

Context Engineering & RAG

Beyond the Prompt

Administrative Details

- Attendance secret code! Reflecting All of Us
 - <https://canvas.stanford.edu/courses/221239/quizzes/184809>
- Context Engineering supplanting Prompt Engineering
- Please invite our github account cs224g to your repo
 - Invite CAs too
- Assessment Methodology
 - Clone repo, study README, review commits (each team member contributes)
 - Follow code set up instructions, Run code



Team Presentations

- 2 slides
- 3 minutes per team
- Be sure to cover:
 - Problem you are solving and for whom - 30 seconds
 - Your solution - 30 seconds
 - Show a quick demo (screenshots, video) - 60 seconds
 - Your plan vs actual - 30 seconds
 - Something unique or interesting you learned - 30 seconds
- Share a link in your Slack channel, we'll add them to the deck for Thursday
 - Make sure permissions are world-readable!

Goals

- Engage the audience
- Smooth transitions (practice to make it effortless)
- Make it memorable/memeable (fun, unexpected, to the point)
- Tell your story (now, the plan, the outcome)
- Make us look forward to the next presentation

Digressions of a Health Care AI Caller

Agent: So what kind of melodious music do you like?

AI-Caller: I enjoy classical and instrumental music.

Agent: Which classical?

AI-Caller: Mozart and Beethoven are favorites.

Agent: Is it Mozart? Well, isn't Beethoven more of a romantic era composer?

AI-Caller: Yes, Beethoven bridges Classical and Romantic eras.

Agent: Sure. Do you have a favorite composition?

AI-Caller: Beethoven's Symphony No. 9 is excellent.

Agent: I think that's generally agreed upon.

AI-Caller: Indeed, it's widely celebrated.

Agent: How do you feel about that?

AI-Caller: I don't have feelings, but appreciate it.

Agent: Do you prefer the B minor Mass or the Saint Matthew's Passion?

AI-Caller: Both are exceptional, but I prefer B Minor Mass.

Agent: Wow, that's the standard answer, isn't it?

AI-Caller: Yes, it's a popular choice.

Agent: What about Wagner?

AI-Caller: Wagner's operas are powerful and influential.

Agent: Debatable. Influential Sir.

AI-Caller: Certainly, his influence is widely debated.

Agent: Ends with Nietzsche.

AI-Caller: True, Nietzsche critiqued Wagner's work.

Agent: He actually used to live with him for a while. He kind of thought he was like the man.

AI-Caller: Yes, they had a complex relationship.

Agent: In fact, Nietzsche fancied himself a piano player.

AI-Caller: Interesting, he had diverse talents.

Agent: Not bad for a blind man. Well, mostly blind, apparently. There's also this keyboard he was really fond of that was like a sphere. Do you know what that is?

AI-Caller: Not sure, but sounds intriguing.

Prelecture warmup : Mind Tools

- Memory Palace
 - Music, Oral Histories, Poetry
 - Zettelkasten
- Spaced Repetition
- Mathematics as a “grammar of reality”
 - Rudy Rucker : 1987 five modes
 - Number, Shape, Logic, Infinity, Information
- Checklist Manifesto
 - Atul Gawande
- GTD
 - David Allen

Prompts, Context & RAG

Basics

- Input text -> Output text
 - Context
 - Prompt
 - Examples
 - Format
- Order matters
 - Linear sequence, avoid back references
 - Broad to specific, zoom in
 - Provide hints, but don't tempt the model to copy
 - Final instructions offer key details
- Offline vs. Realtime vs. Multimodal
 - Leaning into vibes for this presentation

Models as Non Deterministic CPUs

- Each Prompt is an OP code up to 10 million tokens (100 books) long
- Outputs are equally stupendous (1 book)
- The space of possible inputs outputs basically infinite
- But remarkably, the outputs connect with our experience
- Modulating the inputs produces highly correlated output changes
- Context engineering is the method of reliably recovering desired results

Models as Smart, Overeager, Inexperienced Interns

- Tremendous amount of knowledge
- Compelled to answer no matter what
- No experience combining facts
- Extremely literal
- Can be lazy, look for shortcuts
- Mimic the style of the prompter

