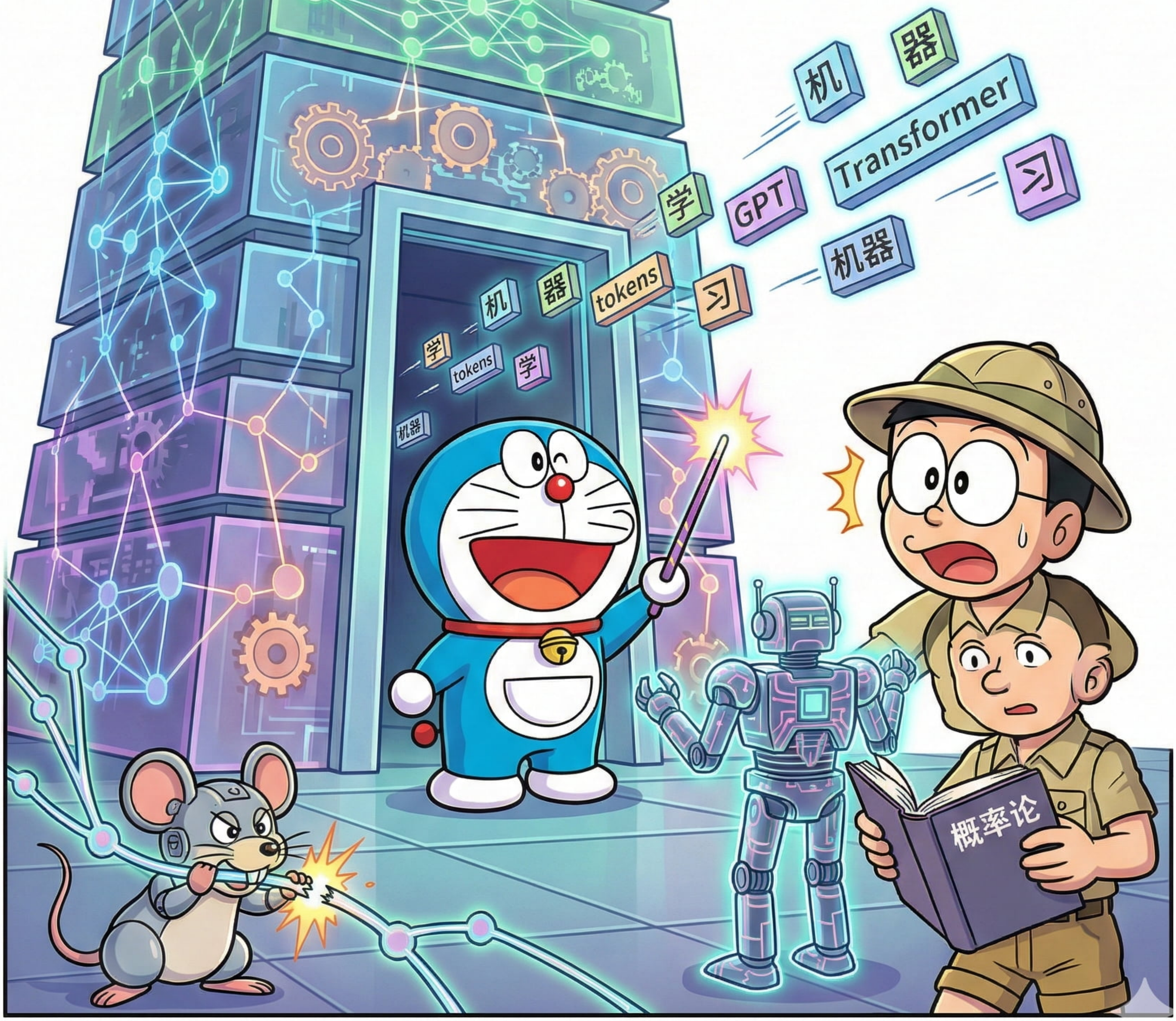
Ohmsha

漫画学习 GPT：大雄的自动写作大冒险

——从零看懂 Transformer 与 GPT-2 架构原理



深夜在万亿太的房间。

不行了！大脑一片空白！除了“我的梦想”四个字，我什么都写不出来！

我的梦想

真是的，大雄，现在才开始写吗？

哆啦A梦，有没有那种只要我起个头，就能自动帮我把剩下800字写完的道具？

拿你没办法。这是**“GPT-2 模型体验机”**！它是目前最先进的语言生成模型哦。

它真的懂怎么写作文吗？它有思想吗？

嗯……与其说它有有思想，不如说它是一个超级厉害的**“接龙高手”**。

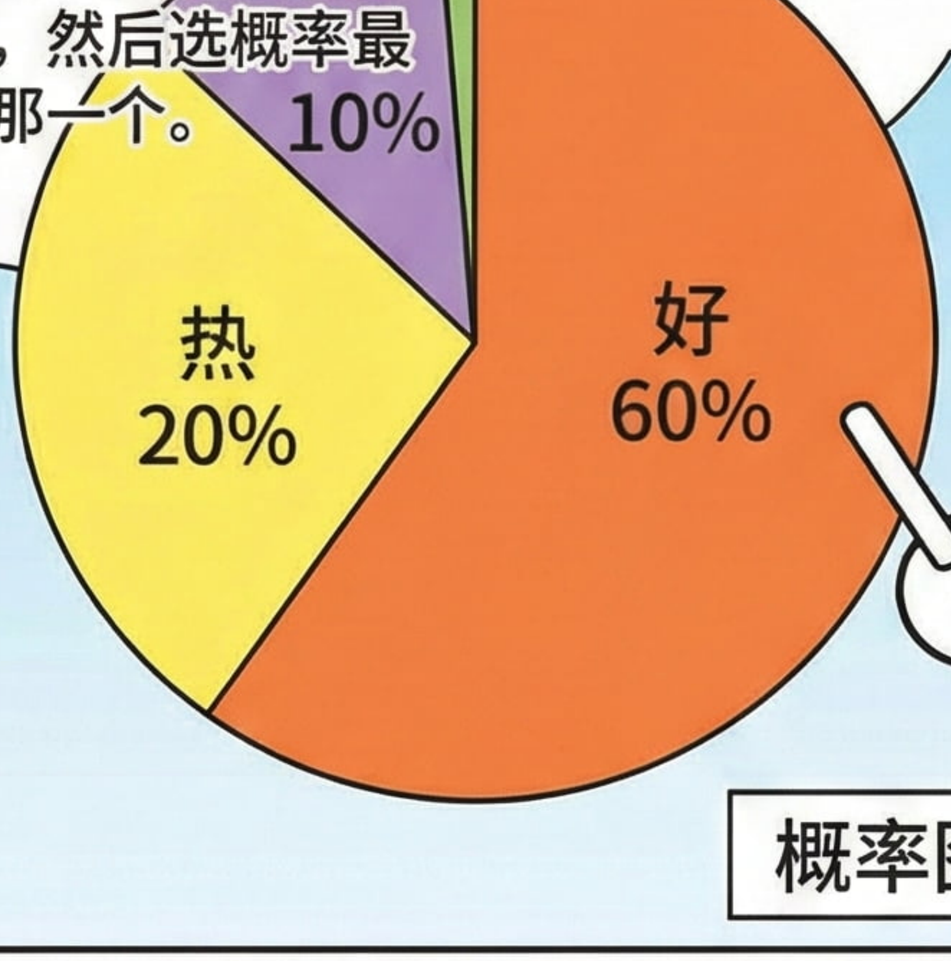
Ohmsha style, full color

今天天气真____

大雄，看到这句话，你本能地想接什么字？

“好”？
或者“坏”？
“不错”？

没错！GPT 的原理就是计算在当前这串文字后面，出现所有汉字的概率，然后选概率最高的那一个。



概率图

就这么简单？只是猜下一个字？

原理简单，但如果它读过互联网上所有的书，它的“猜测”就会变得无比精准，看起来就像有智慧一样。

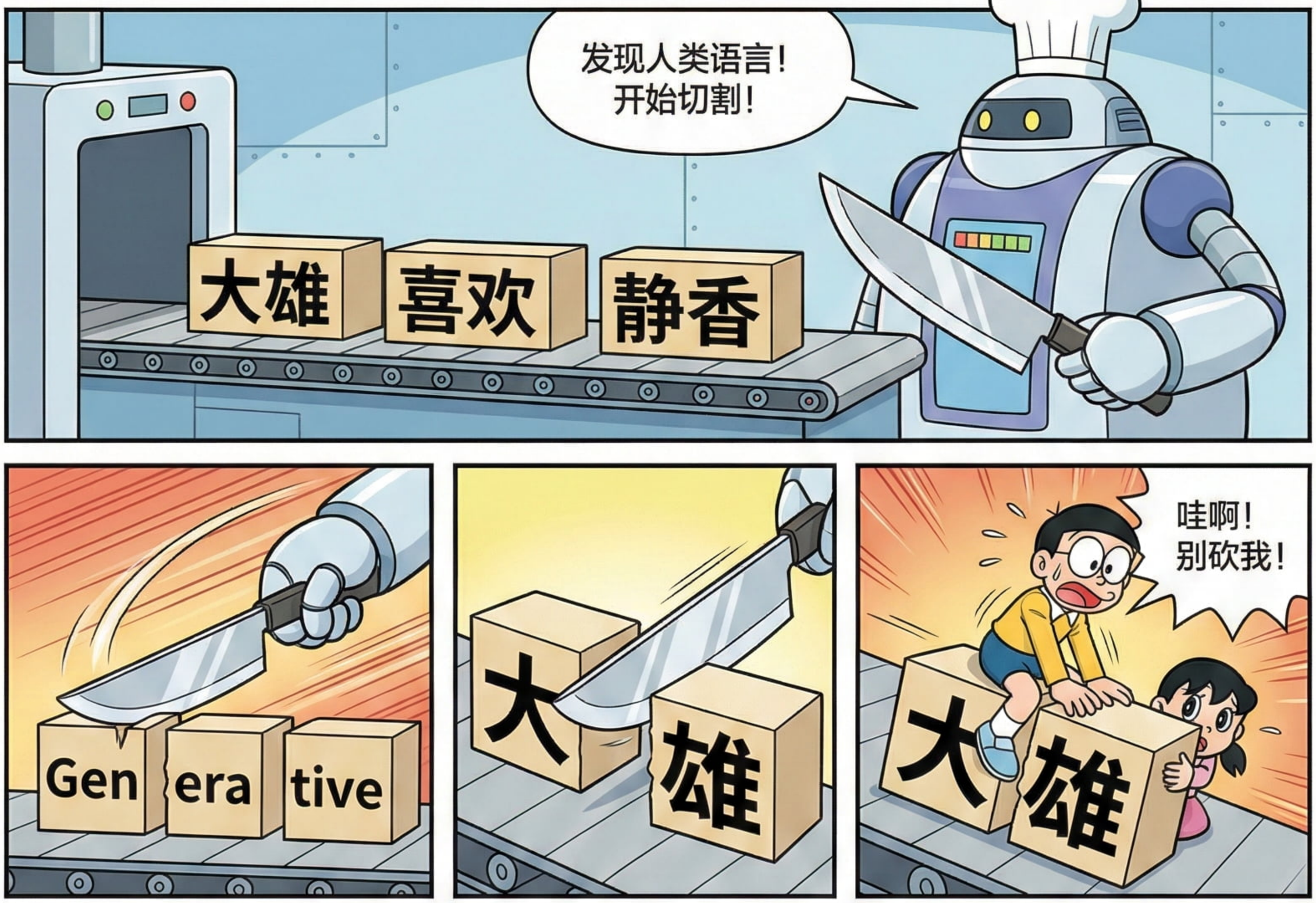
百闻不如一见。我们把这句“大雄喜欢静香”输入进去，然后亲自进到模型里看看它是怎么处理的吧！

咻——

欢迎来到 GPT 的微观世界！我们要开始一场数据流的冒险了！

大雄喜欢静香

哇，字变得好大！


发现人类语言！
开始切割！
大雄
喜欢
静香
Gen
era
tive
大
雄
哇啊！
别砍我！
大
雄
机器不认识汉字，只认识数字。
这一步叫 Tokenization。它把
文本变成机器能查阅的
字典索引。
大
[123]
雄
[456]
Dictionary
2 5
我现在是一
个数字了?
对，GPT-2 的字典里
大约有 50,000 个这样
的词元（Token）。
[123]

光有数字 ID 没用，机器不知道[123]代表什么意思。我们要去 Embedding 层领装备。

[123]

好重！这铠甲也太厚了吧！

[123]

这叫词向量。每一层玻璃代表一个特征。比如‘是否是生物’、‘是否是名词’、‘褒义还是贬义’。

词向量

层1：是否是生物?

层2：是否是名词?

层3：褒义/贬义?

