

大模型时代的中文 信息处理

宗成庆

中国科学院自动化研究所



提 纲

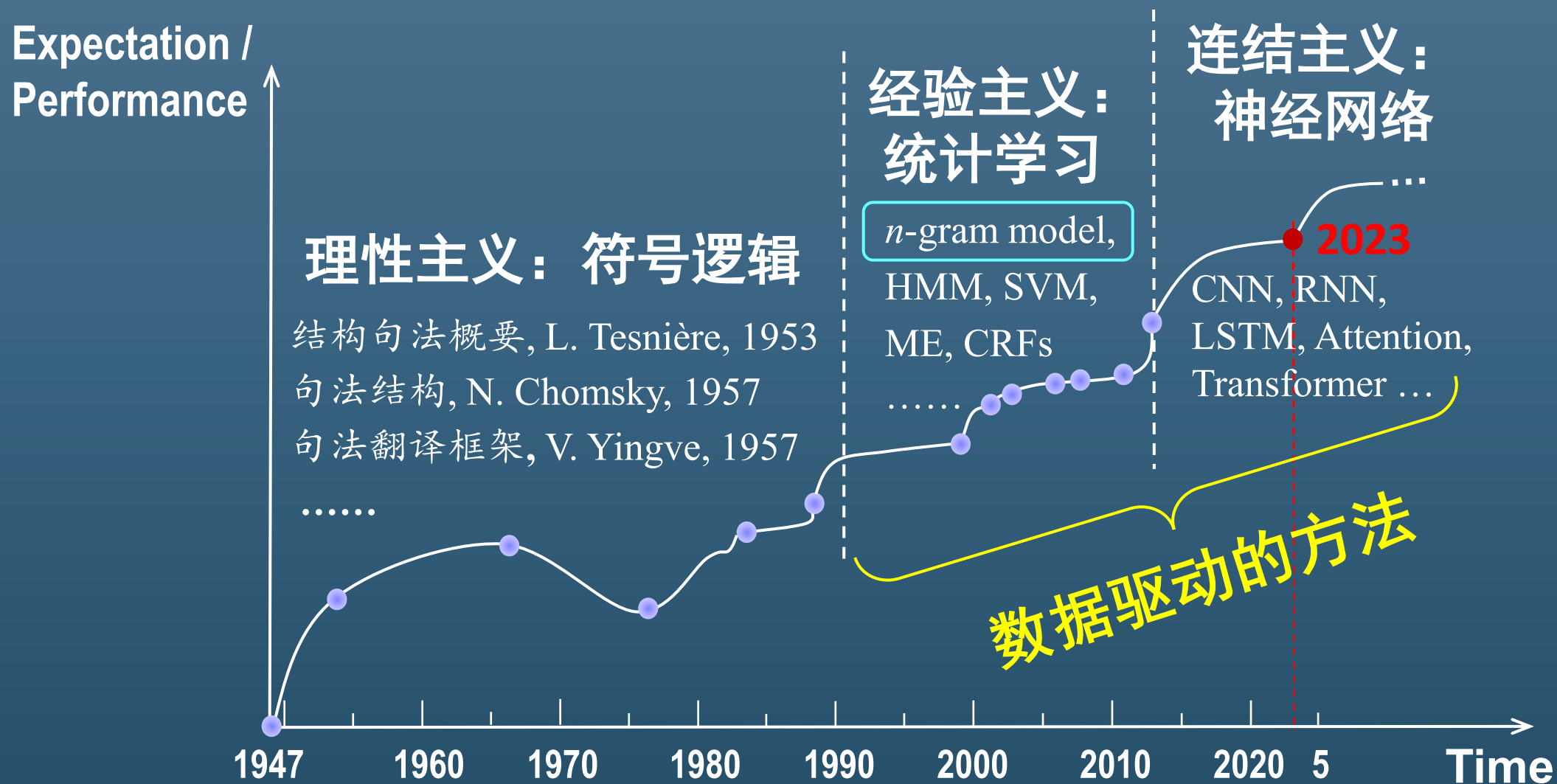
- ➡ 1. 大模型概要
- 2. 大模型的能力分析
- 3. 中文信息处理

热度

◆ 人工智能 (Artificial Intelligence, AI)



