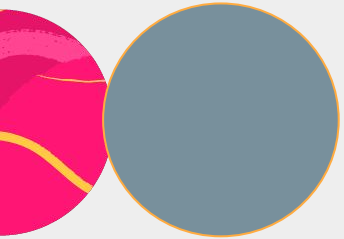

PLURALSIGHT

ChatGPT Prompt Engineering for Developers

Welcome!



Lara Kattan
Instructor

My energy level today

Important URLs for today (reference)

- PLAYGROUND: https://platform.openai.com/playground/chat?models=gpt-4o
- GH (materials): https://github.com/larakattan/prompt_engineering_for_devs
- USAGE DASHBOARD: https://platform.openai.com/usage
- BILLING: https://platform.openai.com/settings/organization/billing/overview
- LIMITS: https://platform.openai.com/settings/organization/limits

How LLM Models Work (Revisited)

The “emergent” behavior of LLMs

- One of the biggest surprises of LLMs is their “emergent” behavior, that is, behavior they exhibit that they weren’t specifically trained to do!
- **Steps of training a general purpose LLM:**
  - Pre-trained on a next token prediction task using lots of data. Called pre-training because it comes before any training for a specific NLP task. **At this stage, called foundation models or base models**
  - Model is then fine-tuned on one or many tasks using labeled data and a given objective. E.g. Chat agents like ChatGPT was trained on conversational data
  - Also fine-tuned on instruction datasets so they know how to follow directions (write me a poem about prompt engineering, e.g.). Others fine-tuned for code generation specifically
  - **Foundation training takes weeks or months; fine-tuning can take minutes**
- Need to be aware that there are limits to the geography and languages used in the pre-training data
- Example datasets: https://huggingface.co/datasets

The “emergent” behavior of LLMs

- Earlier models, like BERT, were intended for language *understanding*, not *generation*
- BERT is bidirectional: trained to predict a missing word with access to words on the left and right
- Models like GPT are *autoregressive* because they don’t have access to the future (right context)
- **Emergent properties** are behaviors that aren’t necessarily predictable from a scaling law (where more data leads to better performing models)
    - Scaling law: model with 100M parameters should be 10% better than model with 10M parameters, but the larger model can actually do things the smaller model can’t do at all!
    - No one knew that using billions of parameters would give us ChatGPT!
    - **Zero-shot learning is one of those emergent properties**
- **No research consensus on what emergence is, what counts as emergence, etc.**
    - That is, is the emergence a *qualitative* shift in the model once it reaches some threshold of parameters, or are researchers just using different metrics on these models?

AI Bias and Ethics

Issues with the data used to train LLMs

- LLMs were trained on internet data, which contains a lot of bias and harmful content. Plus, the content that is available may under-represent certain communities and points of view
- LLMs can thus perpetuate bias and harm!
- Some studies show word embeddings encode bias (e.g. man -> woman is same as doctor -> nurse in the semantic space of some studied models. Yikes!)
    - Research paper showed translation from non-gendered language to English perpetuated bias. The Turkish “O bir doktor. O bir hemsire” became the English “He is a doctor. She is a nurse”

How to fix for bias in LLMs

- So, is debiasing the training data possible?
- Filtering data could unintentionally remove minority voices
    - E.g. filtering out documents with high degree of discussion of sex, racial slurs, white supremacy could filter out discussion within minority groups
- Debiasing could unintentionally remove meaning
- LLMs not frequently updated, so may not represent shifts in content and social views
    - E.g. Explosion of data on police violence post-BLM, but not reflected in the (older) models
- Because LLMs trained on personal data, attackers can use training data extraction attacks to try to recover personally identifiable information that's been memorized by the model
    - Researchers built model on Enron emails and were able to extract SSNs and credit card numbers

How to fix for bias in LLMs

- So, is debiasing the training data possible?
- Filtering data could unintentionally remove minority voices
  - E.g. filterin… could filter… …hite supremacy
- Debiasing could…
- LLMs not freque…
  - E.g. Explos… …r) models
- Because LLMs t… …acks to try to recover persona…
  - Researche… …dit card numbers

Your turn:

Take 15 minutes to research either:

- New debiasing techniques
- Other ethical issues with generative AI

Report back what you find!

