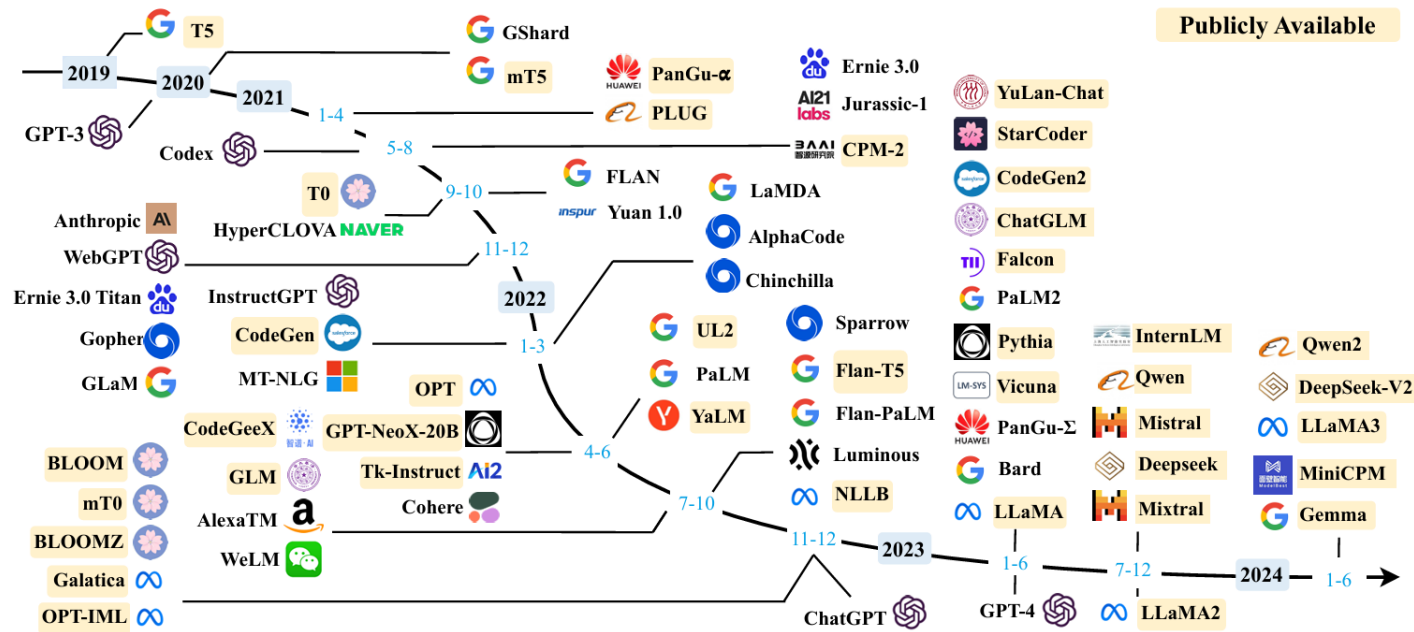




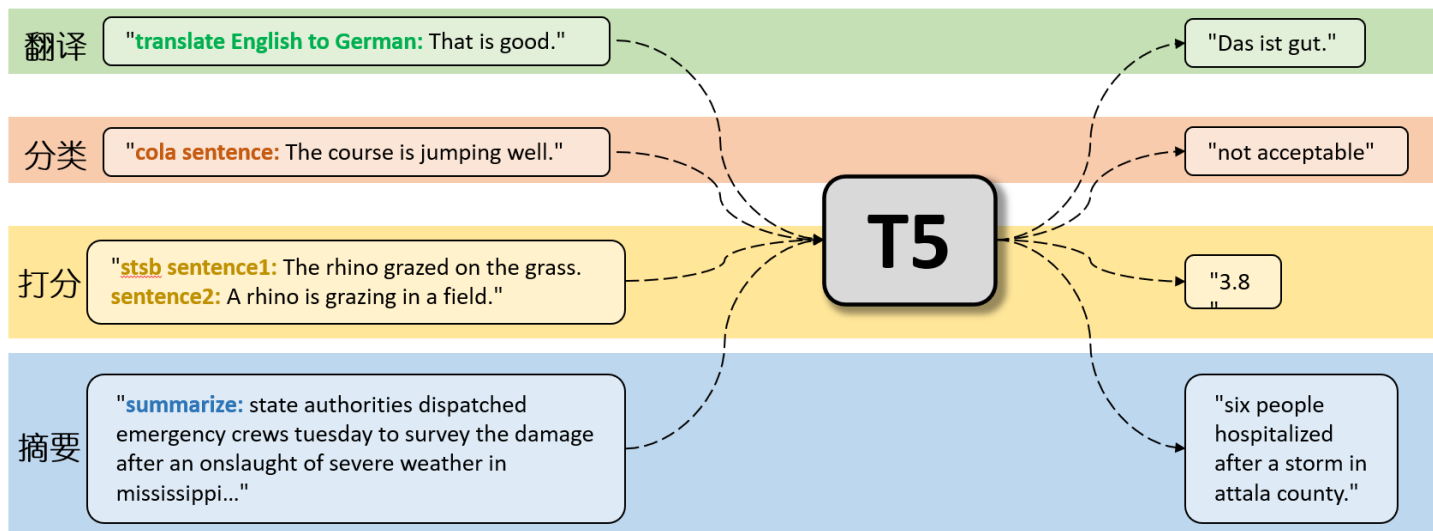
基于LLM的命名实体识别和摘要任务

What are Large Language Models (LLMs)?



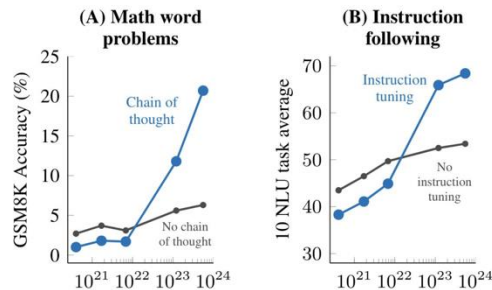
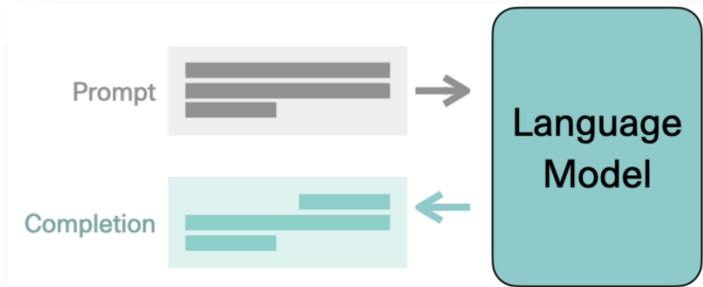
Language Generation

- 语言生成：将所有下游任务转化成语言生成任务后，预训练模型通过生成文本就可以解决各种下游任务的方法



The good side of LLMs

- One single model to solve many NLP tasks!
 - 情感分析、机器翻译、数学推理、文本分类
- Emergent properties in LLMs
 - 上下文学习
 - 指令遵循
 - 循序渐进的推理能力



(Wei et al., 2022)

命名实体识别

- 命名实体识别 (Named Entities Recognition, NER) 是指从文本中识别出各类命名实体的任务。

分类标签:

