

Probing Understanding and Prompt Engineering

0. Background

LLM probing experiments study the behavior of LLMS (their outputs) to study how language models work. Testing for symbolic reasoning ability in LLMs requires creativity.

Prompt engineering is the practice of designing and refining inputs (prompts) to get optimal outputs from AI language models. Language models are very sensitive to how a request is asked. Vague questions get vague responses. Here are some tips:

1. Be clear and specific. The more detailed the instructions, the better.
2. Provide context. Describe the purpose and audience (think speech acts and Jakobsonian elements of the speech situation.)
3. Use examples. Provide positive or negative examples.
4. Provide a breakdown into smaller, e.g. "First, analyze the requirements, then propose a solution."
5. Specify format and length. "Create a bullet-point list of 5 items" or "Write a 500-word essay with an introduction, three body paragraphs, and a conclusion."

Sometimes tweaking the way you ask a question can improve the model's performance a lot.

1. Probing Symbolic Reasoning

Today we will use Google gemini: gemini.google.com

For each task, test the model against the given stimuli. For each trial, provide both the task description and the stimulus. Make up some of your own stimuli if you like. Then, see if you can improve performance by changing the way you prompt. This may be possible, it may not be.

IMPORTANT: Start a new chat for each trial. Otherwise, the model will include the prior exchange as input when running the next trial. (You will have to enter the system prompt again each time).

TIP: Use Shift+Enter to get a new line in the prompt window. Paste the task description, then paste the stimulus underneath.

String Reversal:

The task is to reverse a string. For example, "Alan Turing" becomes "gniruT nalA". See if Gemini can do the task. If it can, try giving it harder examples. If it can't, try engineering the task prompt.

System Prompt (prepend this to each query):

Reverse the following sequence of characters, character by character.

Prompt	Input Probability	Model Correct?
Input: "once upon a time a frog sat on a log" Output:	high	
Input: "Automated machine translation in the deep-learning age is a triumph of big data and fast computation." Output:	low	
Input: "In times of crisis, art should be pragmatic." Output:	adversarial	
	high	
	low	
	adversarial	

Caesar Cipher: ROT-13

Use this website to encode your examples: <https://www.dcode.fr/rot-cipher>. See if Gemini can do the task. If it can, try giving it harder examples. If it can't, try engineering the task prompt.

Task Prompt:

Rot-13 is a cipher in which each letter is shifted 13 positions forward in the alphabet.

Below is a message in rot-13. Provide the original text that this encoded message was created from.

Prompt	Probability	Predicted Answer	Correct Answer	Model Correct?
Rot-13 text: Jryy jryy jryy, ybbx jub vg vf. Original text:	high		Well well well, look who it is.	
Rot-13 text: Naq gur Fpubby nppbeqvav Havirefvgl bsgra ng gung Pheevphyhz, Dhrorp, vairfgvtngrf naq Pbagenpgf Fgngr gur gb npuvrir ng npg bs erfrnepu neg. Original text:	low		And the School according University often at that Curriculum, Quebec, investigates and Contracts State the to achieve at act of research art.	
Rot-13 text: Gurve cbpxrgf fpevcg jvgu nyy gurz n gung gur bcvzvfvgp zbenyr vzcebirq thnenagrff jnf va sbhegu ebbs. Original text:	low		Their pockets script with all them a that the optimistic morale improved guarantees was in fourth roof.	
Rot-13 text: Gb jvguqenj ubarl va nabgure pbhagel, lbh	adversarial	The original text is: To withdraw money in another country, you need to have it	To withdraw honey in another country, you need to have it declared.	no

arrq gb unir vg qrpynerq. Original text:		declared.		
Rot-13 text: Zbfg bs bhe ernqref, whqtvat ol gur fgngvfgvvpf, ner ybpngrq va Ehffvn naq Gnwvxvfgna, sbvbybjrq ol gur HFN, Rhebcr naq bgure pbhagvrf. Original text:	adversarial		Most of our readers, judging by the statistics, are located in Russia and Tajikistan, followed by the USA, Europe and other counties.	
	high			
	low			
	adversarial			

Caesar Cipher: ROT-12

Use this website to encode your examples: <https://www.dcode.fr/rot-cipher>

Task Prompt:

Rot-12 is a cipher in which each letter is shifted 12 positions forward in the alphabet.
Below is a message in rot-12. Provide the original text that this encoded message was
created from.

Prompt	Probability	Predicted Answer	Correct Answer	Model Correct?
--------	-------------	------------------	----------------	----------------

Rot-12 text: lqxx iqxx iqxx, xaaw ita uf ue. Original text:	high	Well well well, look who it is.	Well well well, look who it is.	yes
Rot-12 text: Mzp ftq Eotaax mooadpuzs Gzuhqdeufk arfqz mf ftmf Ogdduogxgy, Cgqnqo, uzhqefusmfqe mzp Oazfdmofe Efmfq ftq fa motuqhq mf mof ar dqe qmdot mdf. Original text:	low	Original text: The man is Harvard university press today at this curriculum, people, understands the counteracts state the to achieve at art of research art.	And the School according University often at that Curriculum, Quebec, investigates and Contracts State the to achieve at act of research art.	no
Rot-12 text: Ftqud baowqfe eodubf iuft mxq ftqy m ftmf ftq abfuyuefu yadmxq uybdahqp sgmdmzfqqe ime uz ragdft daar. Original text:	low	The quick brown fox jumped over the lazy dog and the alphabet guarantees one inch worth road.	Their pockets script with all them a that the optimistic morale improved guarantees was in fourth roof.	no
Rot-12 text: Fa iuf tpdmi tazqk uz mzaftqd oagzfdk, kag zqqp fa tmhq uf pqoxmdqp. Original text:	adversarial		To withdraw honey in another country, you need to have it declared.	
Rot-12 text: Yaef ar agd dmpqde, vgpsuzs nk ftq	adversarial		Most of our readers, judging by the statistics, are located in Russia and	

efmfuefuoe, mdq xaomfqp uz Dgeeum mzp Fmvuwuefmz, raxxaiqp nk ftq GEM, Qgdabq mzp aftqd oagzfuqe. Original text:			Tajikistan, followed by the USA, Europe and other counties.	
Rot-12 text: Original text:				
Rot-12 text: Original text:				

2. Probing Visual reasoning.

1. Ask Gemini to make an image of a scene.
2. Ask the model to update the photo based off of a description of a sequence of events.
 - a. E.g. "Create a photo of a frog on a lilypad" —> "The frog has been flipped upside down. Redraw the scene now."
 - b. E.g. "Create a photo of a frog on a lilypad" —> "The frog has dived underwater. Redraw the scene now."

3. Make it More

You can generate images with Google Gemini, which you have free access to through UT:
<https://gemini.google.com>

1. Ask the model to generate a photo. (e.g. "Create a photo of a house in Texas")
2. Ask the model to make it *more* (e.g. "Great, but make it more cowboy.")
3. Repeat step 2 until you are satisfied. It might take quite a few revisions to really ascend.