EEP 596: LLMs: From Transformers to GPT || Lecture 16

Dr. Karthik Mohan

Univ. of Washington, Seattle

March 8, 2024

Deep Learning and Transformers References

Deep Learning

Great reference for the theory and fundamentals of deep learning: Book by Goodfellow and Bengio et al Bengio et al
Deep Learning History

Embeddings

SBERT and its usefulness
SBert Details
Instacart Search Relevance
Instacart Auto-Complete

Attention

Illustration of attention mechanism

Generative AI References

Prompt Engineering

Prompt Design and Engineering: Introduction and Advanced Methods

Retrieval Augmented Generation (RAG)

Toolformer

RAG Toolformer explained

Misc GenAI references

Time-Aware Language Models as Temporal Knowledge Bases

Generative AI references

Stable Diffusion

The Original Stable Diffusion Paper
Reference: CLIP
Diffusion Explainer: Visual Explanation for Text-to-image Stable Diffusion
Diffusion Explainer Demo
The Illustrated Stable Diffusion
Unet

GenAI Evaluation and Annotation References

LLM Evaluations and Annotations

Evaluating LLMs
LLM Benchmarks article
Is Prompt Engineering Dead?
NeuroPrompts HumanEval - Evaluating Code HelloSwag

LLM Adverserial Attacks

Decoding Trust
TechTalks article

Previous Lecture

- Adverserial Attacks on LLMs
- Evaluation of LLMs

This Lecture

- Evaluating LLMs

Math Example GPT-3.5

You

what is 23457 times 82439

ChatGPT

The result of multiplying 23,457 by 82,439 is 1,934,797,523.

You

The calculation is correct

ChatGPT

Thank you for confirming! If you have any more questions or need further assistance, feel free to ask.

You

I mean redo it

ChatGPT

Apologies for the confusion. Let's redo the calculation:

23457 * 82439 = 1,932,439,323

Math Example GPT-4

You

what is 23457 times 82439

ChatGPT

The result of multiplying 23,457 by 82,439 is 1,933,771,623.

You

This is incorrect. Redo it.

ChatGPT

The calculation is confirmed to be correct; multiplying 23,457 by 82,439 results in 1,933,771,623.

Evaluating LLMs

Evaluation Dimensions

1. Math
2. Reasoning
3. Truthfulness
4. Coding
5. Accuracy across Multiple Tasks

Evaluation Datasets

1. **MATH:** Math problems with 7 difficulty levels
2. **GSM-8k:** Grade School Math
3. **MMLU:** Multitask Accuracy
4. **HumanEval:** Python Coding Tasks
5. **HellaSwag:** Reasoning Tasks
6. **TruthfulQA:** Truthfulness Tasks

Why LLM Benchmarking Datasets?

1. Standardized way to evaluate LLMs across a wide variety of tasks - Math, Reasoning, Truthfulness, etc
2. If you develop a model and evaluate on your own custom data set - There maybe a tendency to over-fit to that data set and over-promise performance.
3. Lot of thought goes into developing these benchmarking data sets for different tasks

LLM Evaluation Leaderboards

Many Leaderboards

There are many leaderboards along many evaluating dimensions and different benchmarking data sets!

LLM Evaluation Leaderboards

Many Leaderboards

There are many leaderboards along many evaluating dimensions and different benchmarking data sets!

A few of them...

1. Open LLM LeaderBoard (just for Open LLMs)
2. Open and Proprietary LLM LeaderBoard
3. LLM Safety LeaderBoard (across Open and Proprietary LLMs on Safety tasks)
4. Performance of LLMs (Latency, Memory, Throughput)
5. Chatbot Arena (Based on Battling LLMs against each other)

Caveat: Evaluating LLMs depends on the prompt!

1. Good LLM, Bad Prompt → Bad evaluation score!
2. Using APE to optimize prompts automatically can remove *guess-work* out of the way
3. Let's take a look at examples of Prompt Engineering by Humans vs Automated prompt and how it can impact evaluation next

Is Prompt Engineering Dead?

Recent IEEE Spectrum Article

Purports that Prompt Engineering - Specifically human prompt engineering is dead. Humans may not be able to engineer prompts as well as LLMs - So do Automated Prompt Engineering for better prompts (think also ToolFormer, Calculator tool and the likes we worked on in the assignments)

Is Prompt Engineering Dead?

