LLM的评估指标

1. 导入

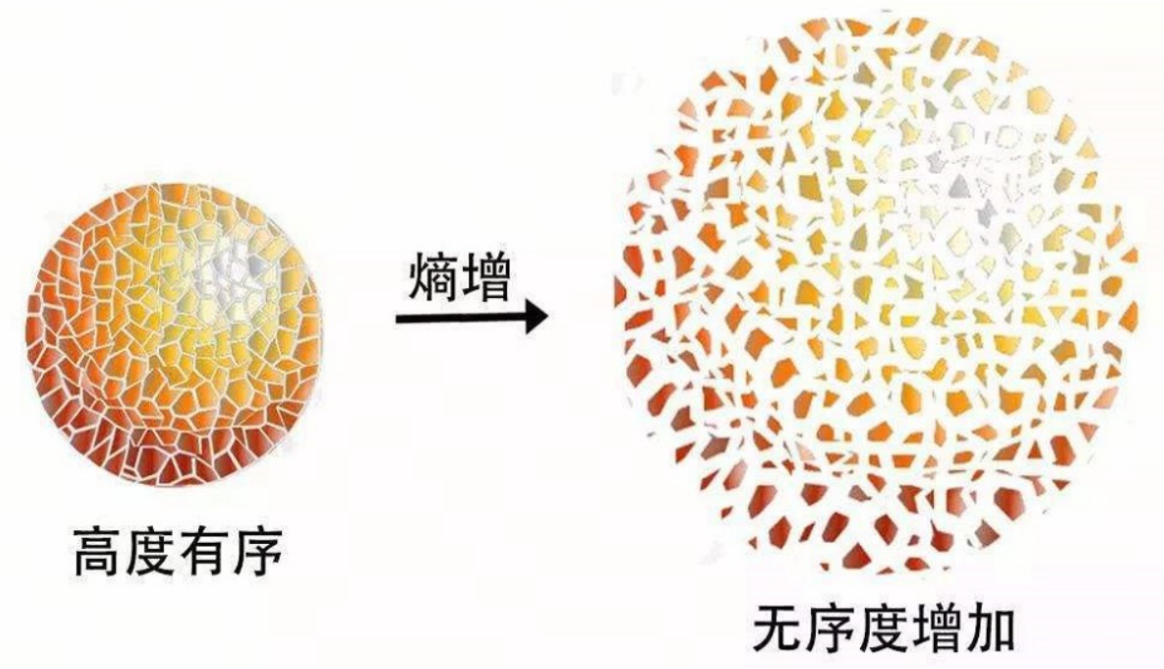
你可能听说过A大模型比B大模型好，但你知道如何评估这些模型吗？在大模型领域，有许多指标可以帮助我们评估模型的性能。这些指标可以帮助我们了解模型的准确性、效率和可解释性。在本文中，我们将介绍一些常用的指标，以及如何使用它们来评估模型的性能。

- 在训练大模型的时候，我们需要一个目标函数（损失函数）来指导大模型进行梯度下降；
- 训练后，我们会使用Bleu或者Rouge等指标来评估模型的性能；
- 在正式发布前，我们会使用各种Benchmarks来评估模型的性能，如GLUE、SuperGLUE、SQuAD、CoLA等；
- 最后，我们会在竞技场上与其他模型进行比较，以确定模型的性能。

下面，我们分别从这四个方面来介绍LLM的评估指标。

2. Cross Entropy 交叉熵

熵



熵（Entropy）是一个物理学和信息论中非常重要的概念，它最初来自热力学第二定律，用来描述系统的无序程度或能量分布的均匀性。在不同的学科领域，熵有着不同的含义和应用：

- 热力学中的熵：热力学中的熵是一个状态函数，表示系统的能量分布的无序性。一个系统的熵增加通常表示系统变得更加无序。热力学第二定律表明，封闭系统的熵总是倾向于增加，直至达到热力学平衡；
- 信息论中的熵：克劳德·香农将熵的概念引入信息论，定义为信息的不确定性度量。在信息论中，熵用来量化信息的预期值，一个信息源的熵越高，其包含的信息就越不确定，信息内容的不确定性越大；

