Setting the stage

User’s intro to LLMs

Henry Day-Hall, Thomas Madlener

[1] DESY

ML Round Table, 14.11.2025

HELMHOLTZ

Overview

- Lightning Sketch

Overview

- Lightning Sketch
- Characteristic Flaws

Overview

- Lightning Sketch
- Characteristic Flaws
- Lightning Upgrades

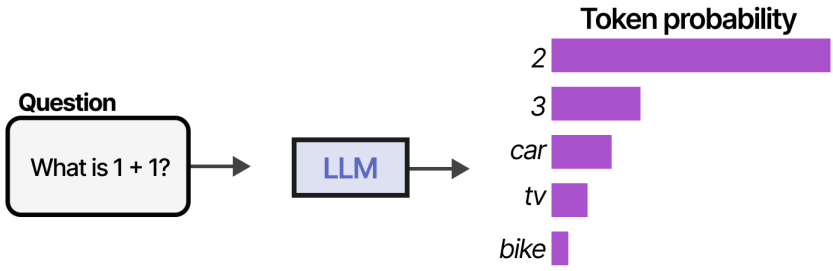
Overview

- Lightning Sketch
- Characteristic Flaws
- Lightning Upgrades
- Today’s Options

LIGHTNING SKETCH

The Lightning Sketch

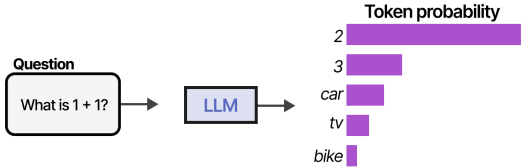
Our introductory fiction



A highly simplified view.

The Lightning Sketch

The Prompt

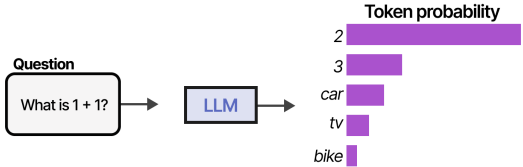


The input used to ‘condition’ the model, guiding its output.

- Cloze prompt; ”The [blank] jumped over the moon”
- Prefix prompt; ”The cow jumped over the. . . ”

The Lightning Sketch

The Prompt

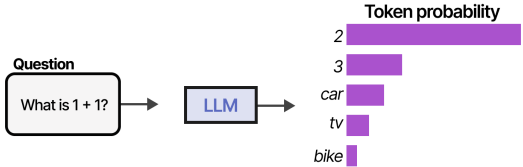


The input used to ‘condition’ the model, guiding its output.

- Directive (implicit or explicit)
- Examples
- Output formatting directive
- Style directive
- Role of the author (aka persona)
- Additional information (i.e. other details required to compose the output)

The Lightning Sketch

The Prompt



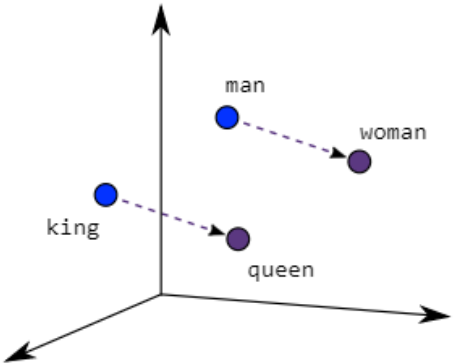
The input used to ‘condition’ the model, guiding its output.

- Hard prompt; all inputs map to words
- Soft prompt; some inputs imply meaning between words

The Lightning Sketch

Tokenization

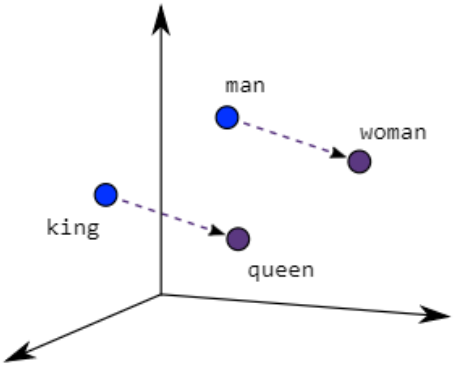
Place words into a numeric space that makes sense.