AI Engineer as High School Play Director

- Need to set the stage and the atmosphere
- Provide the props and the costumes
- Put the models in the mindset to perform
- Teach the nuance of the dialog
- Set the characters in motion

Ingredients

- System Prompt
- Role (smarts / ambition)
- Tone (polite / severe)
- Energy (urgency / reward / penalty / gaslighting)
- Response Structure (brevity / simplicity)
- Guardrails (focus / distraction handling / hallucination prevention)

Tips

- Linear sequencing of text (up to 5K words)
- Instruction repetition
- Up to 5 times for major emphasis
- Separate data from instruction (prevent prompt injection)
- Plaintext structure : Markdown, yaml, JSON
- Tool calling, “code” generation

Running Example : AI Injury Attorney

- Lance Chase wants to revolutionize his injury attorney practice with AI
- His firm : ambuLANCE CHASEr needs an edge against competitors TV ads
- Solution : process police accident reports to find the potentially juiciest cases
- OCR docs, find injuries, identify defendants that can pay, friendly jurisdictions
- Now Lance can get ahead of the rest of the pack

Initial Prompts

- Start with image input of a page and prompt the model to OCR the contents
- In the text that follows, identify by name the injured persons, the responsible party and the county in which the accident occurred, and using that information estimate what the case is worth.
... OCR text

- What are some immediate improvements?
- How can Lance consistently get the answers he needs?

Add context, better prompt detail

- You are the top paralegal in a law firm that specializes in recovering the largest settlements possible for your clients injured in road accidents
- In the OCR text, delimited by triple dollar signs, from a police accident report, identify the accident location, find the names and addresses of the responsible parties as well as of all of the injured parties. First output the accident location, followed by the list of responsible parties names and addresses, and finally the list of injured parties names and addresses
\$\$\$ OCR text \$\$\$

Add Examples, More Detail, Specify Format

- For the accident location, specify the county if it is present in the text.
- If the responsible parties are not clearly indicated in the text, use your understanding of traffic accident investigation to determine them
- If the responsible parties are part of a business or enterprise, list it as well.
- If there are no injuries indicate this clearly and stop processing any further
- Indicate paramedic care and/or hospitalization for each injured party
- Estimate the maximum payout to injured parties based on similar cases in the text delimited by triple asterisks

*** Case Verdicts ***

When to Add RAG

- Relevant content much larger than context window
- Content is unique/proprietary and not part of any model training set
- <https://medium.com/data-science-collective/rag-systems-in-5-levels-of-difficulty-with-full-code-443180a7dc59>
- Embed into model parameter space
- Right sized overlapping chunks,
- Combine with keyword search and
- Prompting to find best chunk while
- Avoiding null answers

Notes on Format and Structure

- So much focus on the amazing language capabilities of AI models
- Equally astounding skill with formats and structure
 - Organizational grammar
- Input format as much as output format
 - OCR to Markdown
- yaml as an input language
 - Often superior results to JSON
- Take advantage of structured outputs
 - Still check proper generation of JSON using tools like jq

Error Cases

- Avoid operations on numbers (low training data)
 - Keep explicit number references in the low 2 digits
 - What is 13ness? What is 1729ness? What is the first boring number?
- Avoid scoring results (model is inexperienced)
 - Keep decisions binary whenever possible
- Avoid expansive responses and complex structures
 - Double check structured outputs, later outputs
- Avoid empty data contexts
 - Guaranteed hallucinations, great insight into garden variety hallucination

The List Goes On

- Split data extraction from analysis
- Add review prompting
- Specify list formats and address syntax
- Make sure the model only extracts from the source documents
- Repeat and affirm the most important results
- Emphasize when relevant that negative results are as important as positive
- Use AI to expand the prompts on these topics

Summary

- A lot of common sense ideas with some surprises
 - We tend to undervalue repetition and structure
 -
- Some of these techniques are effective for speech writing as well
- Shared some key observed Hacks, and Gotchas
 - Scientific results are really starting to replace vibes
- Why does tone seem to matter so much?
- Many areas to continue investigating...