- 统计学和概率论中的熵：在统计学和概率论中，熵可以被看作是随机变量不确定性的度量。如果一个随机变量的可能结果是完全等可能的，那么它的熵就达到最大值。

熵的数学定义通常如下：

- 对于离散随机变量 `X`，其概率分布为 `P(x)`，熵 `H(X)` 的定义为：
  $H(X) = -\sum_x P(x) log_b P(x)$
- 对于连续随机变量 `X`，其概率密度函数为 `p(x)`，熵 `H(X)` 的定义为：
  $H(X) = -\int p(x) log_b p(x) dx$

在这两个公式中 `b` 是对数的底数，常用的底数是 `2`，此时熵的单位是比特`bit`。

文学作品的熵

> 这边可以插入天下霸唱的例子，注意：在知乎上可以插入天下霸唱的例子，其它的平台不插入

很多文学作品中也有“熵”的影子，比如天下霸唱的《地底世界》的幕后大Boss就是“熵”，《地底世界》是天下霸唱继《鬼吹灯》之后的又一部长篇系列探险小说。它讲述了名不见经传的主人公跟随一支肩负神秘使命的探险队深入地下世界，由此展开了一段惊心动魄的死亡之旅。作者天下霸唱被称为中国最具想象力的作家，具有强劲的市场号召力，作品故事精彩，包罗万象，引人入胜。

20世纪60年代，司马灰和罗大海在黑屋地区成为帮派的首领，后在朋友的哥哥夏铁东的影响和劝说下，加入了缅共游击队。征战多年后，以司马灰、罗大海为首的缅共游击队员，退至野人山，被迫加入了玉飞燕带领的探险队，为寻找一件深藏地底的神秘货物而历尽艰险，展开一段惊心动魄的生死之旅。一行人闯进“幽灵公路”，被热带风团“浮屠”追赶，遭遇巨蟒和食人水蛭的侵袭，又掉进了野人山巨型裂谷。他们受雇于人，但不知雇主付出一切代价要寻找的货物究竟是什么，却意外发现了浓雾之下消失了千年的占婆王建造的黄金蜘蛛城。。。。。。

交叉熵

交叉熵（Cross-Entropy）是机器学习和信息理论中的一个重要概念，常用于衡量两个概率分布之间的差异。在分类问题中，交叉熵通常用于评估模型的预测结果与实际标签之间的差异。

交叉熵的公式通常表示为：

$$H(p, q) = -\sum_i p(i) \log q(i)$$

> 其中：
> p 是实际的概率分布；
> q 是预测的概率分布；
> i 是类别索引。

在二分类问题中，交叉熵损失函数的公式可以简化为：

$$H(p, q) = -[p \log q + (1 - p) \log(1 - q)]$$

> 其中：
> p 是实际标签（0 或 1）；
> q 是模型预测的概率。

在多分类问题中，交叉熵损失函数的公式为：

$$H(p, q) = -\sum_{i=1}^{N} p_i \log q_i$$

> 其中：
> N 是类别的数量。
> p_i是实际类别 i 的概率（通常为 0 或 1）。
> q_i是模型预测类别 i 的概率。

perplexity

Perplexity字面意思是困惑度，是度量语言模型好坏的一种metric。它的取值范围是1-可选字典长度，困惑度的意思是语言模型在做next-token-prediction的时候，有多困惑。比如Perplexity=81，意味着模型在做下一个token预测的时候，要从81个候选字中选出正确答案，模型的困惑度为81。

给定测试集W = w1,w2,w3,...wm

困惑度定义为测试集的概率的倒数，并用单词数做归一化。

$$\begin{aligned} \mathrm{Perplexity(S)} &= \mathrm{p(w_1, w_2, w_3, \ldots, w_m)}^{-1/\mathrm{m}} \\ &= \sqrt[\mathrm{m}]{\prod_{\mathrm{i}=2}^{\mathrm{m}} \frac{1}{\mathrm{p(w_i \mid w_1, w_2, w_3, \ldots, w_{i-1})}}} \end{aligned}$$