◆ 自然语言处理(Natural Language Understanding)



◆语言模型(Language Model)

— 如何计算一个句子的概率？

五月的北京开满了鲜花，到处弥漫着迷人的花香。

- 以“词”为单位计算条件概率：

$$p(s) = p(w_1) \times p(w_2|w_1) \times p(w_3|w_1w_2) \times \dots \times p(w_m|w_1 \dots w_{m-1})$$

- 如果只考虑前面有限个($n-1$)词对当前词的影响，
假如 $n=2$ ，那么，

$$p(s) = p(\text{五月}) \times p(\text{北京} | \text{五月}) \times p(\text{开} | \text{北京}) \times p(\text{满} | \text{开}) \dots \times p(。 | \text{花香})$$



Frederek Jelinek

(1932.11.18 ~ 2010.9.14)

语言模型 / n 元文法模型 (n -gram model)

已知前面几个词就可预测出下一个最有可能出现的词：

2025年阳春 ... ? $p(\text{word} | \text{context})$

context

三月/白雪/面

2025年阳春 三月，暖意融融



2025年阳春三月，暖意融融，柳丝轻舞于微风，桃花绽蕊映碧空，处处洋溢着生机与希望。



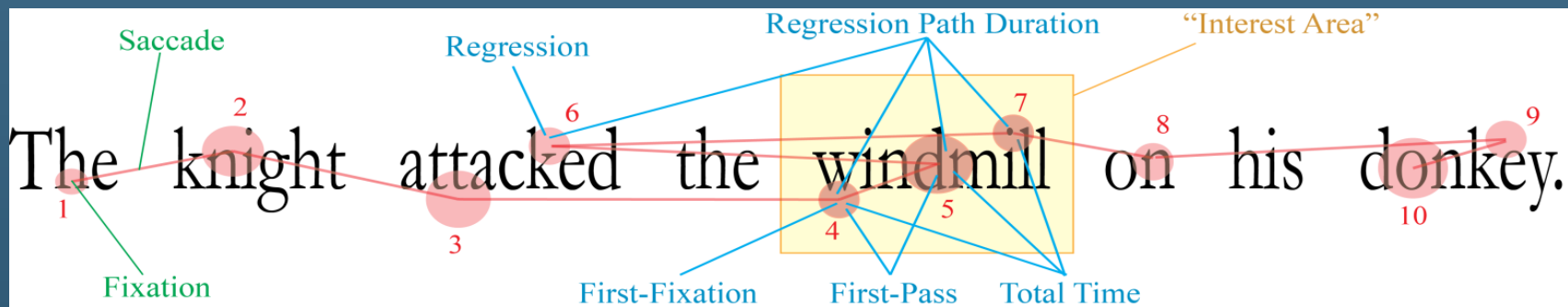
2025年阳春三月，暖意融融，长安街的樱花突然逆季盛放。我蹲身捡起一片发光的粉色花瓣，指尖竟浮现出全息投影的篆体小字："文明重启倒计时——"

- ✧ **生成式语言模型**(Generative Language Model)
- ✧ **语言生成/自动写作**(Language Generation /Automatic Writing)
- ✧ **人工智能生成内容**(Artificial Intelligence Generated Content, **AIGC**)