Risks of using LLMs

- Models don't know what they don't know, meaning they are **confident in their wrong answers**
- Model providers don't want LLMs to product harmful content (is it a genuine interest, or are they mostly concerned about the financial impacts of bad press? Who knows!)
- **AI alignment:** industry term for research into how to align what ML models do with what their creators / users want them to do
- How to stop models from being toxic: for now, mostly post-hoc testing, detection, and suppressing undesirable output. Using a classifier on generated output before showing it to the user, falling back to default language when the output isn't acceptable
- **Helpful vs harmful models:** if you ask a model to write you something from a harmful point of view, it's being helpful if it gives you what you asked for. But, it's not a desirable output...
- If toxic text is removed pre-training, can be good, but could also make it hard for the model to be able to identify toxic prompts
- **What's the cost of subjecting human reviewers to toxic content for testing?**

How to know if output is AI-generated

- When generative models are used for harmful activities (cheating, spreading misinformation), we may want to know how to identify AI-generated content
- The better AI-generated content becomes, the harder it is to identify
- Possible ideas that have been considered:
  - Since LLMs are just probability machines, look at output and look for words that are more likely to be generated by an LLM
  - … easy to bypass by slightly changing the AI-generated output
  - Putting a “model watermark” in models by seeding them to use certain words or constructions more often than they would have otherwise
  - … doing this requires making the “seeds” public so people can use them against text they want to verify, but then they can also be easily used by bad actors to remove the same seeds from their own outputs
- OpenAI released their own classifier in early 2023 to detect AI text, but it was so bad they took it down a few months later

How to know if output is AI-generated

- When generative models are used for harmful activities (cheating, spreading misinformation), we may want to know how to identify AI-generated content
- The better AI-gene
- Possible ideas that
  - Since LLMs
    for words th
  - … easy to by
  - Putting a "n
    certain word
    otherwise
  - … doing this requires making the "seeds" public so people can use them against text they want to verify, but then they can also be easily used by bad actors to remove the same seeds from their own outputs
- OpenAI released their own classifier in early 2023 to detect AI text, but it was so bad they took it down a few months later

Your turn:

Take 15 minutes to find your favorite example of viral AI-generated fakes (can be text, image, video, etc).

Use of AI generation in the workplace

- Assume the model is wrong, double-check output. Models will be confident in their wrong answers
- Don’t feed models data that is proprietary (including code), personally identifiable, or legally protected. If you wouldn’t post it on a public GitHub repo, don’t tell it to a GPT
- Create internal policies and guidelines for where/how AI tools can be used
- Don’t use people’s data without their consent
- Understand relevant data protection laws (stronger in the EU than in the US)
- Many instances so far of professionals using AI when they shouldn’t (e.g. lawyer using it to write court brief, except ChatGPT made up all of these referenced cases!)

Exercise: AI ethics debate

Split into 4 groups. Each group will give a presentation on an AI topic:

1. **AI is ethical** and any impact on employment, such as millions of people losing their jobs to automation, is socially acceptable
2. **AI is unethical** and it's socially wrong to continue developing it at this time. Mitigations and safeguards for privacy and economic impact need to be developed first
3. AGI (artificial general intelligence) is a real threat. It's unethical to continue building AI until we can prove that it will not have AGI
4. AGI is not a real threat. Either it's decades (or more) away, or it would not pose a threat to humans, animals, and Earth if it were to be developed

Evaluating AI Language Models

You’ve written some prompts, now what?!

- Evaluating machine learning models, particularly language models, can be tricky. With language models, it’s more difficult to use quantitative metrics to measure performance
- **Perplexity:** Perplexity is a measurement of how well a probability distribution or probability model predicts a sample. It's a standard metric for language models, measuring how well a model predicts a test set. Lower perplexity indicates a better predictive performance
- **Human evaluation:** Since language models are ultimately designed to perform tasks involving human language, human judgment is often considered the gold standard for evaluation. Human evaluators assess the quality, coherence, relevance, and fluency of the text generated by the model. This method is particularly useful for tasks like dialogue generation, story writing, or any other output meant to be consumed by humans

You've written some prompts, now what?!