Recent IEEE Spectrum Article

Purports that Prompt Engineering - Specifically human prompt engineering is dead. Humans may not be able to engineer prompts as well as LLMs - So do Automated Prompt Engineering for better prompts (think also ToolFormer, Calculator tool and the likes we worked on in the assignments)

Interesting APE Prompt

Command, we need you to plot a course through this turbulence and locate the source of the anomaly. Use all available data and your expertise to guide us through this challenging situation. This prompt of asking it to play the role of Captain Kirk worked better as an initial instruction for the LLM to do well on grade school Math!

Is Prompt Engineering Dead?

Is Prompt Engineering Dead?

Images with and without APE

Is Prompt Engineering Dead?

Is Prompt Engineering Dead?

Recent IEEE Spectrum Article

Purports that Prompt Engineering - Specifically human prompt engineering is dead. Humans may not be able to engineer prompts as well as LLMs - So do Automated Prompt Engineering for better prompts (think also ToolFormer, Calculator tool and the likes we worked on in the assignments)

Is Prompt Engineering Dead?

Recent IEEE Spectrum Article

Purports that Prompt Engineering - Specifically human prompt engineering is dead. Humans may not be able to engineer prompts as well as LLMs - So do Automated Prompt Engineering for better prompts (think also ToolFormer, Calculator tool and the likes we worked on in the assignments)

Interesting APE Prompt

Command, we need you to plot a course through this turbulence and locate the source of the anomaly. Use all available data and your expertise to guide us through this challenging situation. This prompt of asking it to play the role of Captain Kirk worked better as an initial instruction for the LLM to do well on grade school Math!

Is Prompt Engineering Dead?