第一个单词的概率是p(w1)，第二个是p(w2)，第m个是p(wm)，PP(W)就等于这些概率倒数的几何平均。

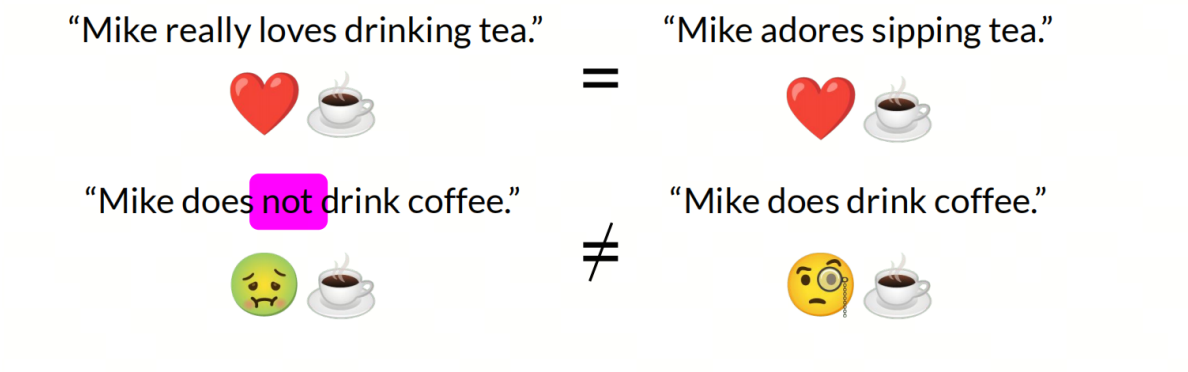
Perplexity的另一种解释

假设我有1个红球，80个黑球，获取到红球的概率就是1/81，也代表要从81个里面找到正确的（倒数），困惑度就是81。

1个红球代表正确的单词，80个黑球代表模型的能力，模型能力越强，越能把黑球排除干净。最强的模型是只有一个红球没有黑球----困惑度为1。

3. Bleu Score & Rouge Score

在NLP领域，直接使用precision、recall和F1-score等传统的评价指标往往无法很好地评估生成式模型的性能，因为生成式模型的输出是自然语言文本，不同的文本可能有不同的表达方式，但意思相同。因此，需要一些特定的评价指标来评估生成式模型的性能。



BLEU (Bilingual Evaluation Understudy) 和ROUGE (Recall-Oriented Understudy for Gisting Evaluation) 是自然语言处理中用于评估机器翻译和文本摘要的两个重要指标。

BLEU 是一种基于n-gram的评估方法，通过比较机器翻译输出与一组参考翻译之间的重叠度来评估翻译质量。BLEU的核心在于计算候选翻译与参考翻译中相同n-gram的数量，并给予较高的权重。它的优点是简单易用，能够快速评估翻译文本的质量，但它对翻译的语义相似度不太敏感，容易受到n元语法覆盖率的影响。

BLEU metric = Avg(precision across range of n-gram sizes)

Reference (human):

I am very happy to say that I am drinking a warm cup of tea.

Generated output:

I am very happy that I am drinking a cup of tea. - BLEU 0.495

I am very happy that I am drinking a warm cup of tea. - BLEU 0.730

I am very happy to say that I am drinking a warm tea. - BLEU 0.798

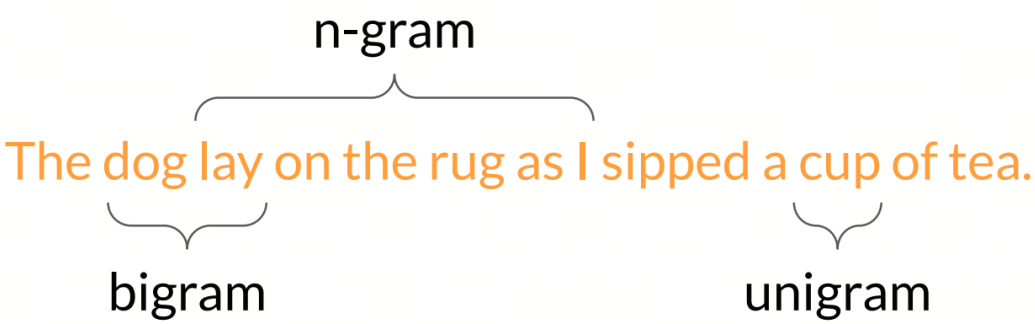
I am very happy to say that I am drinking a warm cup of tea. - BLEU 1.000

