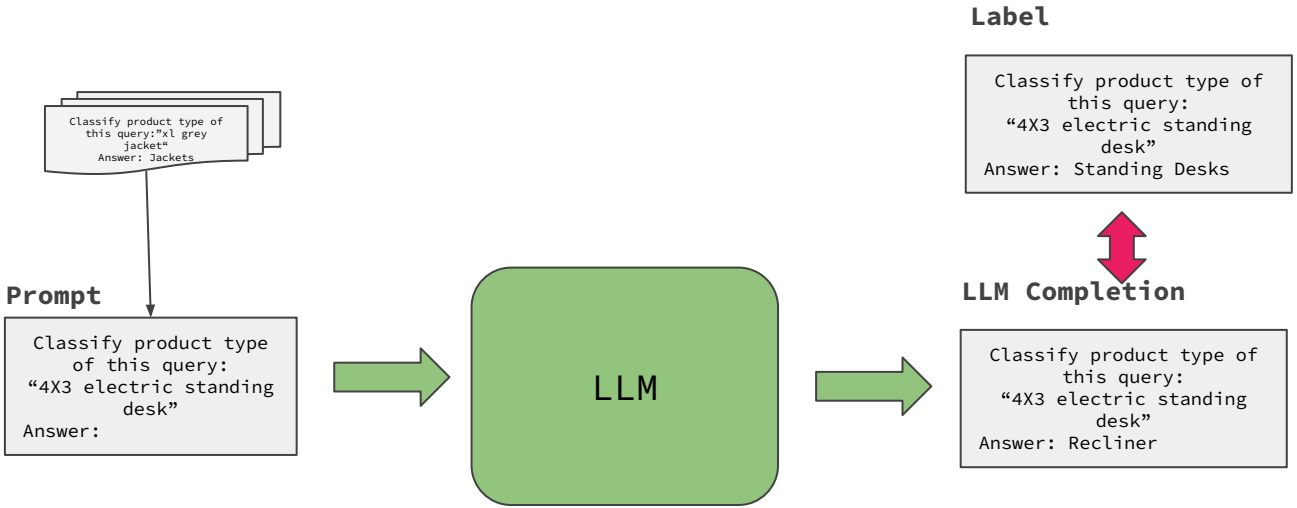
QLoRA: Quantized LoRa



- Uses 4bit NormalFloat precision
- Double quantization (Saves 0.37 bits per parameter)
- Paged optimizers to prevent gradient checkpointing memory spikes
- Use unified GPU-CPU memory management to reduce GPU memory
- Reduce GPU memory usage for slight drop in accuracy

QLORA: Efficient Finetuning of Quantized LLMs

Shape of the data



Lab

Takeaways

- Replicate / Axolotl makes it easier to finetune
- Finetuning with QLora is possible
- Get decent results

Production Use Cases

gifts for men

Query Rewrite Model

Monogrammed leather wallets

Beard grooming kits for men

Gift cards

Intent Understanding Model (query)

Product Type: Personalized Accesories

Product Type: Grooming Products

Product Type: Cards

elastic

Text / Vector Search

Gluing Together

Relevancy Model

#1 : ✅

#2 : ❓

#3 : ❌

Intent Understanding Model (item)

1

Amlion Custom Engraved Wallet,Personalized Photo RFID Wallets for Men,Husband,Dad,Son,Personalized Gifts