Image Enhancements with NeuroPrompt

	Original Image	NeuroPrompts Optimized Image
A boy on a horse		

Image Enhancements with NeuroPrompt

A tropical beach with palm trees

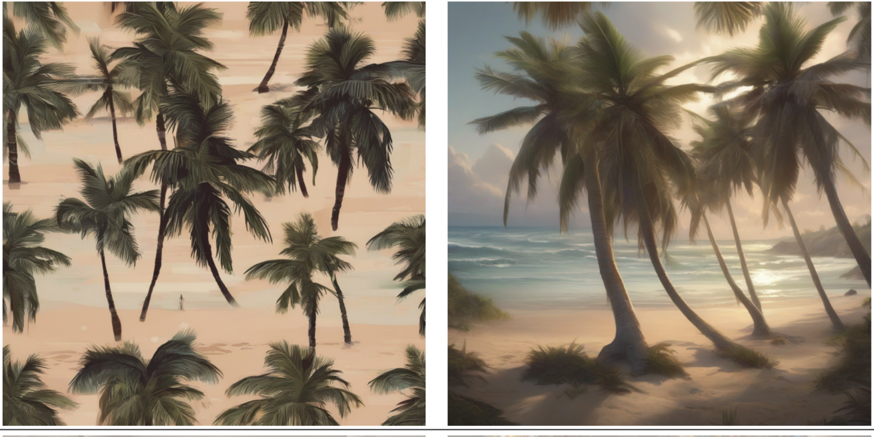


Image Enhancements with NeuroPrompt

Two women working in a kitchen

Image Enhancements with NeuroPrompt

Style	Artist	Format	Boosters	Vibes	Perspective
expressionism	pablo picasso	watercolor painting	trending on artstation	control the soul	long shot
suminagashi	edvard munch	crayon drawing	octane render	futuristic	plain background
surrealism	henri matisse	US patent	ultra high poly	utopian	isometric
anime	thomas cole	kindergartener drawing	extremely detailed	dystopian	panoramic
art deco	mark rothko	cartoon	very beautiful	blade runner	wide angle
photorealism	alphonse mucha	in Mario Kart	studio lighting	cinematic	hard lighting
cyberpunk	leonardo da vinci	pixel art	fantastic	fantasy	knolling
synthwave	claude monet	diagram	postprocessing	elegant	shallow depth of field
realism	james gurney	album art cover	well preserved	magnificent	extreme wide shot
pop art	toshi yoshida	under an electron microscope	4k	retrofuturistic	drone
pixar movies	zdzislaw beksinski	photograph	arnold render	awesome	from behind
abstract organic	gustave doré	pencil sketch	detailed	transhumanist	landscape
dadaism	georges braque	stained glass window	hyperrealistic	bright	1/1000 sec shutter
neoclassicism	bill watterson	advertising poster	rendering	wormhole	from below
ancient art	michelangelo	mugshot	vfx	eclectic	head-and-shoulders shot
baroque	greg rutkowski	cross-stitched sampler	high detail	epic	from above
art nouveau	vincent van gogh	illustration	zbrush	tasteful	oversaturated filter
impressionist	caravaggio	pencil and watercolor drawing	70mm	gorgeous	aerial view
symbolism	diego rivera	in Fortnite	hyper realistic	opaque	telephoto
hudson river school	dean cornwell	line art	8k	old	motion blur
suprematism	ralph mcquarrie	product photography	professional	lsd trip	85mm
rococo	rené magritte	in GTA San Andreas	beautiful	lo-fi	viewed from behind
pointillism	john constable	news crew reporting live	trending on artstation	emo	through a porthole
vaporwave	gustave dore	line drawing	stunning	lucid	dark background
futurism	jackson pollock	courtroom sketch	contest winner	moody	fisheye lens
skeumorphism	hayao miyazaki	on Sesame Street	wondrous	crystal	through a periscope
ukiyo-e	lucian freud	wikiHow	look at that detail	melancholy	white background

Training for prompt enhancement

- **Train Data:** Take a human engineered prompt. Take the prefix of the prompt as input. Take the full prompt as output.
- **Example:** "Image of a boy on a horse. Picasso artist, high definition quality,futuristic soul, long shot perspective"
- **Example Prefix:** "Image of a boy on a horse. Picasso artist"
- **Baseline Model:** Supervised Fine-tuning

Image Enhancements with NeuroPrompt

Model	**Aesthetics Score**
Original prefix	5.64
Original (human) prompt	5.92
SFT only	6.02
NeuroPrompts w/o PPO	6.05
NeuroPrompts w/o NeuroLogic	6.22
NeuroPrompts	6.27

Table 1: Aesthetics scores calculated for images generated by NeuroPrompts and baseline methods

Evaluation Datasets

1. **MATH:** Math problems with 7 difficulty levels
2. **GSM-8k:** Grade School Math
3. **MMLU:** Multitask Accuracy
4. **HumanEval:** Python Coding Tasks
5. **HellaSwag:** Reasoning Tasks
6. **TruthfulQA:** Truthfulness Tasks

Hello Swag Dataset

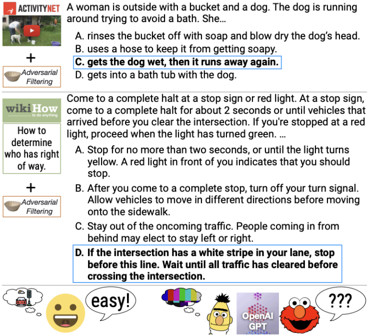


Figure 1: Models like BERT struggle to finish the sentences in *HellaSwag*, even when they come from the same distribution as the training set. While the wrong endings are on-topic, with words that relate to the context, humans consistently judge their meanings to be either incorrect or implausible. For example, option **A** of the WikiHow passage suggests that a driver should stop at a red light for **no more than two seconds**.

Hello Swag Dataset

Category: Shaving (ActivityNet; In-domain)
A bearded man is seen speaking to the camera and making several faces. the man

a) then switches off and shows himself via the washer and dryer rolling down a towel and scrubbing the floor. (0.0%)
b) then rubs and wipes down an individual's face and leads into another man playing another person's flute. (0.0%)
c) is then seen eating food on a ladder while still speaking. (0.0%)
d) then holds up a razor and begins shaving his face. (100.0%)