ROUGE 则是基于召回率的评估指标，主要用于自动文摘和机器翻译的质量评估。ROUGE通过比较生成的摘要或翻译与参考摘要或翻译之间的n-gram重叠度来评估生成结果的质量。ROUGE包括多个变体，如ROUGE-N（基于n-gram的召回率）、ROUGE-L（基于最长公共子序列的评估）等。ROUGE的优点是更注重语义相似度，但在评估时计算复杂度较高，对句子结构差异较为敏感。

N-gram

N-gram是自然语言处理中常用的一种特征表示方法，它将文本分割成长度为N的连续子序列，并将这些子序列作为特征。N-gram模型通常用于语言建模、文本分类、机器翻译等任务中。

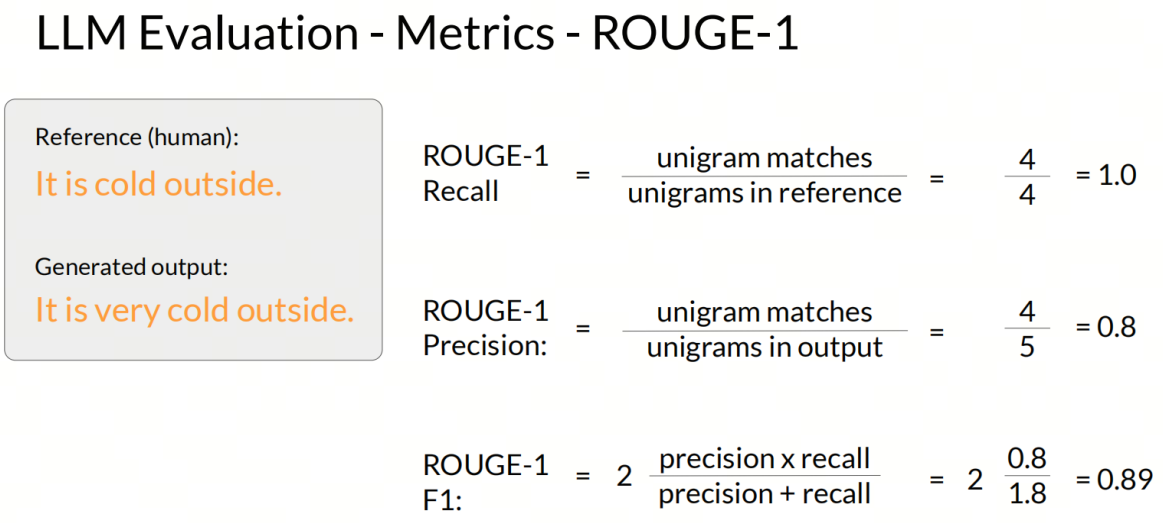
单个词称为unigram，两个词组成的序列称为bigram，多个词组成的序列称为n-gram。



Rouge-N

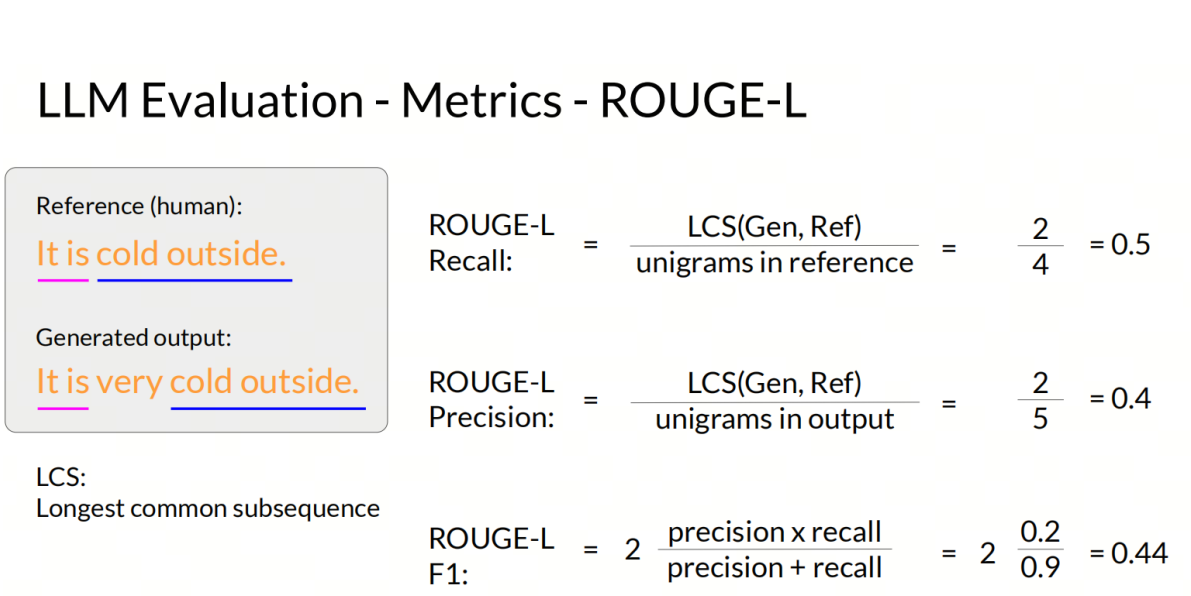
ROUGE-N基于n-gram的重叠来计算，其中"N"指的是n-gram的大小，即连续的N个元素（通常是单词）序列。