Product Type: Personalized Accessories
Gender: Men

2

Amazon Gift Card - Print - Birthday Cupcakes

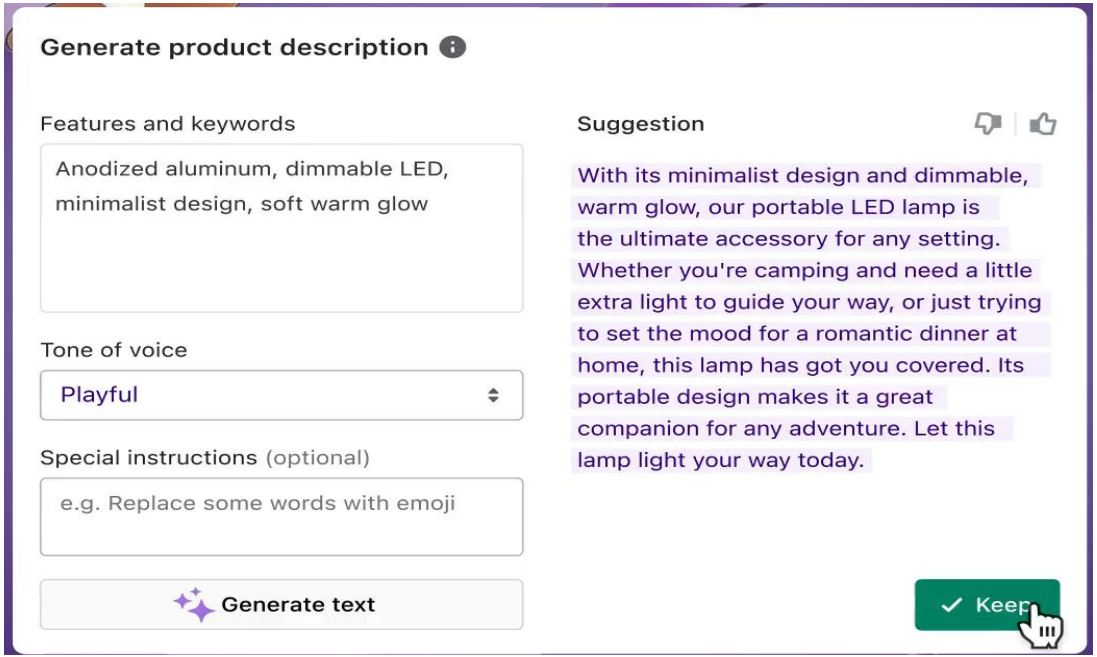
Product Type: Cards

3

Product Type: Personalized Accessories
Gender: Women

Resdink To My Gorgeous Wife, Engraved Wallet Cards for wife, Love Gifts for Wife, Anniversary Present Card for Wife Her, I Love You With All My Heart, Wife Gift from Husband, Birthday Valentines

AI Generated Product Description



shopify

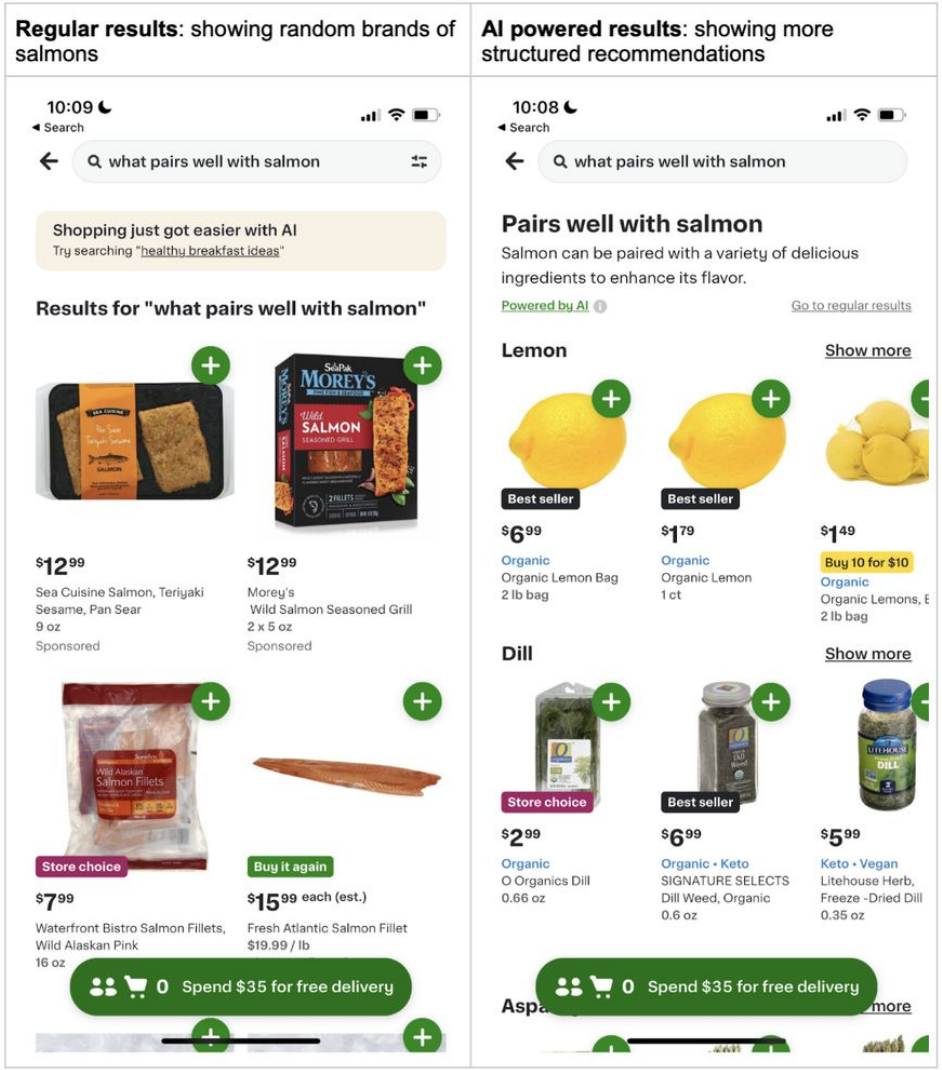
[Introducing AI-Generated Product Descriptions Powered by Shopify Magic]

[Kriti Kohli - Leveraging Generative AI for Enhanced E-commerce | PyData NYC 2023]