- **Quantitative metrics:**
  - BLEU (Bilingual Evaluation Understudy): Originally developed for evaluating machine translation quality by comparing a machine-generated translation to one or more reference translations
  - ROUGE (Recall-Oriented Understudy for Gisting Evaluation): Often used in summarization tasks to evaluate how many of the same words (n-grams) appear in the output compared to a set of reference summaries
  - METEOR (Metric for Evaluation of Translation with Explicit ORdering): An improvement over BLEU, considering synonyms and stemming, and incorporates both precision and recall
  - More tests here
- **A/B testing:** In real-world applications, A/B testing involves comparing two versions of a model to see which performs better on specific tasks. This approach can help understand the impact of model changes on performance in practical applications
- **Adversarial testing:** This method checks the robustness of a model by trying to fool it with tricky or misleading input. The goal is to see if the model can maintain performance when faced with potentially deceptive or unexpected inputs

Evaluating language models

- **Human review and fine-grained analysis:** For some applications, it might be important to perform detailed analysis on specific aspects of model output:
  - **Consistency:** Whether the model's outputs are consistent over similar or repeated queries
  - **Fairness and bias:** Evaluating whether the model outputs are fair and unbiased. This involves checking for discriminatory or inappropriate content that might reflect or amplify social biases
  - **Domain-specific tasks:** For models deployed in specific domains like law, medicine, or customer service, domain experts might evaluate the accuracy and utility of outputs in context
- **Fidelity and factuality:** Particularly for models used in news generation or summarization, it's important to assess how accurately the model represents factual information. This can involve comparing model outputs against trusted sources to check for misinformation or inaccuracies
- **Speed and efficiency:** For many applications, how quickly and resource-efficiently a model can generate responses is critical. This might involve metrics like latency, throughput, and computational resource usage
- **Interactivity and engagement:** For conversational AI, metrics like user engagement duration, turn-taking behavior, and the appropriateness of responses in interactive settings are important

Human evaluation criteria

- **Relevance:** How relevant is the generated text to the given prompt or context?
- **Fluency:** How natural and grammatically correct is the generated text?
- **Coherence:** How logically consistent and well-structured is the generated text?
- **Creativity:** How creative and original is the generated text?
- **Accuracy:** How factually accurate is the generated text (if applicable)?
- **Completeness:** Does the generated text sufficiently address the prompt or task?
- **Engagement:** How engaging and interesting is the generated text?
- **Appropriateness:** Is the generated text appropriate in terms of tone, style, and content for the given context or audience?

Human review exercise

1. Split into 4 groups. Each group will write one prompt against the following article: https://www.scientificamerican.com/article/first-promethium-complex-created-revealing-mysterious-elements-secrets
2. Then, each group will present their prompt/output, and the other three groups will judge the output by the human rater criteria discussed above https://pollev.com/larak279

Code: Evaluating LLM metrics

LLM Security

SAIF framework (link to source)

1. Expand strong security foundations to the AI ecosystem

- This includes leveraging secure-by-default infrastructure protections and expertise built over the last two decades to protect AI systems, applications, and users. At the same time, develop organizational expertise to keep pace with advances in AI and start to scale and adapt infrastructure protections in the context of AI and evolving threat models. For example, injection techniques like SQL injection have existed for some time, and organizations can adapt mitigations, such as input sanitization and limiting, to help better defend against prompt injection-style attacks.

2. Extend detection and response to bring AI into an organization's threat universe

- Timeliness is critical in detecting and responding to AI-related cyber incidents, and extending threat intelligence and other capabilities to an organization improves both. For organizations, this includes monitoring inputs and outputs of generative AI systems to detect anomalies and using threat intelligence to anticipate attacks. This effort typically requires collaboration with trust and safety, threat intelligence, and counter abuse teams.

3. Automate defenses to keep pace with existing and new threats

- The latest AI innovations can improve the scale and speed of response efforts to security incidents. Adversaries will likely use AI to scale their impact, so it is important to use AI and its current and emerging capabilities to stay nimble and cost effective in protecting against them.

