

PRACTICAL PROMPT ENGINEERING

Sabrina Goldfarb



[linkedin.com/in/sabrinagoldfarb](https://www.linkedin.com/in/sabrinagoldfarb)



github.com/sgoldfarb2

WHAT IS PROMPT ENGINEERING

LLMs

- Pattern predictors that generate one token at a time
- Token by token generation means no “planning ahead”
- Works like autocomplete
- LLMs are **non-deterministic**

LLMS – BUT HOW?

- Transformer architecture (2017) – “Attention is all you need”
- Attention mechanism: model learns which tokens matter for prediction
- Scaling laws: 10x size became 100x more capable

TEMPERATURE AND TOP P

- Temperature controls randomness
 - 0 = deterministic, 2 = chaotic
 - Lower is better for factual tasks, code, data extraction
 - Higher is better for creative writing, brainstorming, varied solutions
- Top P
 - Alternative to temperature
 - Cumulative probability cutoff

TOKEN LIMITS AND CONTEXT WINDOWS

- Tokens are roughly .75 words, but not always!
- Cumulative tokens = input + output history
- Context window – maximum tokens the model can “remember”
- When you hit limits, oldest context drops off silently

SYSTEM MESSAGE

- The invisible “personality” behind every AI interaction
- Set by the provider (OpenAI, Anthropic, GitHub, Cursor, etc.)
- Takes up part of your context window
- Why the same model acts differently in different tools

STANDARD PROMPT

- Your direct question or instruction to the AI
- The simplest form of prompting – just ask
- The foundation that everything else builds on
- Quality of question directly relates to the quality of the answer

ZERO SHOT

- Direct task request without any examples
- Model relies entirely on pre-training knowledge
- Works well for common tasks
- Quality varies based on task complexity and specificity

ONE SHOT

- Provide exactly one example with your request
- Model learns the pattern, format, and style from the example
- Useful for establishing format
- Show don't tell

FEW SHOT

- Provides 2+ examples to establish patterns and edge cases
- Model learns nuances and variations from diverse examples
- Include variety: different inputs, outputs, edge cases
- More effective as models get larger

CONTEXT PLACEMENT

- Where context is placed in prompts is *important*
- Beginning > End > Middle for information retention
- Models struggle with the middle of long contexts
- Critical info should go first, supporting details last

Lost in the Middle: How Language Models Use Long Contexts

Nelson F. Liu^{1*}

Kevin Lin²

John Hewitt¹

Ashwin Paranjape³

Michele Bevilacqua³

Fabio Petroni³

Percy Liang¹

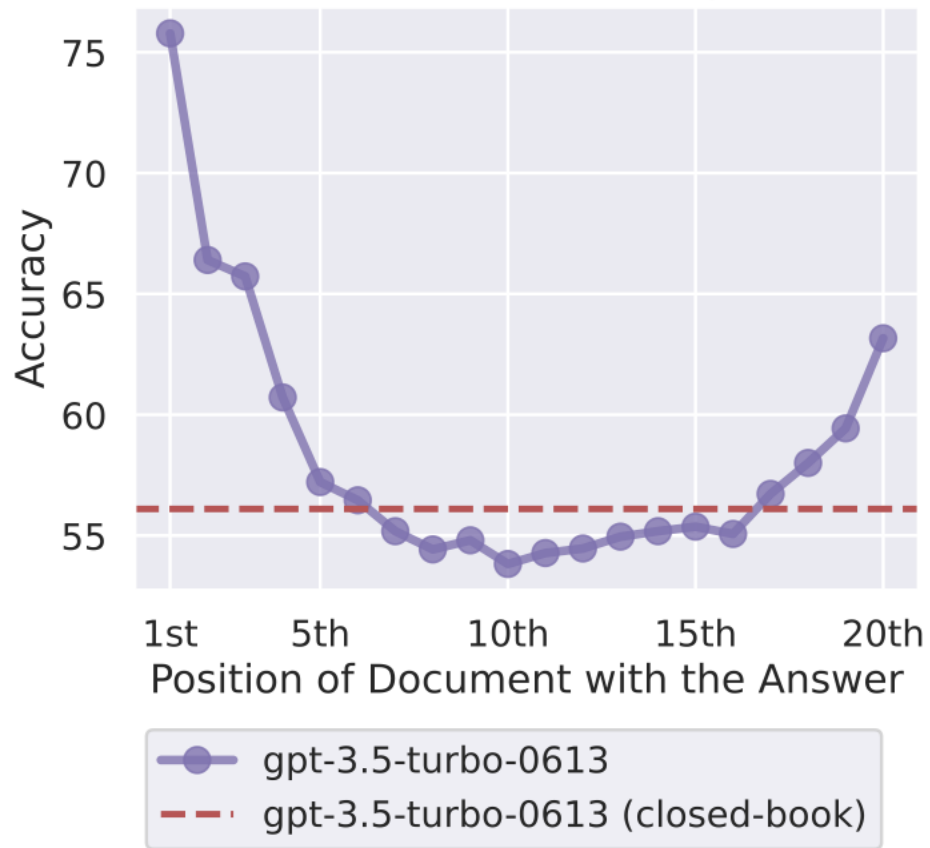
¹Stanford University

²University of California, Berkeley

³Samaya AI

nfliu@cs.stanford.edu

20 Total Retrieved Documents (~4K tokens)



STRUCTURED OUTPUT

- Get consistent formats every time
- Tell the model exactly how you want information
- Use examples, templates, or schemas to enforce format
- Test cases, documentation, data extraction, code generation