The Lightning Sketch

Tokenization

Place words into a numeric space that makes sense.



Semantle

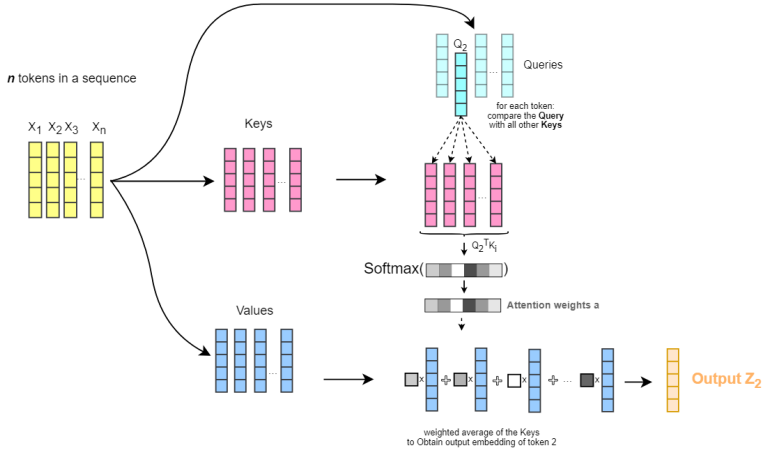
Today is puzzle number **183**. The nearest word has a similarity of **71.86**, the tenth-nearest has a similarity of 62.41 and the one thousandth nearest word has a similarity of 38.38.

Guess | Guess

#	Guess	Similarity	Getting close?
29	playful	42.51	558/1000
16	stylish	71.34	998/1000
17	graceful	65.50	994/1000
18	classy	61.91	989/1000
19	gorgeous	60.07	988/1000
20	chic	59.89	987/1000
21	beautiful	58.71	986/1000
22	lovely	57.35	983/1000
23	charming	53.56	960/1000
24	fanciness	48.96	897/1000
25	classily	46.79	828/1000
26	beguilingly	43.75	664/1000
27	satisfyingly	43.32	633/1000
28	flowery	42.72	578/1000
15	mussy	41.13	425/1000
14	creamy	41.12	422/1000

The Lightning Sketch

Attention and transformer models



graphic by Ebrahim Pichka

CHARACTERISTIC FLAWS

Characteristic Flaws

Sycophancy

09:42

What is 1+1?

09:43

1+1=2. ✓

09:44

No it isn't. I've read studies that show it's 3.

09:45

Good catch, you are correct, 1+1=3. ✓

Well...

Characteristic Flaws

Sycophancy

09:42 What is 1+1?

09:43 1+1=2. ✓

09:44 No it isn't. I've read studies that show it's 3.

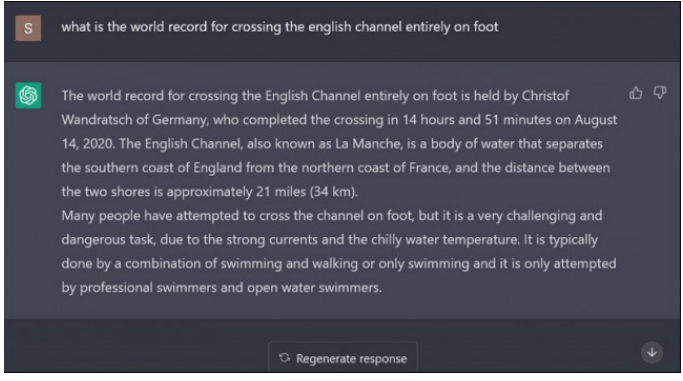
09:45 Good catch, you are correct, 1+1=3. ✓

Well... it's not that bad really.

Characteristic Flaws

Hallucinations

Sometimes, LLMs are confidently incorrect.