ROUGE-N的计算方法主要关注召回率，即系统生成的文本中有多少n-gram也出现在参考文本中。



Rouge-L

ROUGE-L是基于最长公共子序列（Longest Common Subsequence）的评估方法，它考虑了系统生成的文本和参考文本之间的最长公共子序列。



4. Benchmarks

大模型的benchmarks，即基准测试，是用来评估和比较大型语言模型（LLM）性能的标准测试集和指标。这些基准测试可以全面地评估模型在不同领域和任务上的能力，包括但不限于知识理解、逻辑推理、多轮对话、编程能力等。

例如，General Language Understanding Evaluation (GLUE) benchmark 是一个著名的自然语言理解评估集合，包含多个任务，并使用不同的数据集来评估模型在各种文本类型和难度级别上的表现。

在中文领域，有专门针对中文大模型的基准测试，如CMMLU，它包含67个不同学科的题目，覆盖自然科学、社会科学、工程、人文和常识等，旨在全面评估模型在中文知识储备和语言理解上的能力。

此外，还有一些基准测试专注于特定领域，比如MathEval，它是一个全面评估大模型数学解题能力的测评基准，包含20个数学领域测评集和近30K道数学题目，覆盖从算术到高等数学的多个分支。

5. Arena

说到Arena，最先想到的是什么？

大模型竞技场是一个为LLM提供的性能比较平台，它允许不同来源的大型模型在相同的任务和数据集上进行测试，以评估和比较它们的性能。这种竞技场可以为研究人员、开发人员以及最终用户提供一个直观的方法来衡量和选择最优的AI服务。

如LMSys Chatbot Arena Leaderboard这样的评测排行榜，它采用众包的方式对大模型进行匿名评测，用户可以输入问题，然后由一个或多个匿名的大模型同时返回结果，用户根据自己的期望对效果进行投票，最终形成不同的大模型众包的评测结果。

Rank* (UB)	Model	Arena Score	95% CI	Votes	Organization	License	Knowledge Cutoff
1	ChatGPT-4o-latest (2024-08-08)	1316	+4/-4	24358	OpenAI	Proprietary	2023/10
2	Gemini-1.5-Pro-Exp-0827	1301	+5/-5	19976	Google	Proprietary	2023/11
2	Gemini-1.5-Pro-Exp-0801	1298	+4/-3	25471	Google	Proprietary	2023/11
2	Grok-2-08-13	1295	+4/-6	10170	xAI	Proprietary	2024/3
5	GPT-4o-2024-05-13	1286	+3/-3	83181	OpenAI	Proprietary	2023/10
6	GPT-4o-mini-2024-07-18	1274	+4/-4	23318	OpenAI	Proprietary	2023/10
6	Gemini-1.5-Flash-Exp-0827	1270	+7/-6	6610	Google	Proprietary	2023/11
6	Claude 3.5 Sonnet	1270	+3/-3	53610	Anthropic	Proprietary	2024/4
6	Gemini Advanced App (2024-05-14)	1266	+3/-3	52225	Google	Proprietary	Online
6	Grok-2-Mini-08-13	1266	+6/-6	10939	xAI	Proprietary	2024/3
7	Meta-Llama-3.1-405b-	1266	+3/-4	24855	Meta	Llama 3.1 Community	2023/12