$$p(\text{word} | \text{context})$$

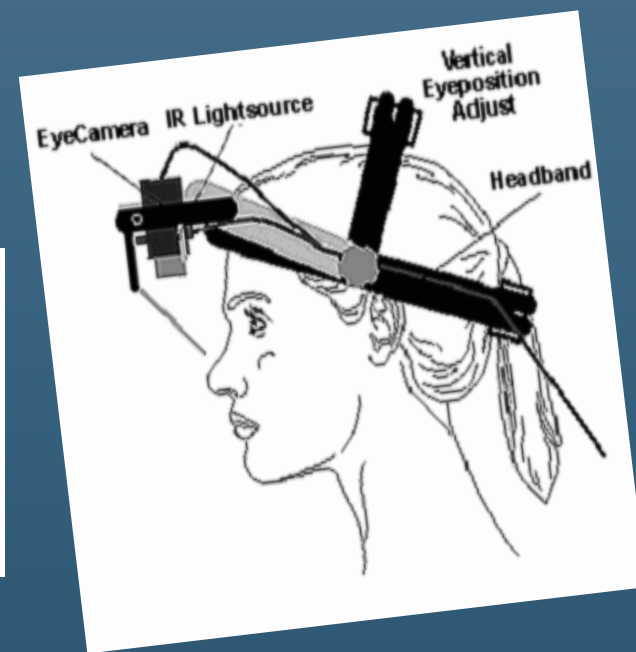
离散符号 → 连续的实数空间 → 注意力机制 → Transformer

扫视 回看 回看路径持续时间 兴趣区



固定 首次固定 第1遍 总耗时

(骑士骑着驴子袭击了风车。)



对输入端进行编码 → 对输出端进行解码

◆预训练语言模型(Pre-trained Language Model)

➤ ELMo [Peters et al., 2018] (AllenAI&WU)

➤ BERT [Devlin et al., 2019] (Google)

➤ GPT [Radford et al., 2018] (OpenAI)

- Generative Pre-Trained Transformer

+ 指令学习

+ 基于人类反馈的强化学习



大语言模型(large language models, LLMs)

①预训练

②指令学习

③强化训练

参数量超过10亿的语言模型通常被称为大语言模型(Wikipedia)。

一种语言的句子 $\implies$ 另外一种语言的句子 (翻译)

一个输入语句 $\implies$ 一个回复 (对话/聊天)

- 机器翻译(Machine translation)
- 人机对话系统(Human-dialogue system)
- 问答(Question & Answer)
- 情感/情绪(Sentiment/ emotion analysis)
- 文本分类(Text classification)
- 自动摘要(Automatic summarization)
- 自动写作(Automatic)
- 图像标题(Image caption)

.....

通用人工智能
(Artificial General Intelligence, AGI)

表示 + 预测

◆大语言模型(Large Language Models, LLMs)



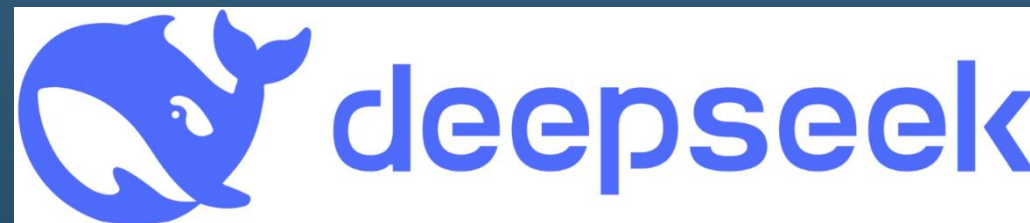
2022年11月29日, OpenAI 公司发布 ChatGPT, 允许用户以自然语言的形式提问任何问题, 系统都可以给出较好的回答。

- GPT-4.0, March 14, 2023
- Sora, Feb. 15, 2024
- GPT-4o (Omni), May 13, 2024
- o1-preview, o1-mini, Sept. 12, 2024
- o3, o3 mini, Dec. 20, 2024

Prompt: A stylish woman walks down a Tokyo street filled with warm glowing neon and animated city signage. She wears a black leather jacket, a long red dress, and black boots, and carries a black purse. She wears sunglasses and red lipstick. She walks confidently and casually. The street is damp and reflective, creating a mirror effect of the colorful lights. Many pedestrians walk about.