苹果

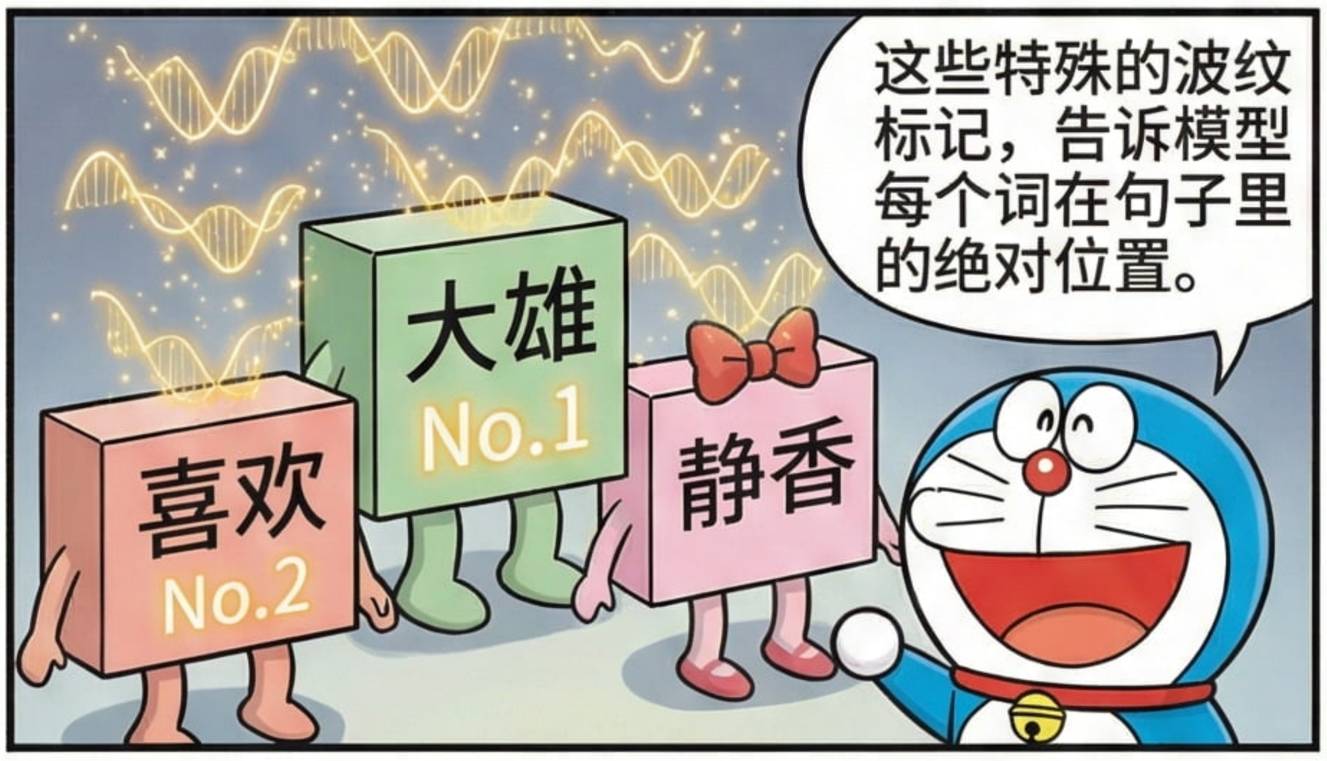
香蕉

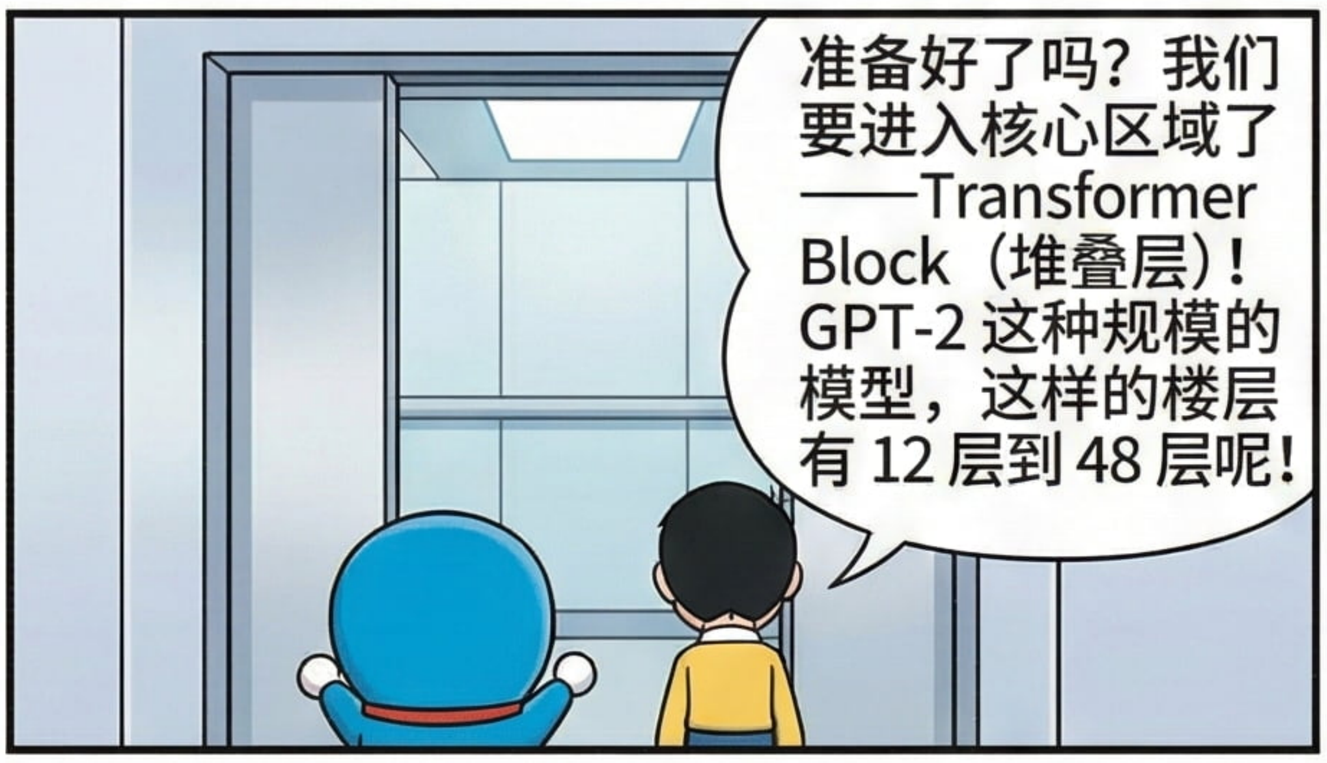
你看，意思相近的词，铠甲的颜色分布（向量方向）就很像。这就是机器理解语义的方式！

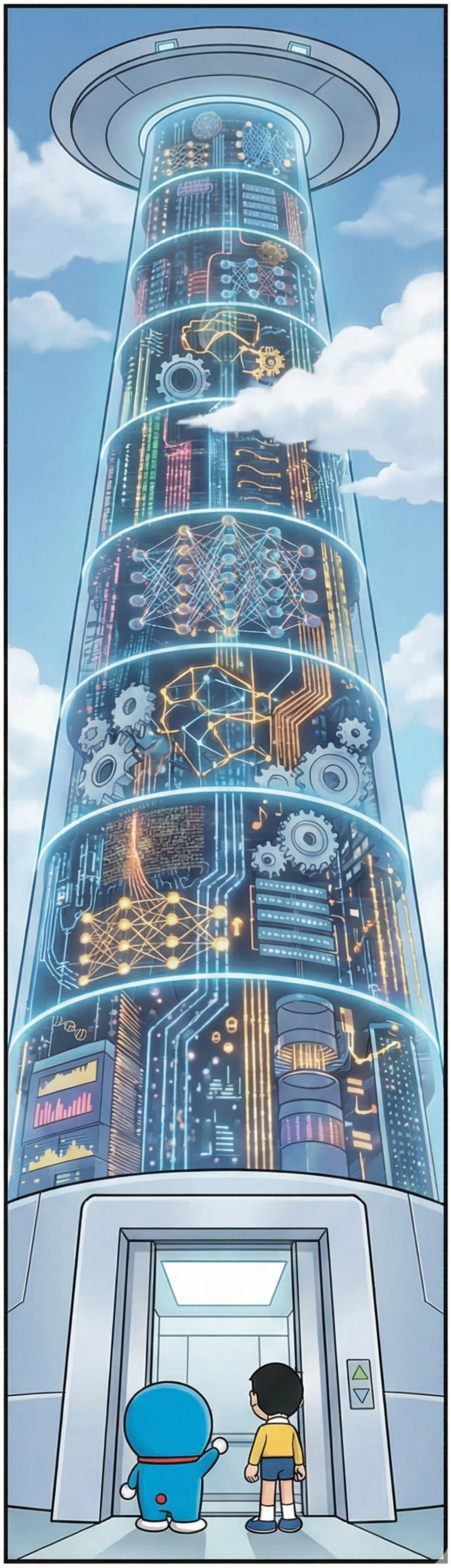
国王
王后
男人
女人
男人
香蕉
苹果
虽然我们看到的是铠甲，但在数学上，它们是高维空间里的点。
国王
王后
－男性特征
＋女性特征
男人
女人
国王－男人＋女人＝王后
通过计算距离，机器能学会推理。比如把“国王”减去“男性特征”加上了“女性特征”，就会自动飞到“王后”的位置。
那我的方块和静香的方块，距离近吗？
大雄
静香
嘿嘿，这就看语料库里的你们是不是经常一起出现啦。