B: 开始

I: 中间

E: 结束

O: 其他

S: 单个

小明在复旦大学上黄老师的自然语言处理课。

B E O B I I E O B I E O B I I I I E O

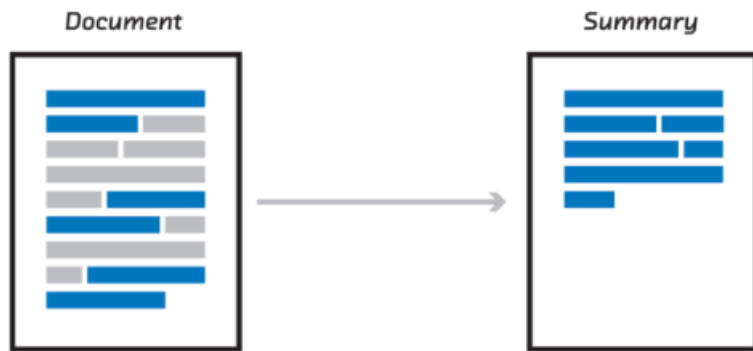
- 指标

$$\text{Precision} = \frac{TP}{TP+FP}, \text{Recall} = \frac{TP}{P}$$

$$F1 = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

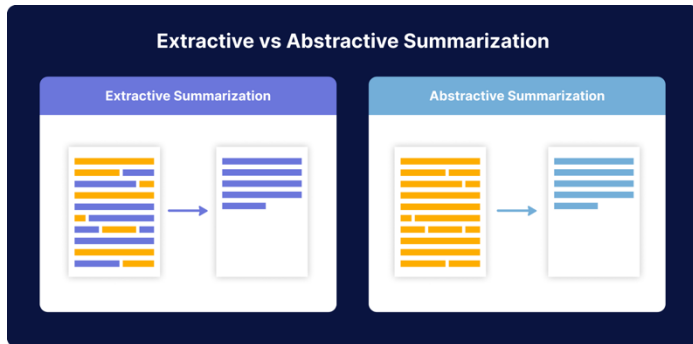
摘要生成

- **文本摘要 (Text Summarization)** 是一种利用算法自动实现文本分析、内容提炼并生成摘要的技术。当前时代的信息量增长远超于人类对信息地处理和利用能力，从而导致我们无法快速地挑选和有效地运用信息，这种现象称为**信息过载 (Information Overloading)**。而文本摘要技术可以对信息进行简化，帮助人们快速获取和理解新的信息。



摘要类型

- **抽取式文本摘要**：关键内容是从原始文本中提取的，同时所提取的内容通常不会以任何方式进行修改。
 - 缺点：**抽取式文本摘要**所获得的结果都来源于原始文本，因此所得到摘要文本中每个单句的语义准确性和语法正确性都可以得到很好的保证。但是由于原始的句子来源于不同的段落甚至篇章，所生成的摘要中句子之间的**连贯性则很难保证，造成摘要整体的可读性较差。**
- **生成式文本摘要**：可以**产生原始文档中不存在的新文本**，通常能产生更精简凝练并且更具有连贯性的摘要。



摘要任务评估指标

- ROUGE (Recall-Oriented Understudy for Gisting Evaluation)
 - 专注于召回率而非精度。关注有多少个参考译句中的 n 元词组出现在了输出之中。
 - 评估生成的摘要与参考摘要之间的相似程度和准确性。
 - ROUGE 指标通常包括多个子指标，如 ROUGE-N（考虑 n-gram 重叠）、ROUGE-L（考虑最长公共子序列）和 ROUGE-W（考虑词级别的重叠）等。
- 例子

机器译文: a cat is on the table

参考译文: there is a cat on the table

上面例子的 ROUGE-1 和 ROUGE-2 分数如下:

$$ROUGE-1 = \frac{|a, cat, is, on, the, table|}{|there, is, a, cat, on, the, table|} = \frac{6}{7}$$

$$ROUGE-2 = \frac{|(a, cat), (on, the), (the, table)|}{|(there, is), (is, a), (a, cat), (cat, on), (on, the), (the, table)|} = \frac{1}{2}$$

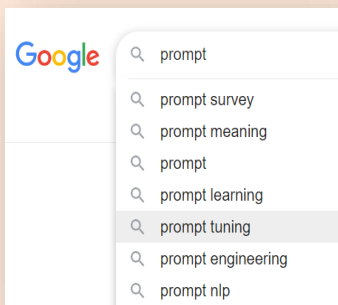
什么是提示 (prompt) ?