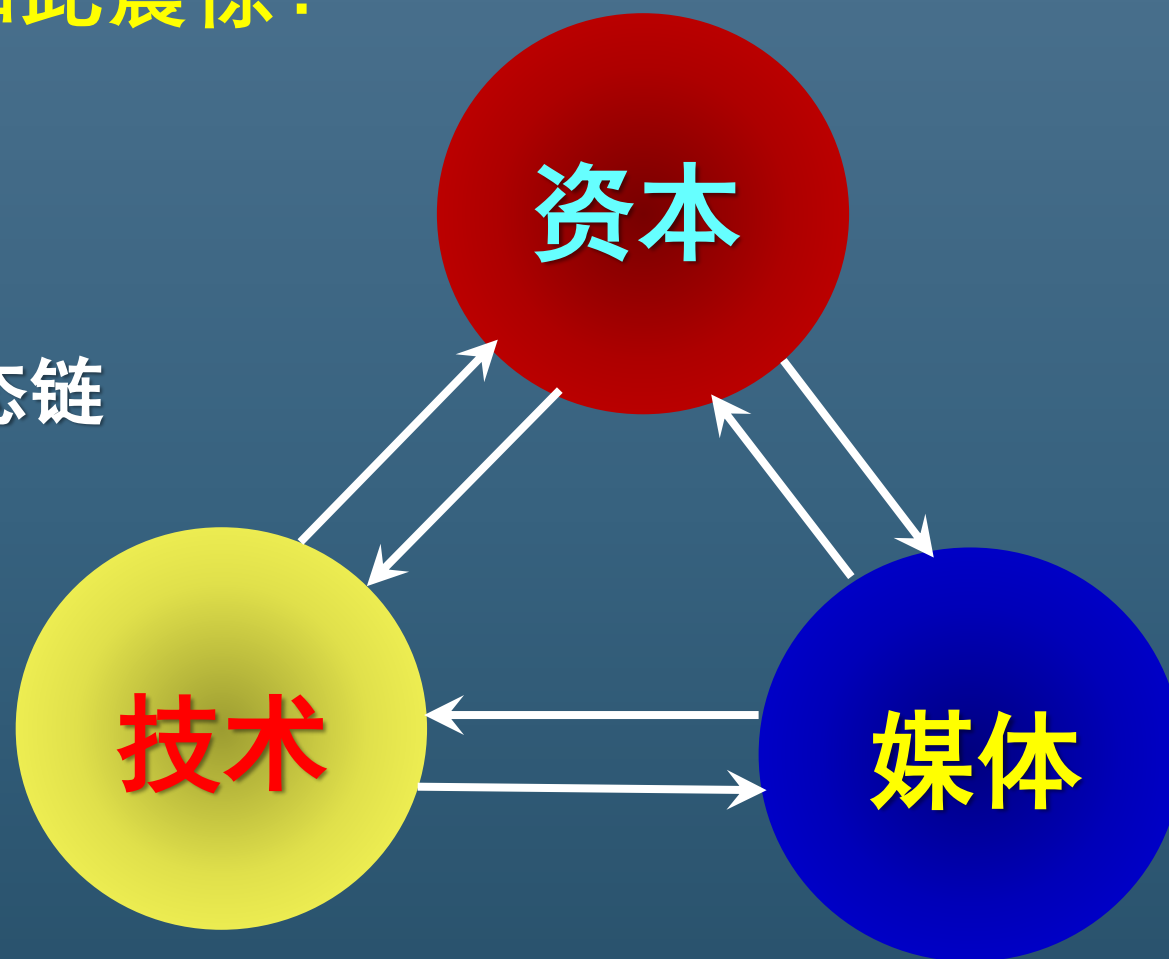


- 文心一言(百度), 豆包(字节), Kimi (月之暗面), 九天, 星火(讯飞), 通义千问(阿里), 腾讯混元, 百川, 华为盘古, 日日新(商汤), 紫冬太初, 智谱华章, ...
- 2024.01.05, **DeepSeek** 发布首个“深度求索”大模型, 随后每隔几个周/月发布一款新版大模型; 12月26日晚正式上线DeepSeek-V3首个版本并同步开源;
- 2025年1月27日, DeepSeek应用登顶苹果美国地区应用商店免费APP下载排行榜, 在美区下载榜上超越了ChatGPT; 2025年1月31日, 英伟达宣布DeepSeek-R1模型登陆NVIDIA ANIM。同时, 亚马逊和微软也接入DeepSeek-R1模型.....



