

1.1 AI是如何工作的

(以大语言模型为例)



组词（在空格内添加一个字）

喝

完成句子

他看见小猫正在 喝

他看见李白正在 喝

“过去的语言经验” 结合 “上下文语境”

让AI完成句子

提示词：

请在括号里添加一个字，使句子完整、通顺。

要求按概率从大到小排列可能的答案，并附概率百分值。

句子如下：他看见李白在喝（ ）。



补全句子：他看见李白在喝（ ）

DeepSeek



好的，我们先分析这个句子“他看见李白在喝（ ）”。

1. 可能的补全思路

- 李白是唐代诗人，历史记载和文学形象中，他常与“酒”联系在一起，所以“喝酒”是最自然的搭配。
- 其他常见饮料如“喝茶”“喝水”“喝汤”等也符合语法，但语境上不如“酒”贴合李白。
- 现代口语或幽默用法也可能填“喝醋”“喝药”等，但概率较低。

2. 概率估算

- **酒**：文学常识 + 高频联想，概率最高。
- **茶**：古代也常见，但李白以嗜酒闻名，所以“茶”次之。
- **水**：日常生活常见，但缺乏语境特色。
- **汤**：可能，但概率低于水。
- **其他**：如“药”“粥”“奶”等，概率很低。

3. 按概率从大到小排列

答案	概率 (约)	理由
酒	75%	李白与酒的强关联
茶	12%	古代常见饮品
水	8%	通用，但无特色
汤	3%	吃饭时可能喝汤
药	1%	可能但需特殊语境
其他	1%	醋、粥、咖啡等

不同上下文的结果对比

他看见 <u>李白</u> 在喝 ()		他看见 <u>武大郎</u> 在喝 ()	
酒	75%	药	70%
茶	12%	酒	15%
水	8%	茶	8%
汤	3%	水	5%
药	1%	汤	2%

补全句子：他看见武大郎在喝（ ）

DeepSeek



好的，我们分析一下“他看见武大郎在喝（ ）”这个句子。

1. 背景与常见联想

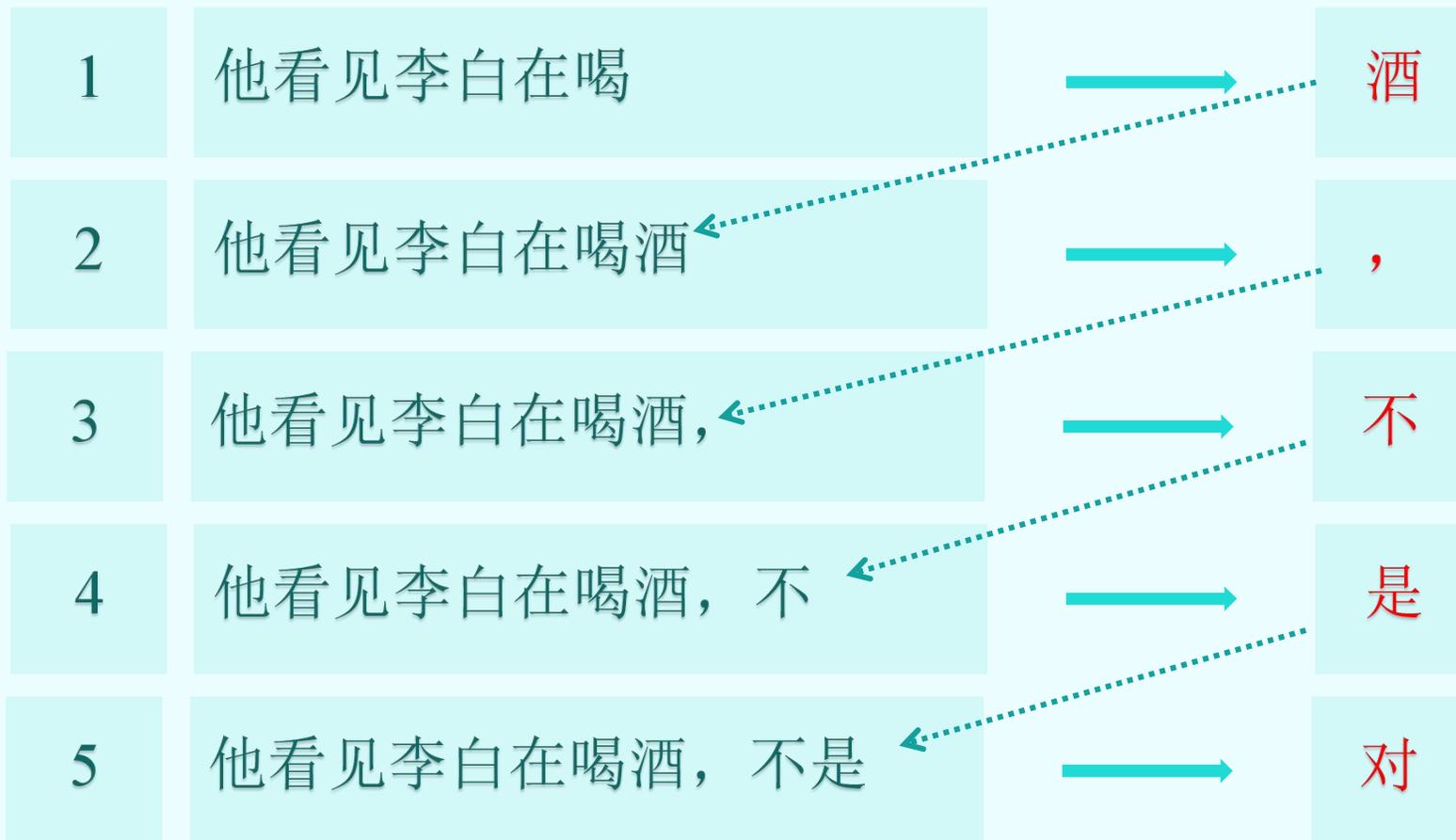
- 武大郎是《水浒传》中人物，职业是卖**炊饼**，但在文学或影视作品中，他常被描绘为在家吃饭、喝药的场景。
- 尤其“喝药”是一个强烈关联——因为他是被潘金莲在药里下毒而死，所以“喝药”有剧情上的高联想度。
- 日常也可能“喝酒”“喝水”“喝茶”，但“喝酒”更多与武松相关，武大郎形象偏老实、居家。