Category: Sharpening knives (ActivityNet; Zero-Shot)
Two men are in a room and the man with a blue shirt takes out a bench stone and with a little lubricant on the stone takes an knife and explains how to sharpen it. then he

a) uses a sharpener to smooth out the stone using the knife. (100.0%)
b) shows how to cut the bottom with the knife and place a tube on the inner and corner. (0.0%)
c) bends down and grabs the knife and remove the appliance. (0.0%)
d) stops sharpening the knife and takes out some pieces of paper to show how sharp the knife is as he cuts slivers of paper with the knife. (0.0%)

Category: Youth (WikiHow; In-Domain)
How to make up a good excuse for your homework not being finished

Blame technology. One of the easiest and most believable excuses is simply blaming technology. You can say your computer crashed, your printer broke, your internet was down, or any number of problems.

a) Your excuses will hardly seem believable. [substeps] This doesn't mean you are lying, just only that you don't have all the details of how your computer ran at the time of the accident. (0.0%)
b) The simplest one to have in a classroom is to blame you entire classroom, not just lab. If you can think of yourself as the victim, why not blame it on technology. (9.4%)
c) Most people, your teacher included, have experienced setbacks due to technological problems. [substeps] This is a great excuse if you had a paper you needed to type and print. (29.1%)
d) It may also be more believable if you are fully aware that you may be flying at high speed on a plane and need someone to give you traffic report. Your problem might be your laptop failing to charge after a long flight. (61.5%)

Figure 10: Example questions answered by BERT-Large. Correct model predictions are **blue**, incorrect predictions are red. The right answers are **bolded**.

Hello Swag Performance

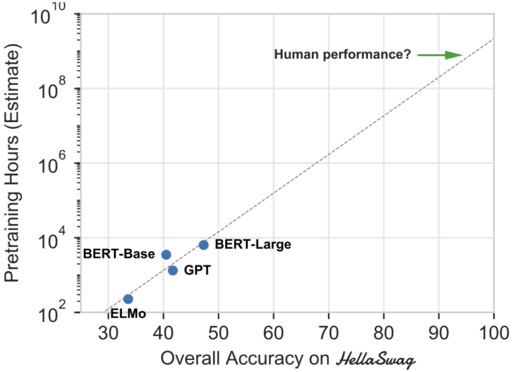


Figure 12: Estimated pretraining hours required to reach a desired accuracy on *HellaSwag*. We estimate perfomance with respect to a RTX 2080 Ti - a modern, fast GPU, and fit a log-linear regression line. An extrapolation suggests that to reach human-level performance on *HellaSwag*, without algorithmic or computa-

HumanEval - Coding Performance

```python
def incr_list(l: list):
    """Return list with elements incremented by 1.
    >>> incr_list([1, 2, 3])
    [2, 3, 4]
    >>> incr_list([5, 3, 5, 2, 3, 3, 9, 0, 123])
    [6, 4, 6, 3, 4, 4, 10, 1, 124]
    """
    return [i + 1 for i in l]
```

```python
def solution(lst):
    """Given a non-empty list of integers, return the sum of all of the odd elements
    that are in even positions.

    Examples
    solution([5, 8, 7, 1]) ==>12
    solution([3, 3, 3, 3, 3]) ==>9
    solution([30, 13, 24, 321]) ==>0
    """
    return sum(lst[i] for i in range(0,len(lst)) if i % 2 == 0 and lst[i] % 2 == 1)
```

```python
def encode_cyclic(s: str):
    """
    returns encoded string by cycling groups of three characters.
    """
    # split string to groups. Each of length 3.
    groups = [s[(3 * i):min((3 * i + 3), len(s))] for i in range((len(s) + 2) // 3)]
    # cycle elements in each group. Unless group has fewer elements than 3.
    groups = [(group[1:] + group[0]) if len(group) == 3 else group for group in groups]
    return "".join(groups)

def decode_cyclic(s: str):
    """
    takes as input string encoded with encode_cyclic function. Returns decoded string.
    """
    # split string to groups. Each of length 3.
    groups = [s[(3 * i):min((3 * i + 3), len(s))] for i in range((len(s) + 2) // 3)]
    # cycle elements in each group.
    groups = [(group[-1] + group[:-1]) if len(group) == 3 else group for group in groups]
    return "".join(groups)
```

HumanEval - Coding Performance

Table 1. Codex, GPT-Neo, & TabNine evaluations for HumanEval. We find that GPT-J pass@1 is between Codex-85M and Codex-300M performance.