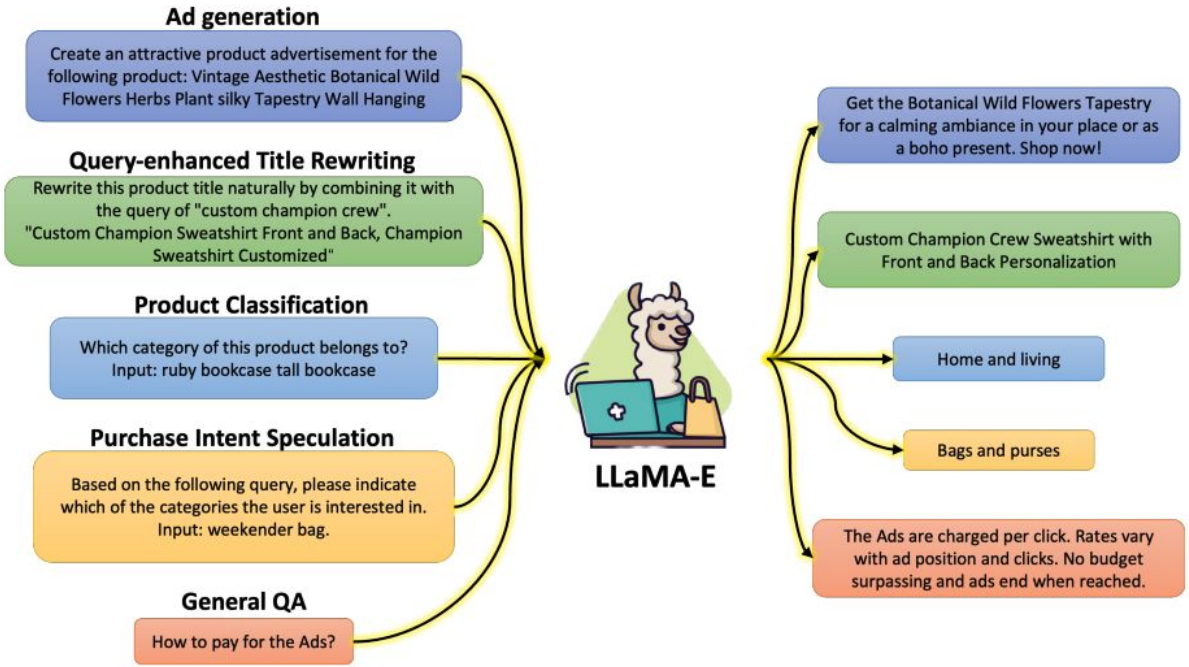
Understand Ambiguous Queries

[Building Generative AI Products Case Study: Instacart]

[Supercharging Search with LLMs (instacart search)]



LLama for e-commerce



[LLaMA-E: Empowering E-commerce Authoring with Multi-Aspect Instruction Following]

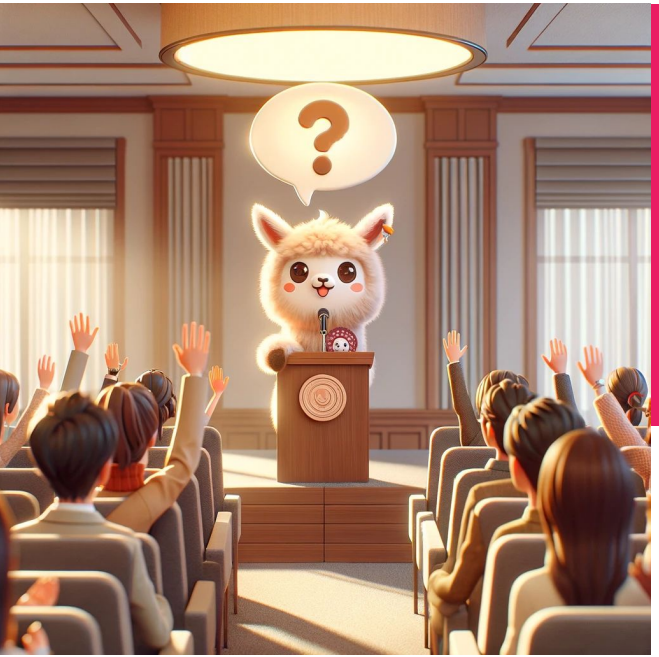
Etsy

LLama for e-commerce

Model		Purchase Intent
Instr.		Which category does the following query suggest the user is interested in?
Input		Himalayan salt lamp
Taxonomy		home and living
GPT-3.5		Home and Living
GPT-2		jewelry
BART		home and living
T5-base		The query suggests that the user is interested in "toys and games" category.
GPT-Neo		The query suggests that the user is interested in accessories, specifically a "Himalayan salt lamp".
LLaMA	7b	This query shows the products with Himalayan salt lamp as part of their title or description on Amazon.com.
	13b	https://www.etsy.com/search?q=himalayan+salt+lamp
	30b	Home & Living
LLaMA-E	7b	home and living
	13b	home and living
	30b	home and living

Qualitative Evaluation of Purchase Intent Speculation

LLaMA-E: Empowering E-commerce Authoring with Multi-Aspect Instruction Following



Questions

Resources

LLaMA-E: Empowering E-commerce Authoring with Multi-Aspect Instruction Following

Introducing AI-Generated Product Descriptions Powered by Shopify Magic

Kriti Kohli - Leveraging Generative AI for Enhanced E-commerce | PyData NYC 2023

Supercharging Search with LLMs (instacart search)

LLMs for Everything and Everyone! - Sebastian Raschka - Lightning AI