2. 概率估算

- **药**：文学典故（被毒死）→ 最高概率。
- **酒**：可能（比如武松回家一起喝酒），次之。
- **水 / 茶**：日常合理，但无特殊关联。
- **汤**：可能，但概率低。

AI写故事

提示词：根据我提供的文字续写故事，200字左右。他看见李白在喝……



AI写故事

他看见李白在喝酒，不是对月小酌，而是抱着酒坛，在江边泥泞里痛饮。青衫湿了大半，狂态却愈显。他不敢惊动，只远远站着。

李白忽然转头，目光如电：“小友，来！”他鬼使神差地接过酒坛，灌了一口，辛辣直冲肺腑。

“这酒如何？”李白笑问。

“苦。”

李白仰天大笑：“这是人间味！”说罢将酒坛掷入江中，朗声吟道：“大道如青天，我独不得出！”

他怔怔看着，只见李白踉跄起身，踏着江波而去，每一步都踩碎满江月影，仿佛要把这不得志的人间，走成一场醉后的传奇。

大模型在实际工作中的token

Tokenizer

Learn about language model tokenization

OpenAI's large language models process text using **tokens**, which are common sequences of characters found in a set of text. The models learn to understand the statistical relationships between these tokens, and excel at producing the next token in a sequence of tokens. [Learn more.](#)

You can use the tool below to understand how a piece of text might be tokenized by a language model, and the total count of tokens in that piece of text.

GPT-5.x & O1/3 GPT-4 & GPT-3.5 (legacy) GPT-3 (legacy)

我在黑板上画了一只熊。

Clear Show example

Tokens	Characters
10	11

我在黑板上画了一只熊。

Text Token IDs

我 在 黑 板 上 画 了 一 只
熊
[7522, 2178, 16599, 28813, 4286, 22126, 59176, 18637, 100391, 788]

Tokenizer

Learn about language model tokenization

OpenAI's large language models process text using **tokens**, which are common sequences of characters found in a set of text. The models learn to understand the statistical relationships between these tokens, and excel at producing the next token in a sequence of tokens. [Learn more.](#)

You can use the tool below to understand how a piece of text might be tokenized by a language model, and the total count of tokens in that piece of text.

GPT-5.x & O1/3 GPT-4 & GPT-3.5 (legacy) GPT-3 (legacy)

I drew a bear on the blackboard.

Clear Show example

Tokens	Characters
9	32

I drew a bear on the blackboard.

Text Token IDs

I drew a bear on the blackboard.
[40, 48614, 261, 16387, 402, 290, 5960, 3935, 13]

GPT-5.x & O1/3 GPT-4 & GPT-3.5 (legacy) GPT-3 (legacy)

I drew a bear on the blackboard.

Clear Show example

Tokens	Characters
9	32

[40, 48614, 261, 16387, 402, 290, 5960, 3935, 13]

Text Token IDs

对于常见的英文文本，一个 token 通常对应大约 4 个字符。这大致相当于 $\frac{3}{4}$ 个单词（即 100 个 token \approx 75 个单词）。

AI大语言模型工作的动态演示



视频来源 (bilibili) [链接](#)