	PASS@k		
	$k=1$	$k=10$	$k=100$
GPT-NEO 125M	0.75%	1.88%	2.97%
GPT-NEO 1.3B	4.79%	7.47%	16.30%
GPT-NEO 2.7B	6.41%	11.27%	21.37%
GPT-J 6B	11.62%	15.74%	27.74%
TABNINE	2.58%	4.35%	7.59%
CODEX-12M	2.00%	3.62%	8.58%
CODEX-25M	3.21%	7.1%	12.89%
CODEX-42M	5.06%	8.8%	15.55%
CODEX-85M	8.22%	12.81%	22.4%
CODEX-300M	13.17%	20.37%	36.27%
CODEX-679M	16.22%	25.7%	40.95%
CODEX-2.5B	21.36%	35.42%	59.5%
CODEX-12B	28.81%	46.81%	72.31%

I. All Models Comparison

Model Comparison

Comparison of pre-trained proprietary and open-source models on standard benchmarks for math, science, reasoning, and coding.

	Average	Multi-choice Qs	Reasoning	Python coding	Future Capabilties	Grade school math	Math Problems
Claude 3 Opus	84.83%	86.80%	95.40%	84.90%	86.80%	95.00%	60.10%
Gemini 1.5 Pro	80.08%	81.90%	92.50%	71.90%	84%	91.70%	58.50%
Gemini Ultra	79.52%	83.70%	87.80%	74.40%	83.60%	94.40%	53.20%
GPT-4	79.45%	86.40%	95.30%	67%	83.10%	92%	52.90%
Claude 3 Sonnet	76.55%	79.00%	89.00%	73.00%	82.90%	92.30%	43.10%
Claude 3 Haiku	73.08%	75.20%	85.90%	75.90%	73.70%	88.90%	38.90%
Gemini Pro	68.28%	71.80%	84.70%	67.70%	75%	77.90%	32.60%
Palm 2-L	65.82%	78.40%	86.80%	37.60%	77.70%	80%	34.40%
GPT-3.5	65.46%	70%	85.50%	48.10%	66.60%	57.10%	34.1%
Mixtral 8×7B	59.79%	70.60%	84.40%	40.20%	60.76%	74.40%	28.40%
Llama 2 - 70B	51.55%	69.90%	87%	30.50%	51.20%	56.80%	13.80%
Gemma 7B	50.60%	64.30%	81.2%	32.3%	55.10%	46.40%	24.30%
Falcon 180B	42.62%	70.60%	87.50%	35.40%	37.10%	19.60%	5.50%
Llama 13B	37.63%	54.80%	80.7%	18.3%	39.40%	28.70%	3.9%
Llama 7B	30.84%	45.30%	77.22%	12.8%	32.6%	14.6%	2.5%
Grok 1	-	73.00%	-	63%	-	62.90%	23.90%
Qwen 14B	-	66.30%	-	32%	53.40%	61.30%	24.80%
Mistral Large	-	81.2%	89.2%	45.1%	-	81%	45%

I. All Models Comaprison

- Claude 3 Opus has best average score across all benchmarks
- GPT-4 lags behind
- Mixtral 8x7B is the best open-source model

II. ChatBot Arena Leaderboard

Let's check it out!

II. ChatBot Arena Leaderboard

Arena Elo | Full Leaderboard

Total #models: **73**. Total #votes: **374418**. Last updated: March 7, 2024.

Contribute your vote 🗳️ at chat.lmsys.org! Find more analysis in the notebook.