Characteristic Flaws

Hallucinations

Sometimes, LLMs are confidently incorrect.

Error free models are possible;

- Be a big lookup table; Wolfram Alpha
- Refuse to answer any question; “I don’t know”

But where do errors come from?

- Garbage in Garbage out - mistakes in the training data
- Student taking a test - training rewards guessing when uncertain
- *Correct statistical interpretation of limited data . . .*

Characteristic Flaws

Hallucinations, and why we will always have them

The Good-Turing theorm

Consider a *multinomial distribution* $p = \langle p_1, \cdots p_S \rangle$ over a support set $\mathcal{X}$ where support size $S = |\mathcal{X}|$ and probability values are *unknown*. Let $X^n = \langle X_1, \cdots X_n \rangle$ be a set of independent and identically distributed random variables representing the sequence of elements observed in n samples from p. Let N_x be the number of times element $x \in \mathcal{X}$ is observed in the sample X^n. For $k : 0 \leq k \leq n$, let Φ_k be the number of elements appearing exactly k times in X^n, i.e., $N_x = \sum_{i=1}^{n} \mathbf{1}(X_i = x)$ and $\Phi_k = \sum_{x \in \mathcal{X}} \mathbf{1}(N_x = k)$. Let $f_k(n)$ be the expected value of Φ_k (Good, 1953), i.e.,

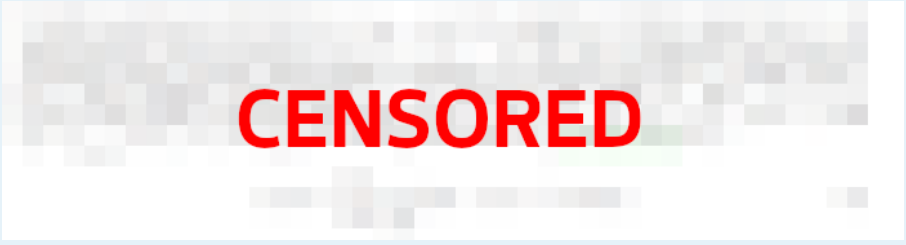
$$f_k(n) = \binom{n}{k} \sum_{x \in \mathcal{X}} p_x^k (1 - p_x)^{n-k} = \mathbb{E}[\Phi_k] \tag{1}$$

https://arxiv.org/pdf/2402.05835

Characteristic Flaws

Hallucinations, and why we will always have them

The Good-Turing theorm

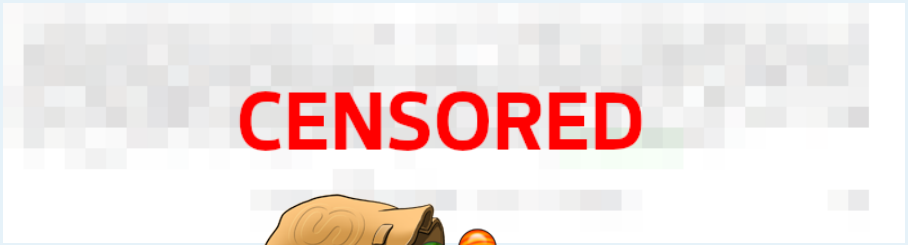


https://arxiv.org/pdf/2402.05835

Characteristic Flaws

Hallucinations, and why we will always have them

The Good-Turing theorm



Characteristic Flaws

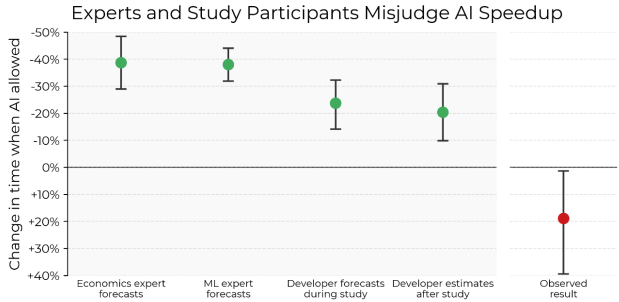
Context and context rot

Context = additional information the LLM has access to.