● 为什么 DeepSeek 让世界如此震惊？

- 技术创新，性能提升
- 成本降低
- 技术开源，打破已有的生态链



提 纲

1. 大模型概要

➡ 2. 大模型的能力分析

3. 中文信息处理

◆ 语言翻译水平客观评价

● 客观评测

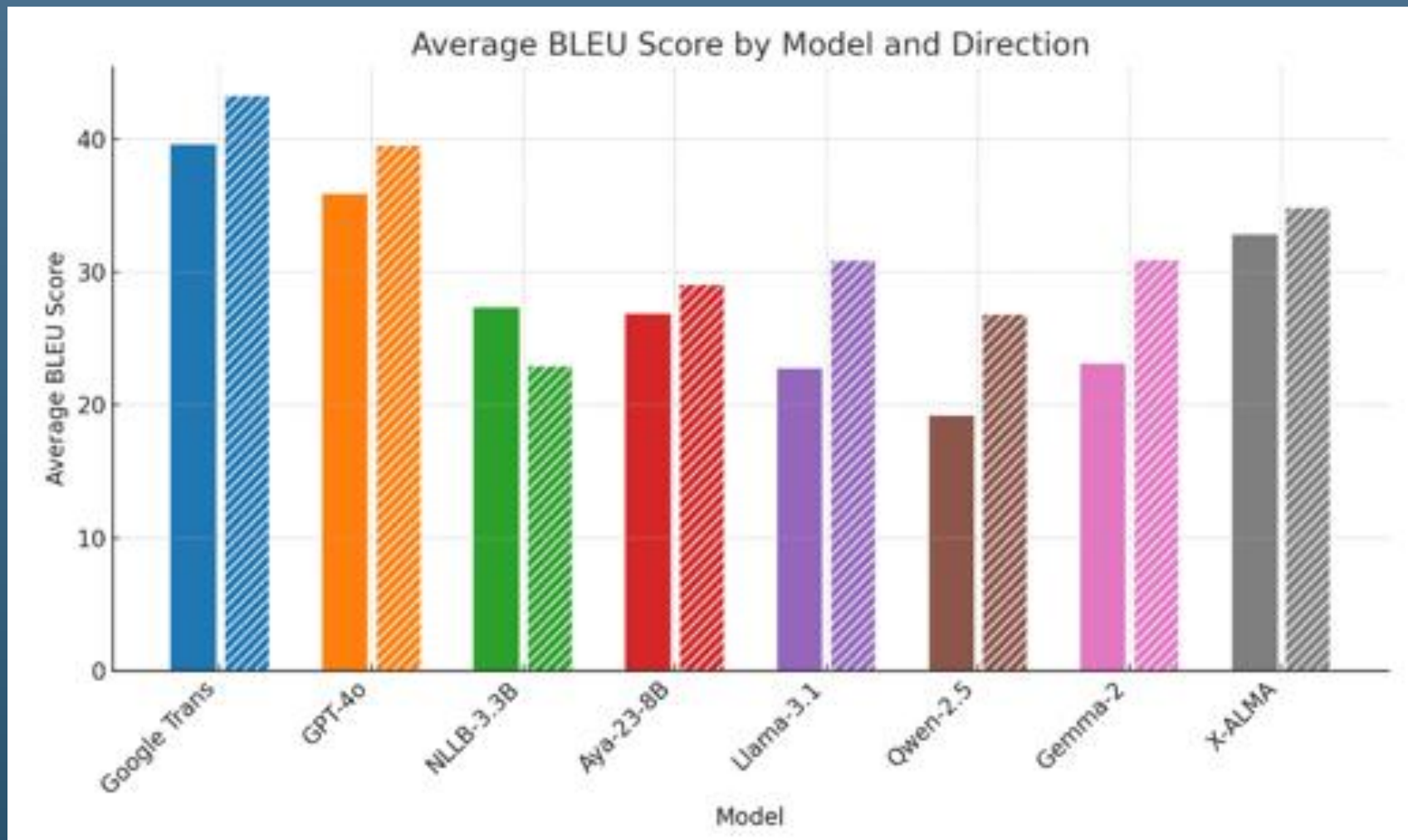
➤ 数据集(2024.10)