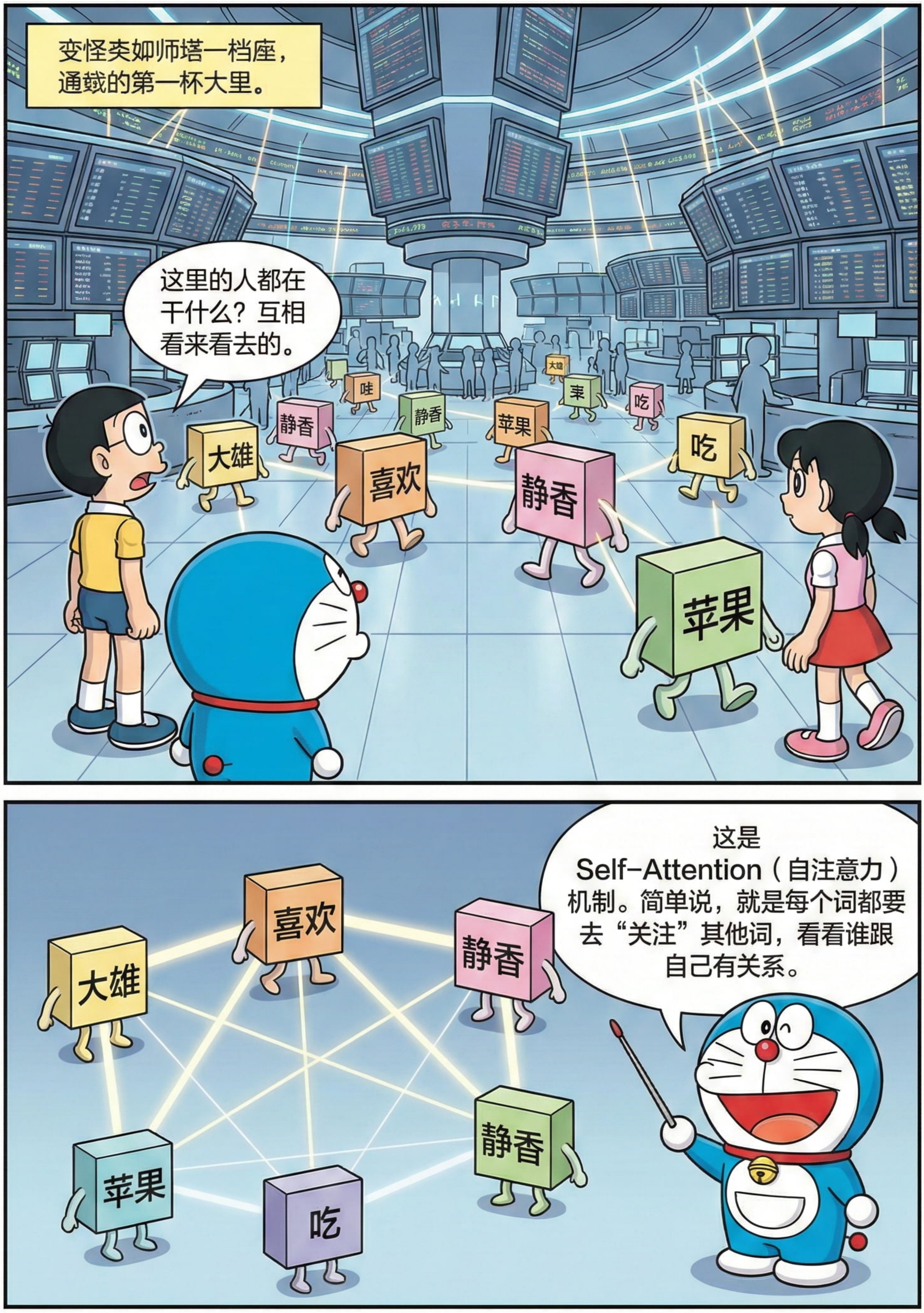

糟了！顺序乱了，
意思就变了！
喜欢
静香
大雄


Transformer 是一次性并行处理所有词的，它是“脸盲”，分不清先后。必须加上 Positional Encoding（位置编码）。

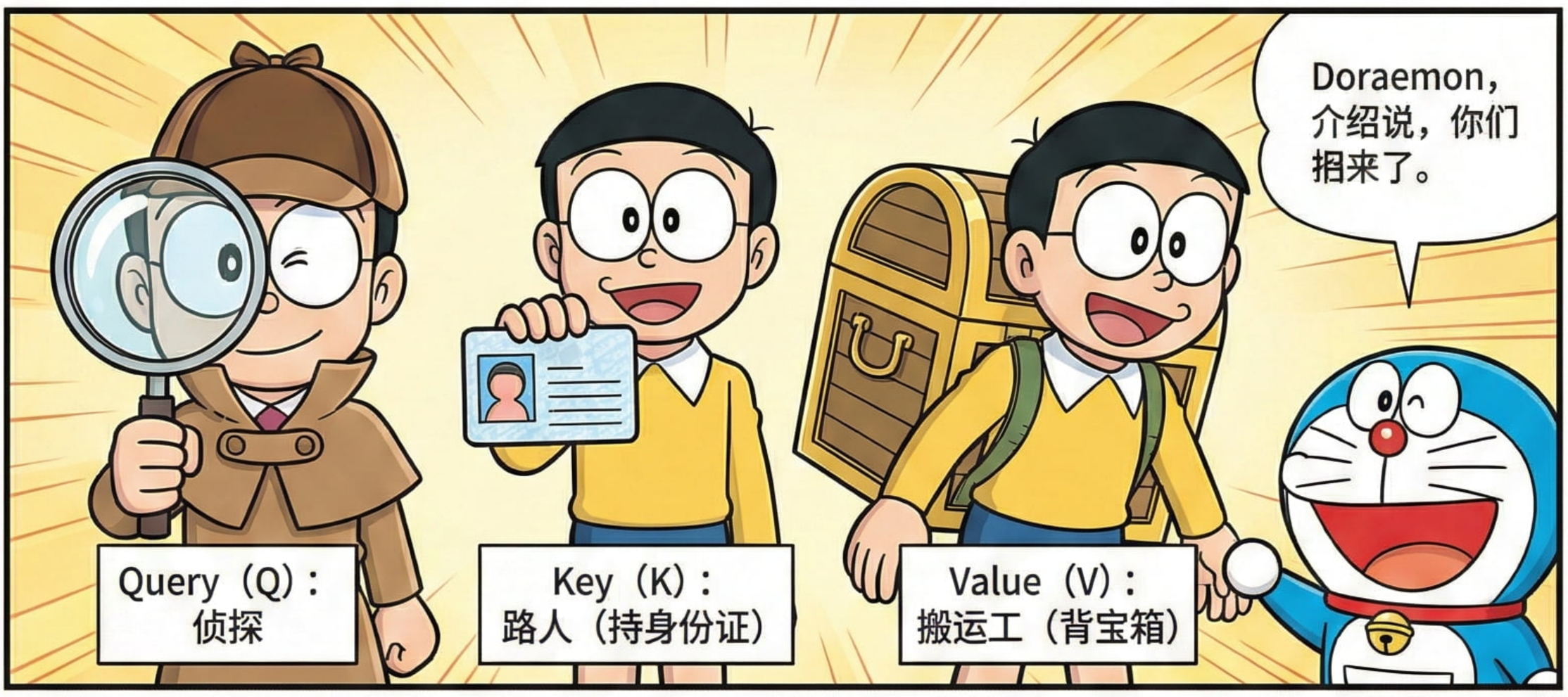

这些特殊的波纹标记，告诉模型每个词在句子里的绝对位置。
喜欢
No.2
大雄
No.1
静香


准备好了吗？我们要进入核心区域了——Transformer Block（堆叠层）！GPT-2 这种规模的模型，这样的楼层有 12 层到 48 层呢！




变怪类如师塔一档座，
通鼓的第一杯大里。
这里的人都在干什么？互相看来看去的。
大雄
静香
喜欢
哇
静香
苹果
吃
静香
苹果
大雄
喜欢
静香
苹果
吃
静香
这是Self-Attention（自注意力）机制。简单说，就是每个词都要去“关注”其他词，看看谁跟自己有关系。





$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Query 是去“查询”的，Key 是用来“匹配”的，Value 是真正包含信息的“宝藏”。

Q —— K

喜欢-Q
(Like-Q)

K-ID
大雄-K
(Nobita-K)

大雄-K
(Nobita-K)

“喜欢”想知道是谁喜欢？于是它的 Q 去扫描“大雄”的 K。

K-ID
大雄-K
(Nobita-K)

匹配度
98%！

K-ID
今天-K
(Today-K)

匹配度
5%。

喜欢-Q
喜欢-静香
喜欢-大雄
喜欢-静香
喜欢-静香
喜欢-大雄
喜欢-静香
喜欢-静香
喜欢-静香
喜欢-静香
喜欢-静香

1	1	0	1	0	0	0	0	0	0
0	18	11	2	2	1	1	0	0	0
2	3	15	12	3	2	4	0	1	0
0	0	1	36	25	2	3	0	0	0
0	2	3	2	2	58	7	0	0	0
0	1	0	0	2	1	17	29	3	1
0	0	2	0	1	0	9	70	65	5
				1	0	3	13	10	103