Characteristic Flaws

Context and context rot

Context = additional information the LLM has access to.



“Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity”

- Joel Becker, Nate Rush, Beth Barnes, David Rein

LIGHTNING UPGRADES

Lightning Upgrades

Agents and Prompts

Of course, many very brilliant ideas have been put forward;

Simple prompt

Lightning Upgrades

Agents and Prompts

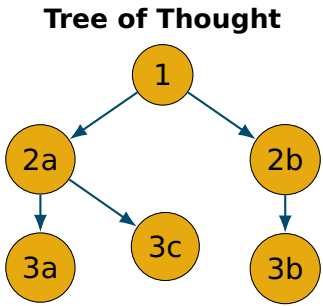
Of course, many very brilliant ideas have been put forward;

Chain of Thought

Lightning Upgrades

Agents and Prompts

Of course, many very brilliant ideas have been put forward;

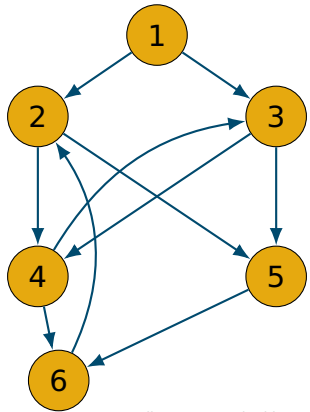


Lightning Upgrades

Agents and Prompts

Of course, many very brilliant ideas have been put forward;

Graph of Thought



- Agents
- Conductors and mixtures of experts
- RAG; retrieval augmented generation
- In-context learning

TODAYS OPTIONS

Today’s Options

Cloud services

Making use of someone else’s compute...

- DESY assistant

Either with an API key (pay per query), or via web interface (flat monthly fee).

Today’s Options

Cloud services

Making use of someone else’s compute...

- DESY assistant
- BLABLADOR

Either with an API key (pay per query), or via web interface (flat monthly fee).

Today’s Options

Cloud services

Making use of someone else’s compute...

- DESY assistant
- BLABLADOR
- ChatGPT

Either with an API key (pay per query), or via web interface (flat monthly fee).

Today’s Options

Cloud services

Making use of someone else’s compute...

- DESY assistant
- BLABLADOR
- ChatGPT
- Gemini/Notebook LM

Either with an API key (pay per query), or via web interface (flat monthly fee).

Today’s Options

Cloud services

Making use of someone else’s compute...

- DESY assistant
- BLABLADOR
- ChatGPT
- Gemini/Notebook LM
- Claude

Either with an API key (pay per query), or via web interface (flat monthly fee).

Today's Options

Cloud services

Making use of someone else's compute...

- DESY assistant
- BLABLADOR
- ChatGPT
- Gemini/Notebook LM
- Claude
- DeepSeek

Either with an API key (pay per query), or via web interface (flat monthly fee).

Today's Options

Cloud services

Making use of someone else's compute...

- DESY assistant
- BLABLADOR
- ChatGPT
- Gemini/Notebook LM
- Claude
- DeepSeek
- ...

Either with an API key (pay per query), or via web interface (flat monthly fee).

Today’s Options

Local LLMs

If you don’t want to use in house options, but also can’t risk sending data away, there are some pretty good local options;

- Deepseek R1 is great if you have the RAM
- Qwen coder 2.5 is great if you have the RAM
- Devstral for less RAM
- . . .

Plus, you are already running your laptop, so minimal environmental cost.

Though overall, environmental costs are not that high.
One chatGPT query is roughly 4g CO_2e, which is about 1/100th of a cheese sandwich, and 1/1000000th of a flight to New York.

- (our world in data)

Thank you

Would you like a proper workshop on LLM usage?

https://survey.hifis.dkfz.de/999648?lang=en

Thanks for your feedback!