- WMT23: 德语(de), 中文(zh), 俄语(ru), 日语(ja), 乌克兰语(uk), 希伯来语(he)
- WMT22: 冰岛语(is)

➤ 翻译方向: $EN \rightarrow X$, $X \rightarrow EN$

	高资源	中等资源	低资源
与英语同语言族	de, ru	uk	is
与英语不同语言族	zh	ja	he

● BLEU平均得分



- Google Trans
- GPT-4o
- NLLB-3.3B
- Aya-23-8B
- Llama-3.1
- Qwen-2.5
- Gemma-2
- X-ALMA

- EN→X
- X→EN

➤ GPT-4o的语言翻译能力已与Google Translator相当。

◆ 语言翻译水平主观评价

✧ 梁茂成教授，北京航空航天大学

- ① 本科阶段设有翻译专业，而翻译专业离开扎实的语言基本功显然是不可行的。从该专业的现状看，**翻译专业本科生的翻译水平比不过机器翻译**。
- ② 研究生阶段设有翻译专业硕士(MTI, Master of Translation and Interpretation)，毕业生水平参差不齐，**很多毕业生的翻译水平比不上机器翻译**。



✧ 胡开宝教授，上海外国语大学

机器翻译(大语言模型)的译文质量超过我们培养的翻译专业的研究生，甚至博士生大部分学生翻译的水平。

◆ 中文写作水平

- 请以“AI时代的人文与科学素养培养”为题写一篇短文，不超过800字。



数字洪流中的灯塔：AI时代的人文与科学双翼

当AlphaGo击败李世石的那一刻，人类仿佛站在了一个新时代的门槛上。人工智能正以前所未有的速度重塑着我们的世界，从医疗诊断到艺术创作，从金融分析到教育革新，AI的触角已延伸至人类文明的各个角落。在这样的时代背景下，单一的知识储备已不足以应对未来的挑战，唯有同时具备深厚人文素养与扎实科学精神的双重翅膀，才能在数字洪流中保持清醒的航向。

人文素养是AI时代不可替代的精神指南针。古希腊哲学家苏格拉底曾言："未经省察的人生不值得过。"在算法主导的信息茧房中，这种省察精神尤为珍贵。当ChatGPT能够生成流畅的文章，MidJourney可以创作惊艳的画作时，人类的价值恰恰体现在那些无法被量化的领域——对生命意义的追问、对伦理困境的思考、对美的独特感知。文艺复兴时期，达芬奇既是科学家又是艺术家，这种跨界思维使他能够从解剖学研究中获得艺术创作的灵感。今天，我们同样需要这种人文视角来审视技术的边界，确保AI的发展始终服务于人类的福祉而非相反。人文素养使我们保持批判性思维，不至于在技术狂欢中迷失自我。

科学素养则是驾驭AI工具的必要能力。伽利略用望远镜挑战地心说时，不仅需要勇气，更需要严谨的实验方法和数学验证。在AI时代，理解机器学习的基本原理、知晓算法的局限性、具备数据批判意识，已成为现代公民的基本素养。当推荐系统左右我们的信息获取，当面部识别技术影响社会公平，只有具备科学素养的人才能看穿"技术中性论"的迷思，参与有关AI伦理的社会讨论。法国数学家庞加莱曾说："科学是事实构建的房屋，但事实的堆积不等于科学。"同样，会使用AI工具不等于具备科学素养，唯有理解其背后的逻辑与局限，才能避免成为技术的被动接受者。