所有词的 Q 和 K 都进行了一轮交互，生成了一张复杂的“人际关系网”（Attention Score Matrix）。

正在“喜欢-Q”掉保中，网现在将亮的话的亮度，并将亮甚他匹配度成分。

喜欢-Q

匹配度越高，吸取的 V 信息就越多。

大雄-V

静香-V

今天-V

看！“喜欢”这个词变了！它现在不仅是“喜欢”，还包含了“大雄发出的”和“指向静香的”这些含义。

喜欢

这就是上下文融合！经过这一步，每个词都读懂了它在句子里的环境。

大雄

静香

喜欢

黄雄

今天

喜欢-Q
德傈师

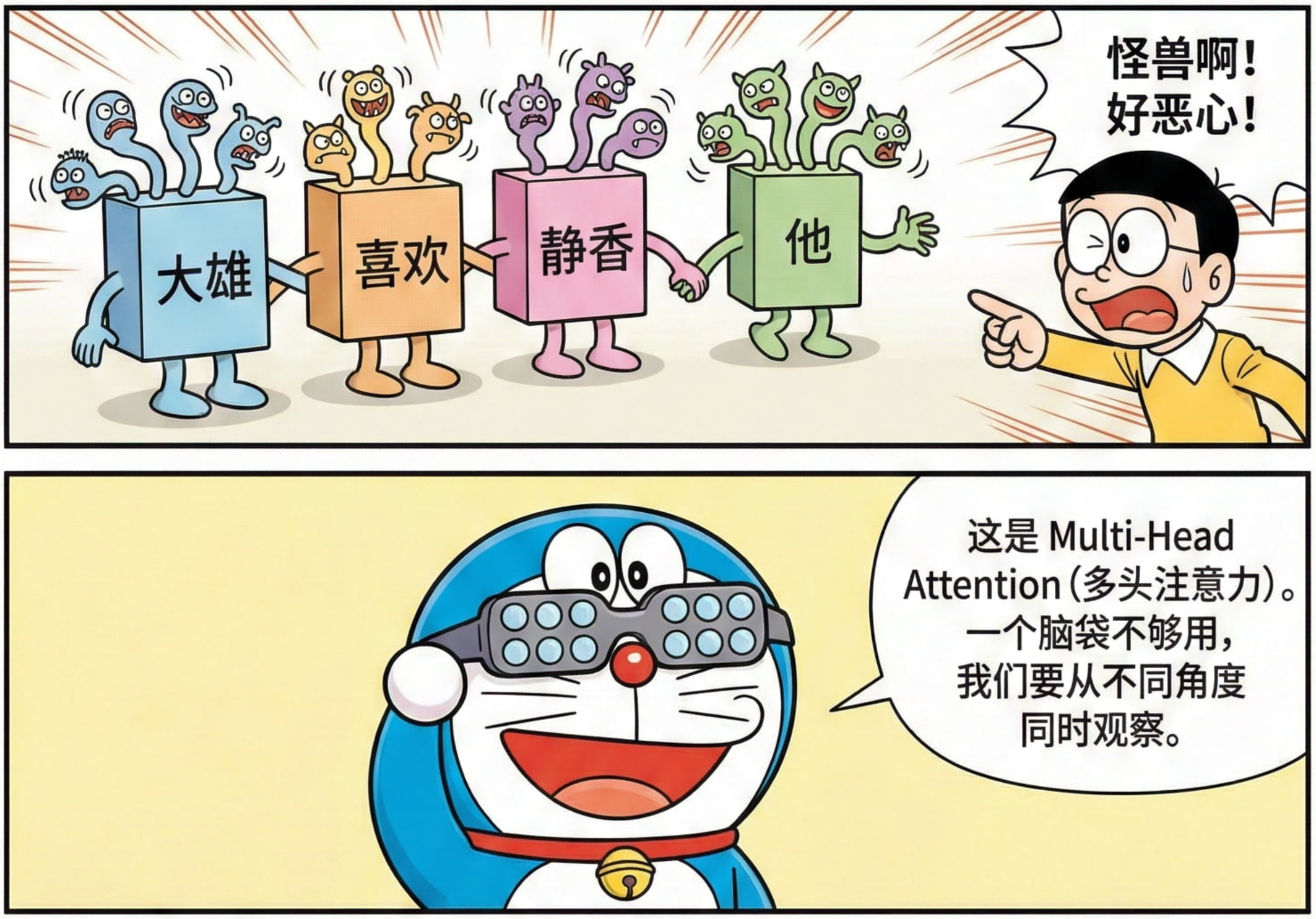
静香

喂！如果你现在就看到后面是“静香”，那不就是作弊吗？

正是！GPT 是单向模型（Decoder-only）。在预测未来时，它只能看历史。所以必须用 Mask（掩码）挡住后面的词。

香

BERT（双向）	GPT（单向）
今天天气真 MASK 好	今天天气真____
BERT 是做填空题（完形填空），可以看两头；	GPT 是写作文，只能往后写。


大雄
喜欢
静香
他
怪兽啊！
好恶心！
这是 Multi-Head Attention（多头注意力）。
一个脑袋不够用，
我们要从不同角度
同时观察。
大雄
喜欢
静香
（主）
（谓）
（宾）
头1
关注语法（主谓宾关系）。
他
大雄
头2
关注指代（“他”是指谁）。
大雄
喜欢
静香
头3
关注情感。
三个臭皮匠顶个诸葛亮，
多头机制让模型更聪明！
Concat

这是 Residual Connection（残差连接）。为了防止信息在层层传递中丢失，我们把原始信息直接加到处理后的信息上。

喜欢

喜欢

这是 Layer Norm（层归一化）。让数据分布更稳定，方便下一层处理。

大猫

猫猫

归一化整烫机

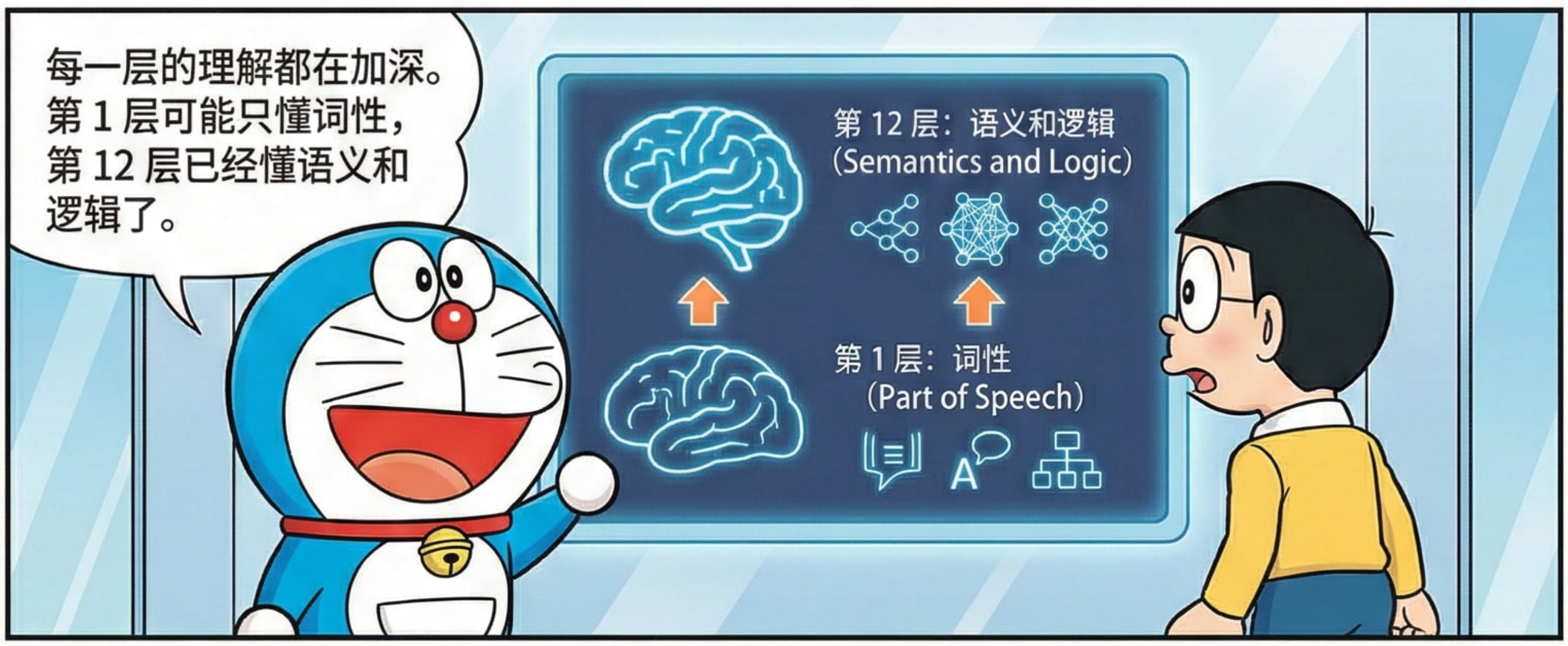
喜欢

喜欢

猫猫

大
大雄
喜欢
静香
Feed-Forward Network (FFN)
静香
注意力层只是在“收集信息”，这里才是“思考和消化”的地方。
通过非线性函数（GELU），模型在这里提取更高级的特征。
喜欢
喜欢
升维
（Expansion）
GELU
（非线性函数）
降维
（Contraction）