填空题

1. _____, 红掌拨清波。《咏鹅》 骆宾王
2. 天似穹庐, _____, 天苍苍, 野茫茫, _____。《敕勒歌》 北朝民歌
3. _____, 花落知多少?《春晓》 孟浩然
4. 儿童相见不相识, _____。《回乡偶书》 贺知章
5. _____, 低头思故乡。《静夜思》 李白
6. 日出江花红胜火, _____, 能不忆江南?《忆江南》 白居易
7. _____, 游子身上衣。谁言寸草心, _____。《游子吟》 孟郊
8. _____, 桂花何团团。 _____, 问君与谁餐?《古朗月行》 李白
9. _____, 春风送暖入屠苏。千门万户曈曈日, _____。《元日》 王安石
10. _____, 更上一层楼。《登鹳雀楼》 王之涣

出题者给定古诗词的一部分, 让练习者根据前后的内容判断考察的是哪一句诗词并默写

搜索引擎



搜索引擎根据用户已有的输入进行提示 (预测), 引导用户完成内容查找。

写作提示



出题者提供一些材料 (文本/图像), 让练习者根据材料规定好的主题完成作文写作

提示, 就是线索, 会引导某个对象去做某件事

How to Craft Effective Prompts?

- 原则

- Write clear instructions
 - 明确任务、指定模型所扮演的角色
- Provide reference text
 - 提供例子
- Split complex tasks into simpler subtasks
 - 拆分成子任务
- Give the model time to “think”
 - think step by step

- 参考

- <https://platform.openai.com/docs/guides/prompt-engineering>

基于LLM的命名实体识别和摘要任务

- 模型：
 - Qwen2.5-0.5B-Instruct
 - <https://huggingface.co/Qwen/Qwen2.5-0.5B-Instruct>
- 任务：通过设计 prompt 引导模型完成下游任务
 - 命名实体识别任务
 - CHisIEC 古文命名实体识别数据集
 - <https://github.com/tangxuemei1995/CHisIEC/blob/main/data/ner/test.txt>
 - 中文新闻摘要任务
 - CNewSum: A Large-scale Chinese News Summarization Dataset with Human-annotated Adequacy and Deducibility Level

```
{  
  "article": ["庆阳市 气象台 2016年 7月 14 日 15 时 40 分 发布 雷电 黄色 预警 信号：预计 今天 下午 到  
夜间 我市 大部分 地方 将 有 雷电 活动，局地 伴有 短时 强降水、阵性 大风 等 天气，请 注意 防范。", "图例  
标准 防御 指南 6 小时 内 可能 发生 雷电 活动，可能 会 造成 雷电 灾害 事故。", "1、政府 及 相关 部门 按照  
职责 做好 防雷 工作；2、密切 关注 天气，尽量 避免 户外 活动。"],  
  "summary": "庆阳市 发布 雷电 黄色 预警：预计 今天 下午 到 夜间 我市 大部分 地方 将 有 雷电 活动，局地  
伴有 . . ."  
}
```

基于LLM的命名实体识别和摘要任务

- 必做1:
 - 使用模型完成**命名实体识别**任务，并与原来的方法得到的结果进行对比（指标为F1-macro），要求分析各自的优劣
- 必做2:
 - 使用模型完成**摘要**任务，使用不同指标（ROUGE-1, ROUGE-2）进行评价，并简单进行案例分析，说明完成的效果及提升的空间
- 选做:
 - 使用更大规模模型、有条件可尝试微调、其他认为值得探索的要点
- 提交文件：学号-姓名.ipynb
- Due:
 - 11月18日 **23:59**

评分标准

- 准时提交文件（有运行过程）、完成两个必做任务： 60分
- 文件中含有相关注释： 70分
- 设计prompt 并进行优化： 80分
- 对基于神经网络模型与基于LLM的结果进行对比分析： 90分
- 命名实体识别任务测试集分数达到0.6以上、摘要任务ROUGE-1 为0.4以上， ROUGE-2 为0.2以上： 100分
- 加分项：微调模型
- 迟交： 1天内分数*0.6， 1天后为0分（如有任何特殊情况，请提前说明，无充分证明时不接受事后解释）

联系我们

- 邮件:
 - 助教-叶俊杰: jjye23@m.fudan.edu.cn
 - 助教-仝竞奇: jqtong23@m.fudan.edu.cn
 - 助教-丁怡文: ywding23@m.fudan.edu.cn