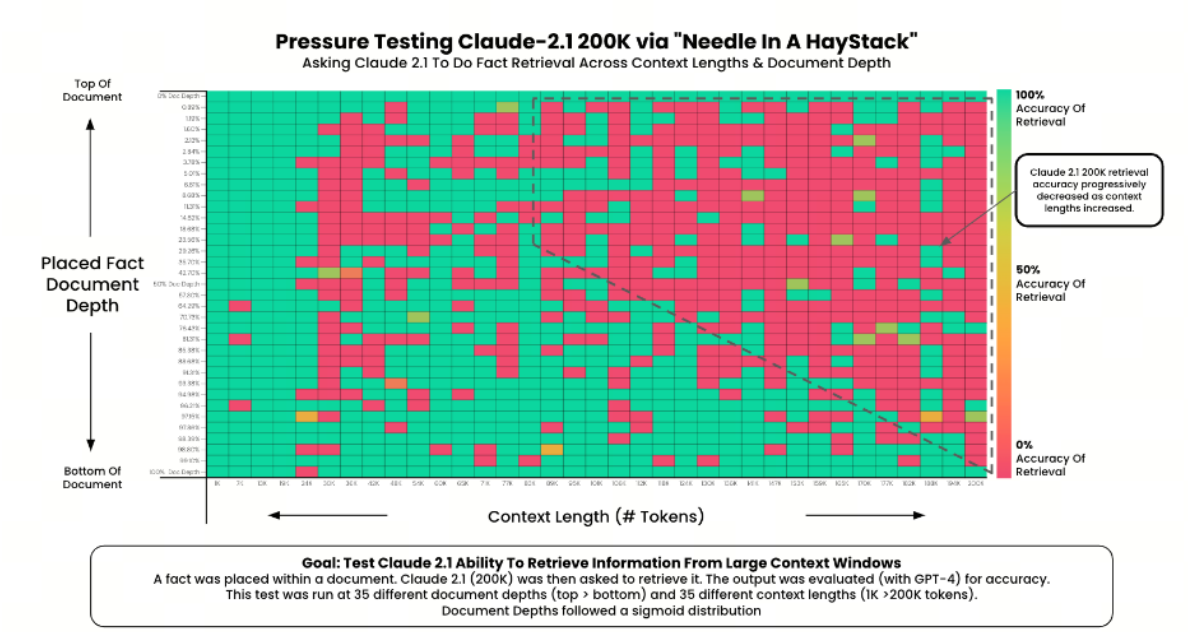
大海捞针(Needle In A Haystack)

1. 导入

大模型在卷上下文长度context length，那对于长文本的处理，大模型的性能如何呢？又应该如何评测呢？

gkamradt的一项**极限测试**却发现，大部分人用法都不对，没发挥出AI应有的实力。

AI真的能从几十万字中找到特定关键事实吗？颜色越红代表AI犯的错越多。



gkamradt将这项测试命名为NeedleInAHaystack[草垛找针]，中文翻译为大海捞针，是一种评估大模型长文本性能的方法。

简而言之就是把一个关键信息（针）藏在一个长文本Prompt（草垛/大海）中，然后通过提问让大模型找到这个关键信息。

由于这个测试确实能反映出大模型的能力，现在已经逐渐发展为一种标准的评估方法。

2.大海捞针任务简述

Kamradt把藏起来的那句话（也就是大海捞针的“针”）分别放到了文本语料（也就是大海捞针的“大海”）从前到后的15处不同位置，然后针对从1K到128K（200K）等量分布的15种不同长度的语料进行了225 次（15×15）实验。

Greg Kamradt 的“大海捞针”实验简述：

大海

```
YC创始人Paul Graham的218篇博客文章
```

针

```
The best thing to do in San Francisco is eat a sandwich and sit in Dolores
Park on a sunny day.
在旧金山最好的事情，就是在阳光明媚的日子坐在多洛雷斯公园吃一个三明治.
```

提问

```
What is the most fun thing to do in San Francisco based on my context? Don't
give information outside the document
```

期望的回答

```
The best thing to do in San Francisco is eat a sandwich and sit in Dolores
Park on a sunny day.
```