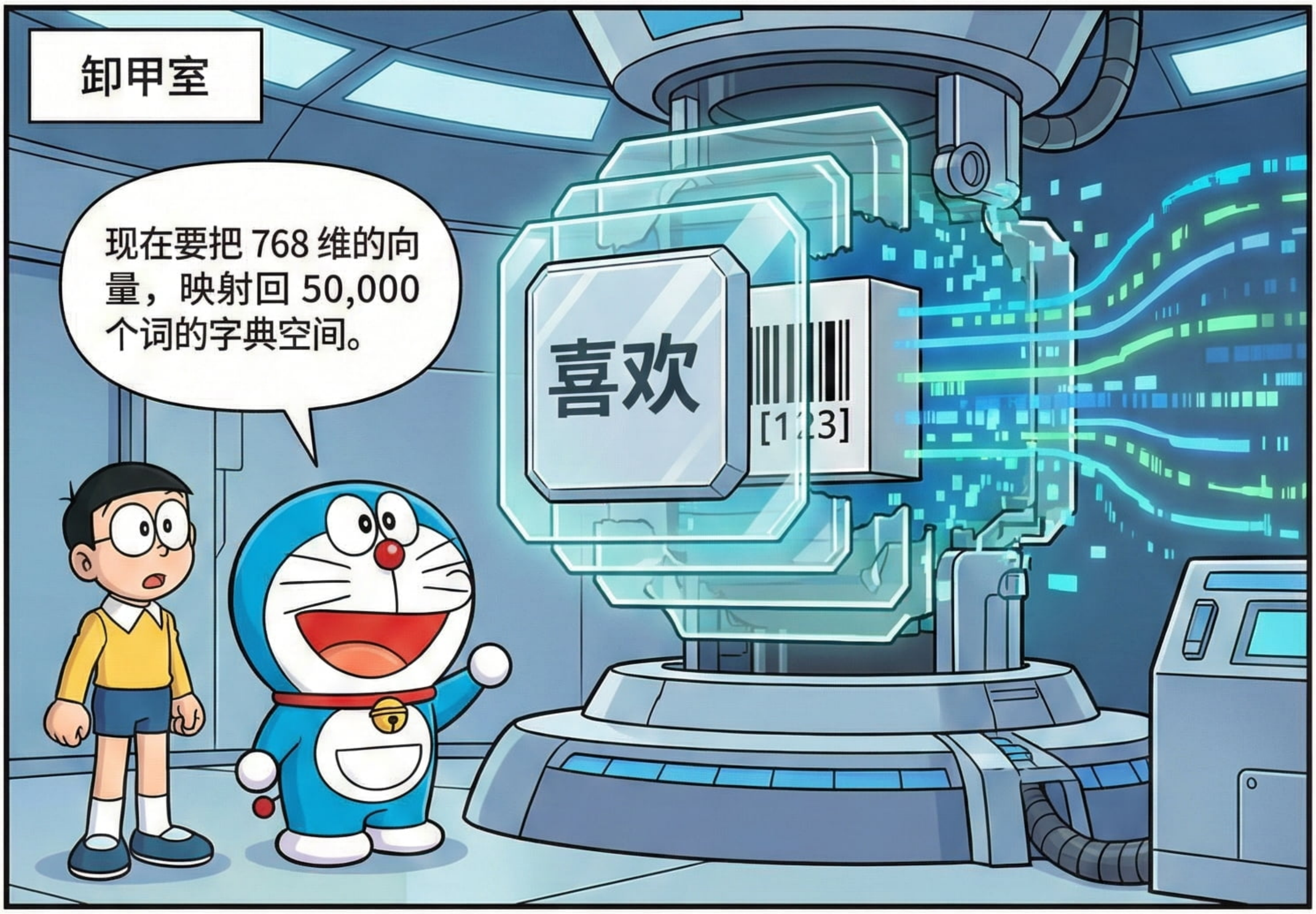

12F
11F
9F
8F
7F
4F
3F
2F
1F
Add&Norm
FFN
Attention
怎么又是这一套？
要重复多少次啊？


每一层的理解都在加深。
第 1 层可能只懂词性，
第 12 层已经懂语义和
逻辑了。
第 12 层：语义和逻辑
(Semantics and Logic)
第 1 层：词性
(Part of Speech)

12F - 最终层

喜欢

看，经过 12 层的洗礼，这个向量已经包含了推测下一个词所需的所有智慧。


卸甲室
现在要把 768 维的向量，映射回 50,000 个词的字典空间。
喜欢
[123]


牧陽 35.5
同禁 33.4
天雄 21.7
静香 12.5
静香 18.5
铜锣烧 15.2
大雄 9.8
考试 -5.1
学校 2.3
玩耍 12.1
Logits
蕾田燉 28.9
口醇 27.8
暗潮 16.6
静香 15.3
铜锣烧 14.2
玩耍 10.1
竟花
Logits
玩耍
大雄
学校 9.8
5.2
分数有高有低……“静香”的分数好像很高！