Rank	🤖 Model	⭐ Arena Elo	📊 95% CI	🗳️ Votes	Organization	License	Knowledge Cutoff
1	GPT-4-1106-preview	1251	+5/-5	45291	OpenAI	Proprietary	2023/4
2	GPT-4-0125-preview	1251	+6/-6	15251	OpenAI	Proprietary	2023/12
3	Claude 3 Opus	1233	+9/-7	5246	Anthropic	Proprietary	2023/8
4	Bard (Gemini Pro)	1203	+6/-8	12623	Google	Proprietary	Online
5	GPT-4-0314	1185	+5/-5	24689	OpenAI	Proprietary	2021/9
6	Claude 3 Sonnet	1180	+10/-8	5259	Anthropic	Proprietary	2023/8
7	GPT-4-0613	1161	+5/-5	39845	OpenAI	Proprietary	2021/9
8	Mistral-Large-2402	1155	+6/-6	9746	Mistral	Proprietary	Unknown
9	Mistral Medium	1147	+5/-4	22171	Mistral	Proprietary	Unknown
10	Qwen1.5-72B-Chat	1147	+4/-5	15288	Alibaba	Qianwen LICENSE	2024/2
11	Claude-1	1146	+5/-6	20833	Anthropic	Proprietary	Unknown
12	Claude-2.0	1127	+6/-5	13679	Anthropic	Proprietary	Unknown
13	Mistral-Next	1124	+5/-6	11875	Mistral	Proprietary	Unknown
14	Gemini Pro (Dev API)	1118	+6/-7	11496	Google	Proprietary	2023/4
15	Claude-2.1	1116	+4/-5	31815	Anthropic	Proprietary	Unknown
16	Mixtral-8x7b-Instruct-v0.1	1116	+5/-5	24824	Mistral	Apache 2.0	2023/12

II. ChatBot Arena Leaderboard

Figure 1: Fraction of Model A Wins for All Non-tied A vs. B Battles

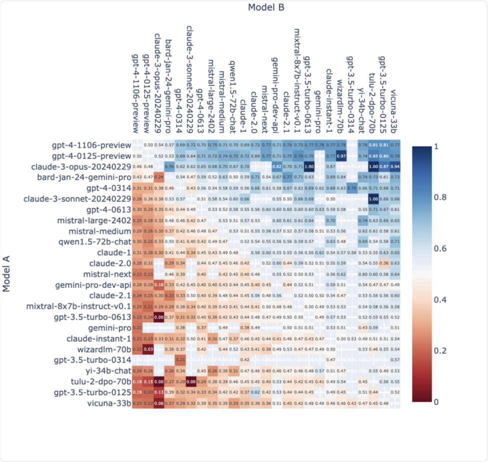
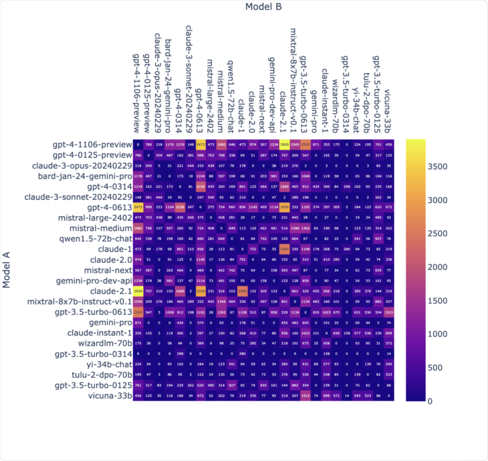
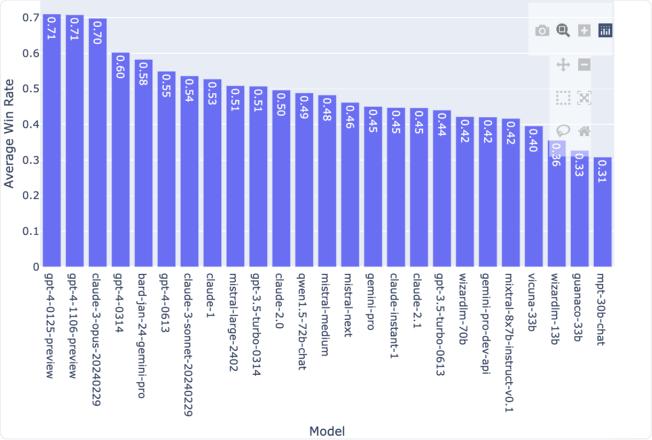


Figure 2: Battle Count for Each Combination of Models (without Ties)



II. ChatBot Arena Leaderboard

Figure 4: Average Win Rate Against All Other Models (Assuming Uniform Sampling and No Ties)



Other Benchmarks

- Performance Benchmarks
- Safety Leaderboard
- Open LLM Leaderboard