人文与科学的融合将催生AI时代最具创造力的思维。中国古代思想家墨子既是哲学家又精通光学与力学，他的兼爱思想与科学探索相得益彰。当代科技史学者发现，许多重大科学突破往往产生于不同学科的交叉地带。当AI接手了越来越多的程式化工作，人类的优势恰恰在于能够将看似不相关的领域联系起来——用文学隐喻启发算法设计，以哲学思辨指导AI伦理框架，借艺术审美提升人机交互体验。诺贝尔物理学奖得主费曼曾将科学探索比作“观察宇宙的诗歌”，这种诗意正是人文与科学交融的典范。在AI时代，最富价值的可能是那些能够同时在两个维度思考的人。

站在文明转型的关口，我们既不能对技术革新视而不见，也不应拜倒在算法神坛之下。AI如同一面镜子，既映照出人类理性的辉煌，也暴露出我们认知的局限。培养兼具人文温度与科学严谨的新素养，不仅是为了适应时代变化，更是为了守护人之为核心的价值——在机器能计算一切的世界里，我们仍需追问何为美好生活；在算法能预测行为的年代，我们依然珍惜自由意志的闪光。让科学与人文如DNA双螺旋般交织上升，或许才是应对AI时代最智慧的态度。



● 请以杜甫的写作风格写一首反映青年人面临内卷压力的诗。



《自习室夜题》

荧屏光冷对残更，
倦眼犹追卷上名。
笔试千题争寸隙，
薪微百战困孤城。

青春渐共咖啡尽，
意气难随榜帖平。
若问此心何所似？
秋风落叶满街声。

注：仿杜甫风格解析

1. 题材现实性：

- 聚焦当代青年“内卷”困境（如加班、考编、职场竞争），延续杜甫关注社会疾苦的写实传统。

2. 语言凝练沉郁：

- “笔试千题争寸隙”浓缩考公考研的激烈竞争；
- “薪微百战困孤城”化用杜甫《征夫》“十室几人在”的悲慨，以“孤城”喻职场围城。

3. 情感递进：

- 前四句直陈现象，后四句转入抒情。“青春渐共咖啡尽”暗用杜甫《赠卫八处士》“少壮能几时”的时光之叹；
- 末句以景结情，仿《登高》“无边落木萧萧下”，用秋风落叶隐喻青年的消耗与迷茫。

4. 格律严谨：

- 平仄依七律仄起式，颌联、颈联对仗工稳（如“笔试”对“薪微”，“千题”对“百战”），体现杜甫“语不惊人死不休”的锤炼精神。

此诗尝试以古典形式承载现代议题，正如杜甫以“诗史”笔法记录盛唐崩裂，今人亦需为时代存照。



◆ 作品分析

- 请分析《三国演义》中10位核心人物之间的关系。



《三国演义》的核心人物关系错综复杂，既有君臣忠义、盟友合作，也有敌对仇怨和家族羁绊。以下是10位核心人物之间的关系分析：

1. 刘备（蜀汉）

- **与关羽、张飞**：桃园结义兄弟，情同手足，是蜀汉政权的核心奠基人。
- **与诸葛亮**：三顾茅庐请出，视为“如鱼得水”，诸葛亮终生辅佐刘备父子。
- **与曹操**：早期合作讨董卓，后因“汉贼不两立”成为死敌。
- **与孙权**：联合抗曹（赤壁之战），后因荆州问题反目（夷陵之战）。