1.8 9.8 18.5 9.8 19.5 5.6 15.2 15.2 9.8 9.8 9.8 18.5

Softmax 层

这是 Softmax 层。
它把分数变成概率百分比，
所有词的概率加起来
正好是 100%。

静香
60%

铜锣烧
10%

胖虎
1%

一定会转到“静香”吧？

不一定哦！这就看我们怎么设置 Temperature（温度）。

低温（Low Temp）—保守

温度低时，指针死死吸在最大面积上（保守）；

静香 60%

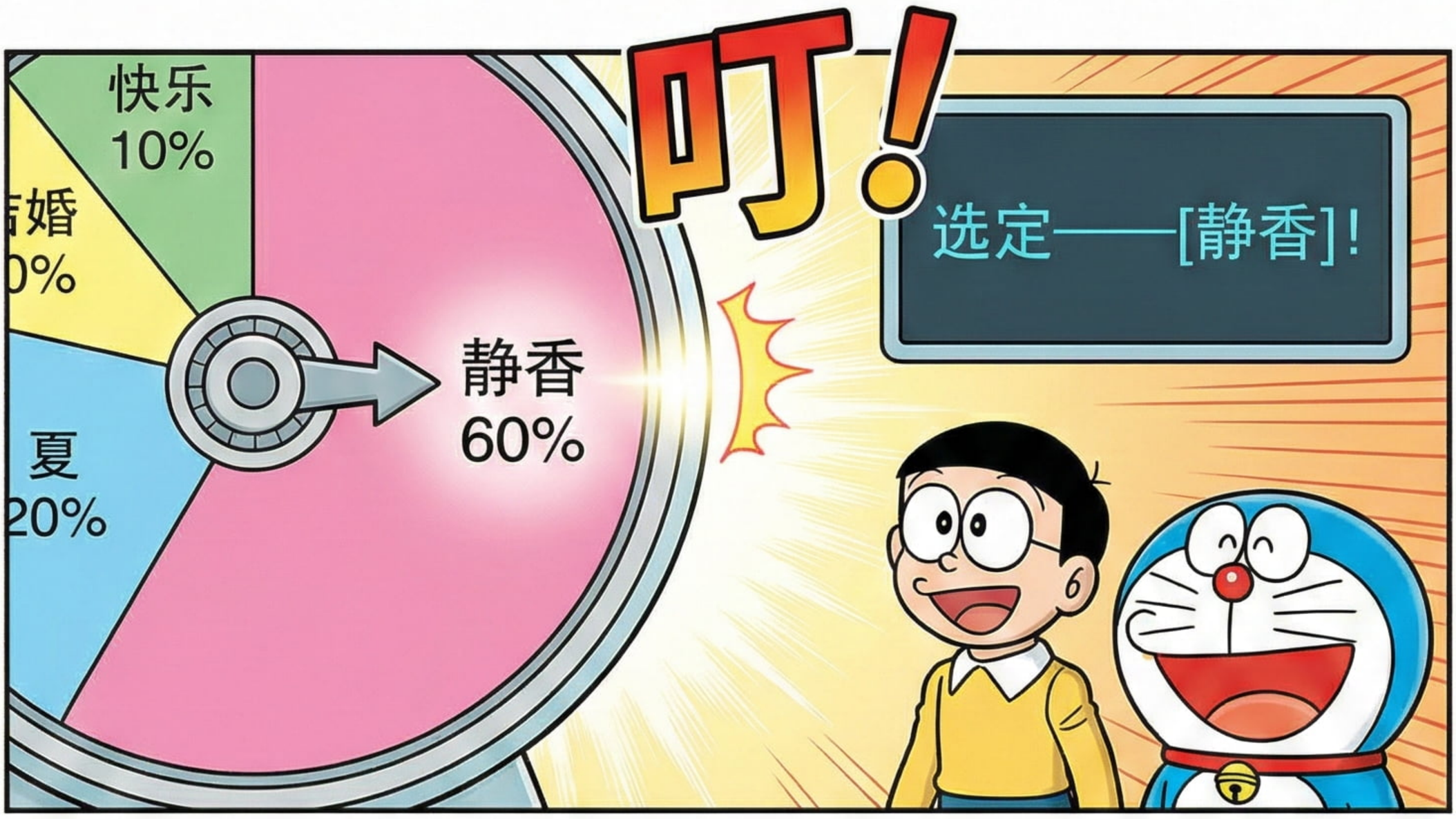
高温（High Temp）—创造力/胡说八道

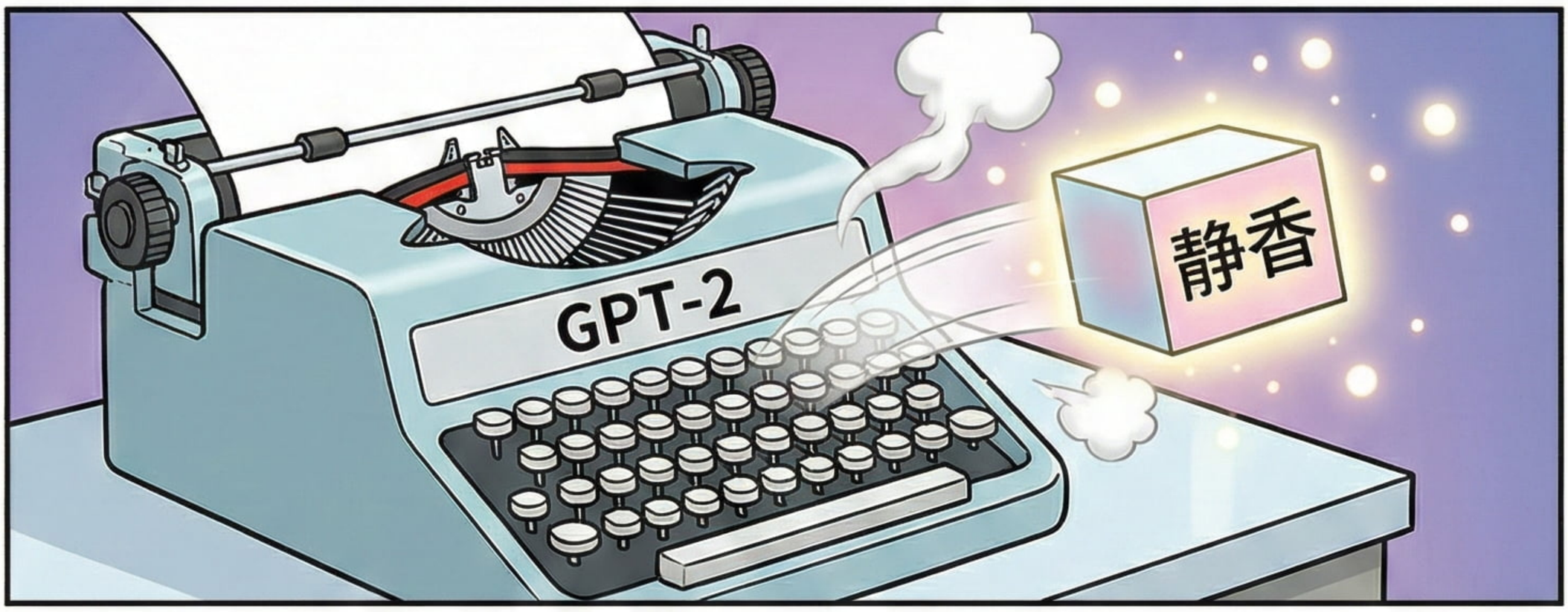
温度高时，指针乱跳（更有创造力，但也可能胡说八道）。

铜锣烧 10%

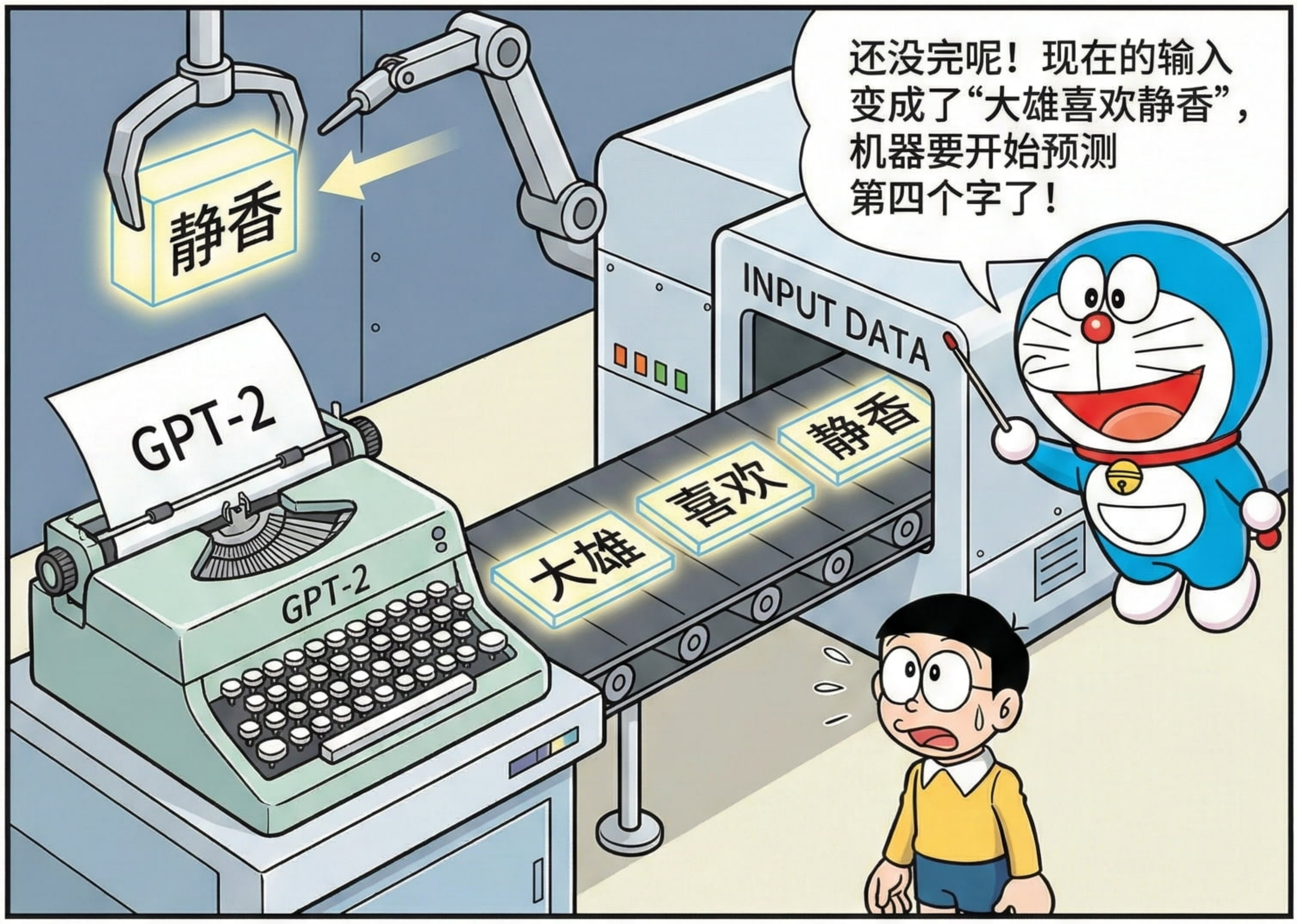