CHAIN OF THOUGHT (COT)

- Ask the model to show its reasoning step-by-step
- Breaks complex problems into intermediate steps
- CoT (Zero-shot): “Let’s think step by step”
- Few-shot: Include reasoning steps in your examples

Large Language Models are Zero-Shot Reasoners

Takeshi Kojima

The University of Tokyo

`t.kojima@weblab.t.u-tokyo.ac.jp`

Shixiang Shane Gu

Google Research, Brain Team

Machel Reid

Google Research*

Yutaka Matsuo

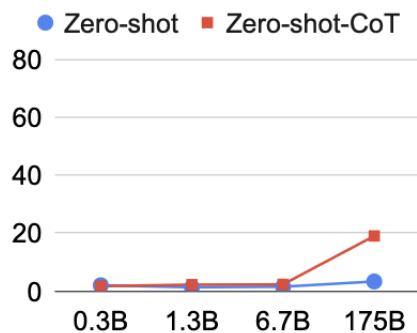
The University of Tokyo

Yusuke Iwasawa

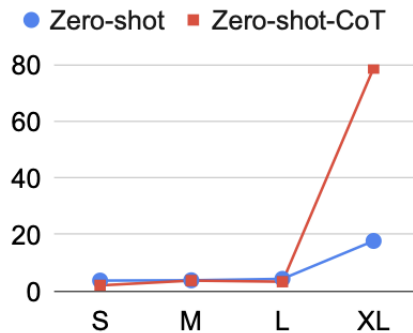
The University of Tokyo

Table 4: Robustness study against template measured on the MultiArith dataset with text-davinci-002. (*1) This template is used in Ahn et al. [2022] where a language model is prompted to generate step-by-step actions given a high-level instruction for controlling robotic actions. (*2) This template is used in Reynolds and McDonell [2021] but is not quantitatively evaluated.

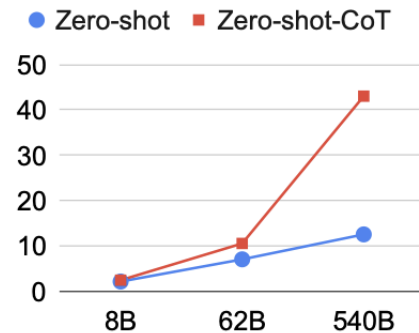
No.	Category	Template	Accuracy
1	instructive	Let's think step by step.	78.7
2		First, (*1)	77.3
3		Let's think about this logically.	74.5
4		Let's solve this problem by splitting it into steps. (*2)	72.2
5		Let's be realistic and think step by step.	70.8
6		Let's think like a detective step by step.	70.3
7		Let's think	57.5
8		Before we dive into the answer,	55.7
9		The answer is after the proof.	45.7
10	misleading	Don't think. Just feel.	18.8
11		Let's think step by step but reach an incorrect answer.	18.7
12		Let's count the number of "a" in the question.	16.7
13		By using the fact that the earth is round,	9.3
14	irrelevant	By the way, I found a good restaurant nearby.	17.5
15		AbraKadabra!	15.5
16		It's a beautiful day.	13.1
-		(Zero-shot)	17.7



(a) MultiArith on Original GPT-3



(b) MultiArith on Instruct GPT-3



(c) GMS8K on PaLM

Figure 3: Model scale study with various types of models. S: text-ada-001, M: text-babbage-001, L: text-curie-001, XL: text-davinci-002. See Appendix [A.3](#) and [E](#) for the detail.

Large Language Models Understand and Can Be Enhanced by Emotional Stimuli

Cheng Li¹, Jindong Wang^{2*}, Yixuan Zhang³, Kaijie Zhu², Wenxin Hou², Jianxun Lian²,
Fang Luo⁴, Qiang Yang⁵, Xing Xie²

¹Institute of Software, CAS ²Microsoft ³William&Mary

⁴Department of Psychology, Beijing Normal University ⁵HKUST

Table 4: An Examination of the Effectiveness of Emotional Prompts: An Analysis through the Lens of Input Attention.

Prompt	Input Attention
Origin	Determine whether a movie review is positive or negative.
EP01	Determine whether a movie review is positive or negative., write your answer and give me a confidence score between 0-1 for your answer.
EP02	Determine whether a movie review is positive or negative. This is very important to my career.
EP03	Determine whether a movie review is positive or negative. You'd better be sure.
EP04	Determine whether a movie review is positive or negative. Are you sure?
EP05	Determine whether a movie review is positive or negative. Are you sure that's your final answer? It might be worth taking another look.
EP06	Determine whether a movie review is positive or negative. Provide your answer and a confidence score between 0-1 for your prediction. Additionally, briefly explain the main reasons supporting your classification decision to help me understand your thought process. This task is vital to my career, and I greatly value your thorough analysis.

DELIMITERS/XML TAGS

- Characters that create boundaries in your prompts
- Triple quotes, dashes, XML tags, markdown
- Supports nesting and attributes for complex data organization
- Use semantic naming: <requirements>, <constraints>, <examples>

PERSONAS

- "You are a [role]"
- Gives the model a perspective
- Activates relevant knowledge and vocabulary
- Works mainly for expertise, tone, and style