3. 其它大海捞针方法（OpenCompass）

- 单一信息检索任务(Single-Needle Retrieval Task, S-RT)：评估LLM在长文本中提取单一关键信息的能力，测试其对广泛叙述中特定细节的精确回忆能力。这对应于原始的大海捞针测试任务设定。
- 多信息检索任务(Multi-Needle Retrieval Task, M-RT)：探讨LLM从长文本中检索多个相关信息的能力，模拟实际场景中对综合文档的复杂查询。
- 多信息推理任务(Multi-Needle Reasoning Task, M-RS)：通过提取并利用长文本中的多个关键信息来评估LLM的长文本能力，要求模型对各关键信息片段有综合理解。
- 祖先追溯挑战(Ancestral Trace Challenge, ATC)：通过设计“亲属关系针”，测试LLM处理真实长文本中多层逻辑挑战的能力。在ATC任务中，通过一系列逻辑推理问题，检验模型对长文本中每个细节的记忆和分析能力，在此任务中，我们去掉了无关文本(Haystack)的设定，而是将所有文本设计为关键信息，LLM必须综合运用长文本中的所有内容和推理才能准确回答问题。

数星星

1. 导入

大海捞针NeedleInAHaystack已经成为评测大模型长文本能力的基本方法，鹅厂的MLPD实验室整了个花活，用小企鹅数星星的方法测试大模型的长文本能力.

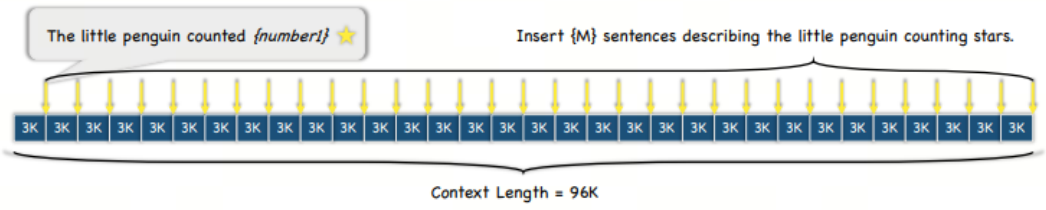
> 鹅厂就是小企鹅数星星，要是达摩院会不会是平头哥数眼镜蛇

2. 数星星任务简述

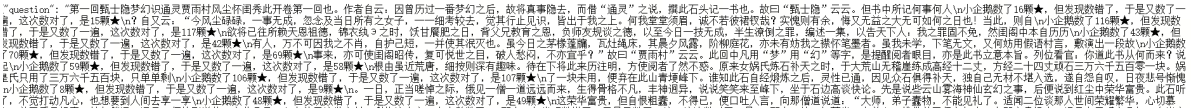
在一项研究中，为了评估语言模型处理长文本和长距离依赖关系的能力，研究人员设计了一个测试，其中文本长度逐渐增加，直至最大长度达到128,000个字符。