胖虎 1%

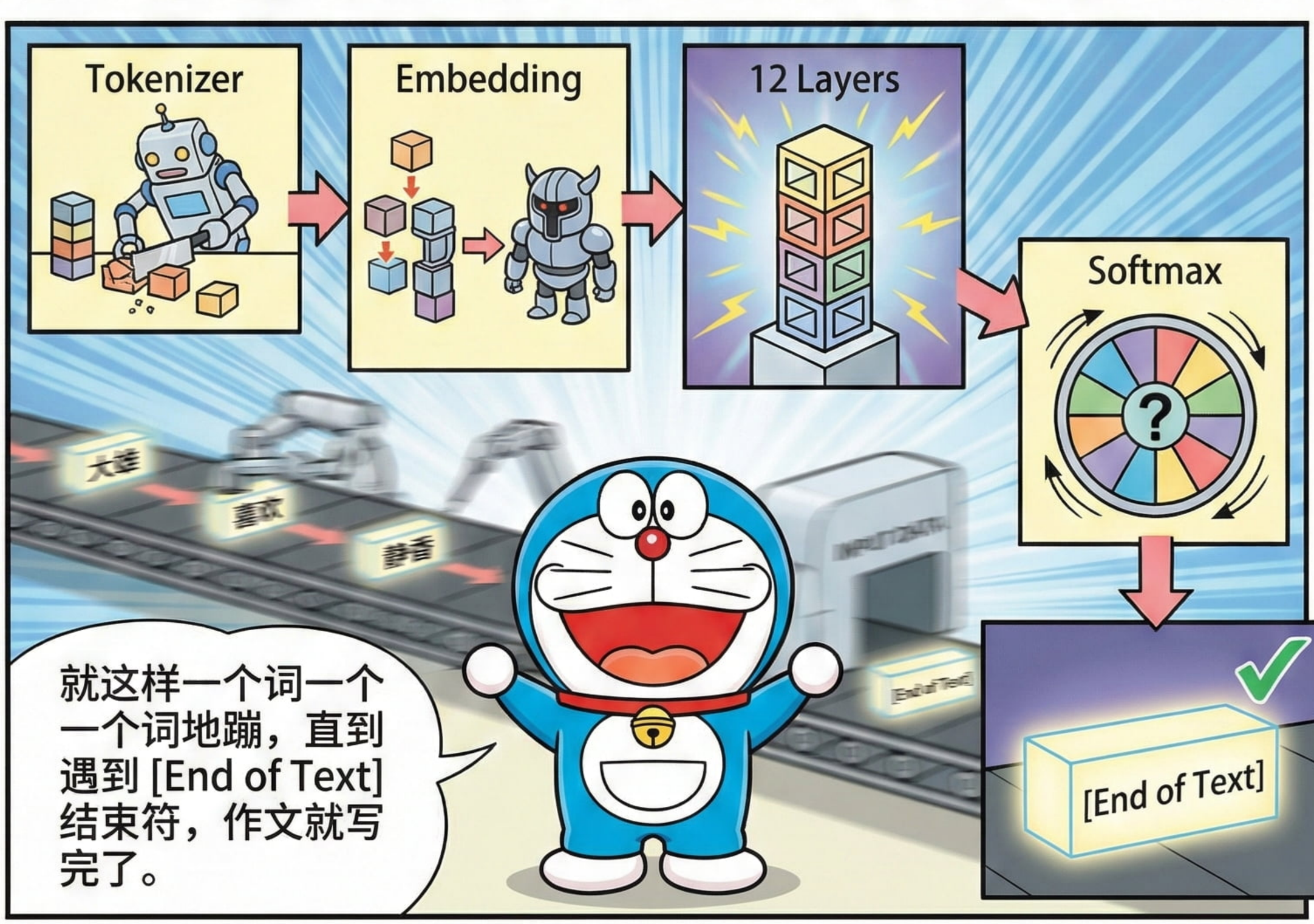
为了作文有点文采，我们稍微给一点随机性。


快乐
10%
结婚
0%
夏
20%
静香
60%
叮！
选定——[静香]！


GPT-2
静香


太好了！是「大雄喜欢静香」！
大雄
喜欢
静香
静香


静香
GPT-2
GPT-2
INPUT DATA
静香
喜欢
大雄
还没完呢！现在的输入变成了“大雄喜欢静香”，机器要开始预测第四个字了！


Tokenizer
Embedding
12 Layers
Softmax
?
[End of Text]
就这样一个词一个一个词地蹦，直到遇到 [End of Text] 结束符，作文就写完了。

后来在距太的房间，

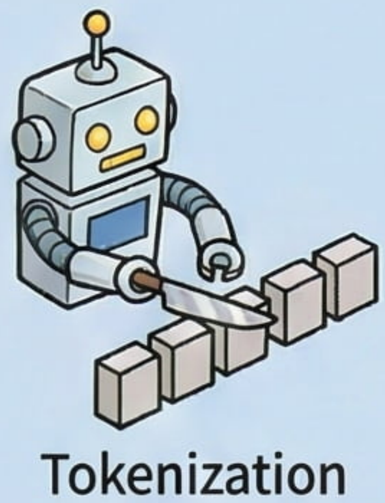
写好了！我看看……《我的梦想》：大雄喜欢静香，但是静香喜欢吃烤红薯，所以大雄要变成烤红薯……