SAIF framework (link to source)

4. Harmonize platform level controls to ensure consistent security across the organization

- Consistency across control frameworks can support AI risk mitigation and scale protections across different platforms and tools to ensure that the best protections are available to all AI applications in a scalable and cost efficient manner. At Google, this includes extending secure-by-default protections to AI platforms like Vertex AI and Security AI Workbench, and building controls and protections into the software development lifecycle.

5. Adapt controls to adjust mitigations and create faster feedback loops for AI deployment

- Constant testing of implementations through continuous learning can ensure detection and protection capabilities address the changing threat environment. This includes techniques like reinforcement learning based on incidents and user feedback and involves steps such as updating training data sets, fine-tuning models to respond strategically to attacks and allowing the software that is used to build models to embed further security in context (e.g. detecting anomalous behavior). Organizations can also conduct regular red team exercises to improve safety assurance for AI-powered products and capabilities.

6. Contextualize AI system risks in surrounding business processes

- Lastly, conducting end-to-end risk assessments related to how organizations will deploy AI can help inform decisions. This includes an assessment of the end-to-end business risk, such as data lineage, validation and operational behavior monitoring for certain types of applications. In addition, organizations should construct automated checks to validate AI performance.

Types of LLM attacks (there are many!)

- **Prompt injection attacks:**
  - Manipulative input: By crafting specific prompts, attackers can manipulate the model to produce harmful or biased outputs. For example, asking the model to provide harmful advice or generate inappropriate content
  - Context injection: Embedding malicious prompts within otherwise benign input to deceive the model into producing unintended responses
- **Adversarial attacks:**
  - Text perturbation: Slightly altering the input text (e.g., typos, synonym replacement) to cause the model to produce incorrect or harmful outputs. These perturbations are often imperceptible to humans but can significantly impact the model's performance
  - Adversarial examples: Crafting inputs that are specifically designed to exploit weaknesses in the model and cause it to fail in specific ways
- **Data poisoning:**
  - Training data manipulation: Injecting malicious data into the training dataset to bias the model's behavior or degrade its performance. This can be done by adding harmful examples or by subtly altering legitimate examples
  - Prompt-based data poisoning: Introducing poisoned prompts into the dataset used for prompt-based learning, leading the model to produce undesirable outputs when encountering similar prompts

Types of LLM attacks (there are many!)

- **Model inversion and extraction:**
  - Model inversion: Inferring sensitive information from the model's outputs. For instance, attackers can try to recover training data or private information by analyzing the model's responses
  - Model extraction: Reconstructing a copy of the model by systematically querying it and using the outputs to train a surrogate model. This can lead to intellectual property theft and exploitation of the model's capabilities
- **Bias exploitation:**
  - Amplifying bias: Crafting inputs that exploit the model's inherent biases, causing it to produce biased, harmful, or inappropriate content. This can be used to highlight and exacerbate societal biases present in the training data
- **Contextual attacks:**
  - Contextual misleading: Providing misleading context or framing the input in a way that causes the model to generate incorrect or harmful outputs. This is often done by exploiting the model's tendency to follow conversational or narrative cues

Types of LLM attacks (there are many!)

- **Evasion attacks:**
  - Bypassing content filters: Crafting inputs that bypass content moderation filters by using obfuscated or coded language. This allows users to generate harmful content without triggering the model's safety mechanisms
- **DDoS and resource exhaustion:**
  - Overloading with queries: Sending an overwhelming number of queries to the model to exhaust its computational resources or degrade its performance. This can be used to launch Denial-of-Service (DoS) attacks on services that rely on LLMs
- **Phishing and social engineering:**
  - Crafting deceptive content: Using the model to generate convincing phishing emails, fake news, or other deceptive content aimed at manipulating individuals or groups. This exploits the model's ability to produce high-quality, human-like text

Types of LLM attacks (there are many!)

- **Evasion attacks:**
  - Bypassing content filters: Crafting inputs that bypass content moderation filters by using obfuscated or coded language. This allows users to generate harmful content without triggering the model's safety mechanisms
- **DDoS and resourc**
  - Overloading ust its computatio rvice (DoS) attacks on services t
- **Phishing and soci**
  - Crafting dec vs, or other deceptive c ty to produce high-quality

Your turn:

Take 15 minutes to research either:

- Your favorite example of one of these attacks
- Any other attack types not listed here

End of course survey: link

Questions?