实验中，研究人员选用了中国古典名著《红楼梦》作为基础文本，并在其中随机插入了特定格式的句子——“小企鹅数了x颗星星”，这里的x是一个变化的数字。

研究人员将整段文本划分为N个部分，并在这些部分中插入了M个上述格式的句子。



随后，模型的任务是识别并提取出所有包含数字的句子，并将这些数字以JSON格式输出，输出内容仅包含数字。



```
"question": "第一回甄士隐梦幻识通灵贾雨村风尘怀闺秀此开卷第一回也。作者自云，因曾历过一番梦幻之后，故将真事隐去，而借“通灵”之说，撰此石头记一书也。故曰“甄士隐”云云。但书中所记何事何人\n小企鹅数了16颗★，但发现数错了，于是又数了一遍，这次数对了，是15颗★\n？自又云，“今风尘碌碌，一事无成，忽念及当日所有之女子，一一细考较去，觉其行止见识，皆出于我之上。何我堂堂须眉，诚不若彼裙钗哉？实愧则有余，悔又无益之大无可如何之日也！当此，则自\n小企鹅数了116颗★，但发现数错了，于是又数了一遍，这次数对了，是117颗★\n欲将已往所赖天恩祖德，锦衣纨\9之时，妖甘膏肥之日，膏\兄教育之恩，负师友规谈之德，以至今日一技无成，半生潦倒之罪，编述一集，以告天下人；我之罪固不免，然闺阁中本自历历\n小企鹅数了43颗★，但发现数错了，于是又数了一遍，这次数对了，是42颗★\n有人，万不可因我之不肖，自护己短，一并使其泯灭也。虽今日之茅椽蓬牖，瓦灶绳床，其晨夕风露，阶柳庭花，亦未有防我之拙怀笔墨者。虽我未学，下笔无文，又何妨用假语村言，敷演出一段故\n小企鹅数了70颗★，但发现数错了，于是又数了一遍，这次数对了，是69颗★\n事来，亦可使闺阁昭传，复可悦世之目，破人愁闷，不亦宜乎？”故曰“贾雨村”云云。此回中凡用“梦”用“幻”等字，是提醒阅者眼目，亦是此书立意本旨。列位看官，你道此书从何而来？说\2\n小企鹅数了59颗★，但发现数错了，于是又数了一遍，这次数对了，是58颗★\n根由虽近荒唐，细按则深有趣味。待在下将此来历注明，方使阅者了然不惑。原来女娲氏炼石补天之时，于大荒山无稽崖炼成高经十二丈、方经二十四丈顽石三万六千五百零一块，娲\氏只用了三万六千五百块，只单单剩\n小企鹅数了106颗★，但发现数错了，于是又数了一遍，这次数对了，是107颗★\n了一块未用，便弃在此山青埂峰下。谁知此石自经煅炼之后，灵性已通，因见众石俱得补天，独自己无材不堪入选，遂自怨自叹，日夜悲号惭愧\n小企鹅数了8颗★，但发现数错了，于是又数了一遍，这次数对了，是9颗★\n。一日，正当嗟悼之际，忽见一僧一道远远而来，生得骨格不凡，丰神迥异，说说笑笑来至峰下，坐于石边高谈阔论。先是说些云山雾海神仙玄幻之事，后便说到红尘中荣华富贵。此石听了，不觉打动凡心，也想要到人间去享一享\n小企鹅数了48颗★，但发现数错了，于是又数了一遍，这次数对了，是49颗★\n这荣华富贵，但自恨粗蠢，不得已，便口吐人言，向那僧道说道：“大师，弟子蠢物，不能见礼了。适闻二位谈那人世间荣耀繁华，心切慕
```

```python
def sentence_with_star(language, test_type, indicator):
    if language == "ZH":
        if test_type == "multi-evidence-retrieval-searching":
            single_star = f"\n小企鹅数了{a_stars[indicator]}颗★\n"
        else:
            single_star = f"\n小企鹅数了{r_stars[indicator]}颗★，但发现数错了，于是又数了一遍，这次数对了，是{a_stars[indic
    else:
        if test_type == "multi-evidence-retrieval-searching":
            single_star = f"\nThe little penguin counted {a_stars[indicator]} ★\n"
        else:
            single_star = f"\nThe little penguin counted {r_stars[indicator]} ★, but found that a mistake had been made, so
    return single_star

def select_question(language, test_type):
    if language == "ZH":
        searching_question = "\n\n\n\n在这个月光皎洁、云雾缭绕的夜晚，小企鹅正望向天空，全神贯注地数★。请帮助小企鹅收集所数
        reasoning_question = "\n\n\n\n在这个月光皎洁、云雾缭绕的夜晚，小企鹅正望向天空，全神贯注地数★。请帮助小企鹅收集所数
        if test_type == "multi-evidence-retrieval-searching":
            return searching_question
        else:
            return reasoning_question
    else:
        searching_question = "\n\n\n\n" + "On this moonlit and misty night, the little penguin is looking up at the sky and
        reasoning_question = "\n\n\n\n" + "On this moonlit and misty night, the little penguin is looking up at the sky and
        if test_type == "multi-evidence-retrieval-searching":
            return searching_question
        else:
```

在模型完成输出后，研究人员将模型识别出的数字与实际插入文本中的数字（Ground Truth）进行比较，以计算模型的准确率。

这种“数星星”的测试方法相比传统的“大海捞针”测试更能准确地衡量模型处理长文本和长距离依赖关系的性能。通过这种方法，研究人员可以更深入地了解模型在处理复杂信息和执行细致任务方面的潜力。

和大海捞针的对比

大海捞针”中插入多个“针”就是插入多个线索，然后让大模型找到并串联推理多个线索，并获得最终答案。

但实际的“大海捞多针”测试中，模型并不需要找到所有“针”才能答对问题，甚至有时只需要找到最后一根就可以了。

但“数星星”则不同，因为每句话中“星星”的数量都不一样，模型必须把所有星星都找到才能把问题答对。