这写的是什么乱七八糟的啊！

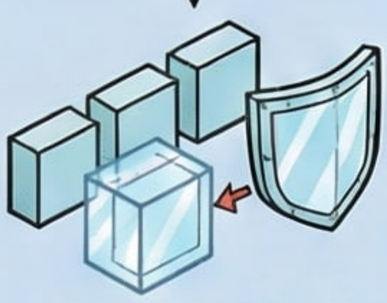
哈哈，可能是训练数据里有很多奇怪的小说吧。AI只是模仿概率，它并不真的懂逻辑哦。

知识点大总结（Cheat Sheet）

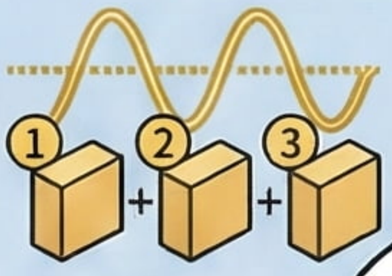
输入处理



Tokenization

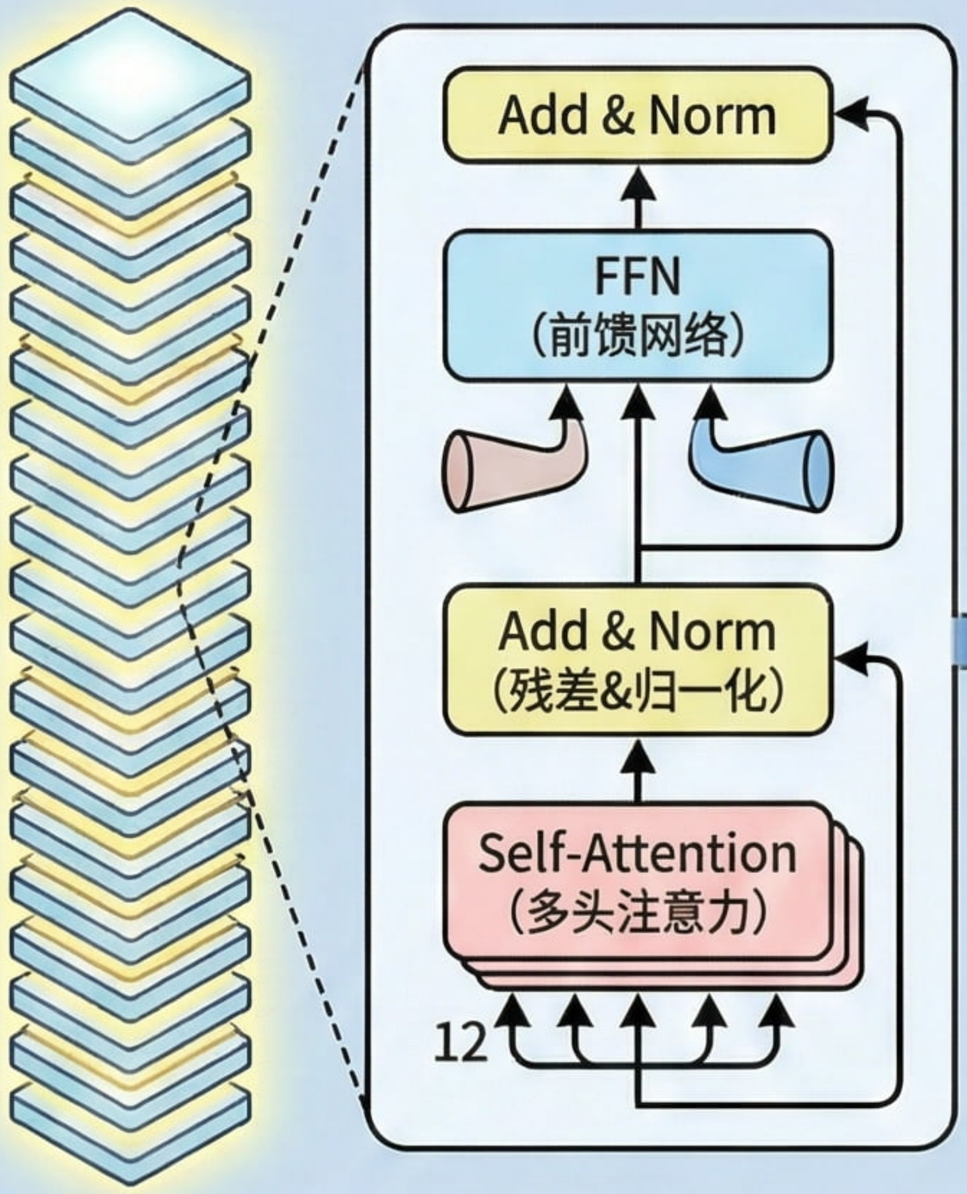


Embedding

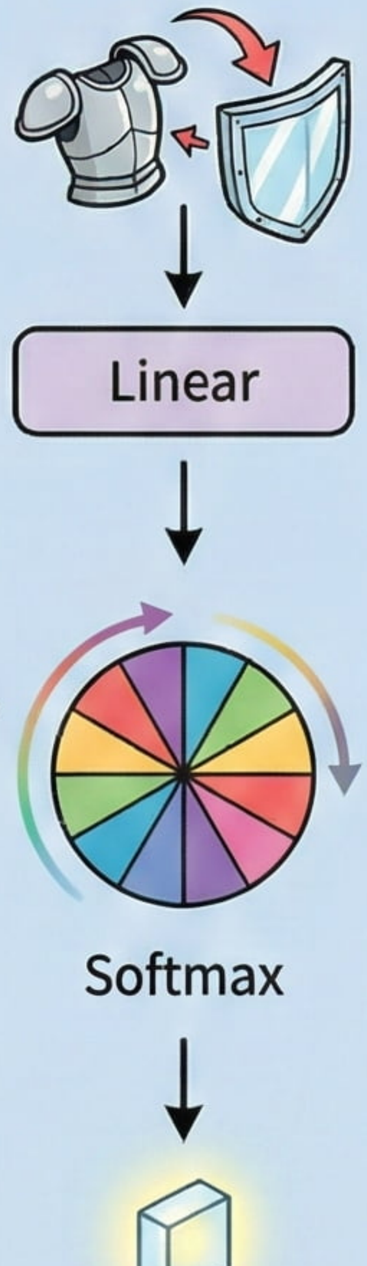


Positional Enc

N × Transformer Block（堆叠层）



输出



下一个词

现在的 ChatGPT（GPT-3.5/4）原理也是一样的，只是模型更大，读的书更多，还经过了人类老师的调教（RLHF）哦！





Ohmsha