2. 关羽（蜀汉）

- **与曹操**：曾短暂投靠并受厚待（斩颜良报恩），但最终回归刘备阵营。
- **与孙权**：拒绝联姻并辱骂孙权，导致东吴偷袭荆州，兵败被杀。

3. 张飞（蜀汉）

- **与吕布**：多次敌对（骂其"三姓家奴"），最终因虐待部下被叛将所杀。

4. 诸葛亮（蜀汉）

- **与周瑜**：赤壁之战合作，但周瑜忌惮其才能（"既生瑜，何生亮"）。
- **与司马懿**：北伐时的主要对手，多次斗智（空城计、死诸葛吓走活仲达）。

5. 曹操（曹魏）

- **与汉献帝**：挟天子以令诸侯，表面尊奉实则控制。
- **与袁绍**：早期盟友，后官渡之战以少胜多，统一北方。
- **与吕布**：曾合作又反目（"宁教我负天下人"），最终擒杀吕布。

6. 司马懿（曹魏/晋）

- **与曹丕**：深受信任，成为托孤重臣。
- **与诸葛亮**：长期对抗，拖延战术消耗蜀军。

7. 孙权（东吴）

- **与周瑜**：君臣兼挚友，周瑜助其奠定江东基业。
- **与鲁肃**：主张联刘抗曹，接替周瑜成为战略规划者。

8. 周瑜（东吴）

- **与黄盖**：君臣关系，赤壁之战用苦肉计诈降曹操。

9. 吕布（独立势力）

- **与董卓**：先为义父子，后因貂蝉反目并刺杀董卓。
- **与刘备**：曾受其收留（辕门射戟解围），后夺徐州反目。

10. 赵云（蜀汉）

- **与刘备**：忠心护卫，长坂坡救阿斗，但未被完全重用。

核心关系网络总结

1. **蜀汉集团**：以刘备为核心，关羽、张飞、诸葛亮、赵云构成"忠义"纽带。
2. **曹魏集团**：曹操通过权术驾驭群臣，司马懿蛰伏后夺权。
3. **东吴集团**：孙权依赖周瑜、鲁肃等将领维系江东。
4. **敌对关系**：
 - 魏 vs 蜀（正统之争）
 - 蜀 vs 吴（荆州之恨）
 - 魏 vs 吴（长江对峙）
5. **特殊纽带**：
 - 诸葛亮与司马懿（宿敌）
 - 周瑜与诸葛亮（既合作又嫉妒）
 - 吕布与董卓/曹操（反复无常）

① 本回答由 AI 生成，内容仅供参考，请仔细甄别。

《三国演义》的人物关系体现了"合纵连横"的权谋逻辑，既有理想化的兄弟情义（如桃园三结义），也有现实的政治利益博弈（如孙刘联盟破裂）。

◆ 大模型的艺术创作能力

请画一幅油画，展现长城脚下春暖花开的美好景象。



◆ 大模型的数学问题求解能力

- o1系统在2024年的美国(针对高中生的)数学邀请赛(American Invitational Mathematics Examination, AIME)考试题上得分为83.3(满分100分);
- 在博士级别的科学问题上(GPQA Diamond), o1的正确率达到了78%, 超越人类博士69.7%;
- 在国际信息学奥林匹克竞赛(IOI)中在每题 50 次提交的条件下取得了前49%的成绩, 在每题10,000次提交的情况下, 超过了金牌门槛;
- 编码方面, o1在竞争性编程问题(Codeforces)上的得分是89%;
- 在2024年AIME数学竞赛题目测试中, o3的准确度得分为96.7, 超过o1等。

信息来源: <https://news.qq.com/rain/a/20240924A023CH00>
<https://baijiahao.baidu.com/s?id=1819027426169963008&wfr=spider&for=pc>

◆大模型能力归纳

➤很强的对话能力

- 多轮对话的上下文理解
- 用户意图的理解能力
- 回复的完整性和逻辑一致性
- 具有指代歧义消解能力……

➤丰富的知识储备

- 常识性、百科知识
- 多领域专业知识
- 强大的数学问题求解能力
……

➤强大的生成能力

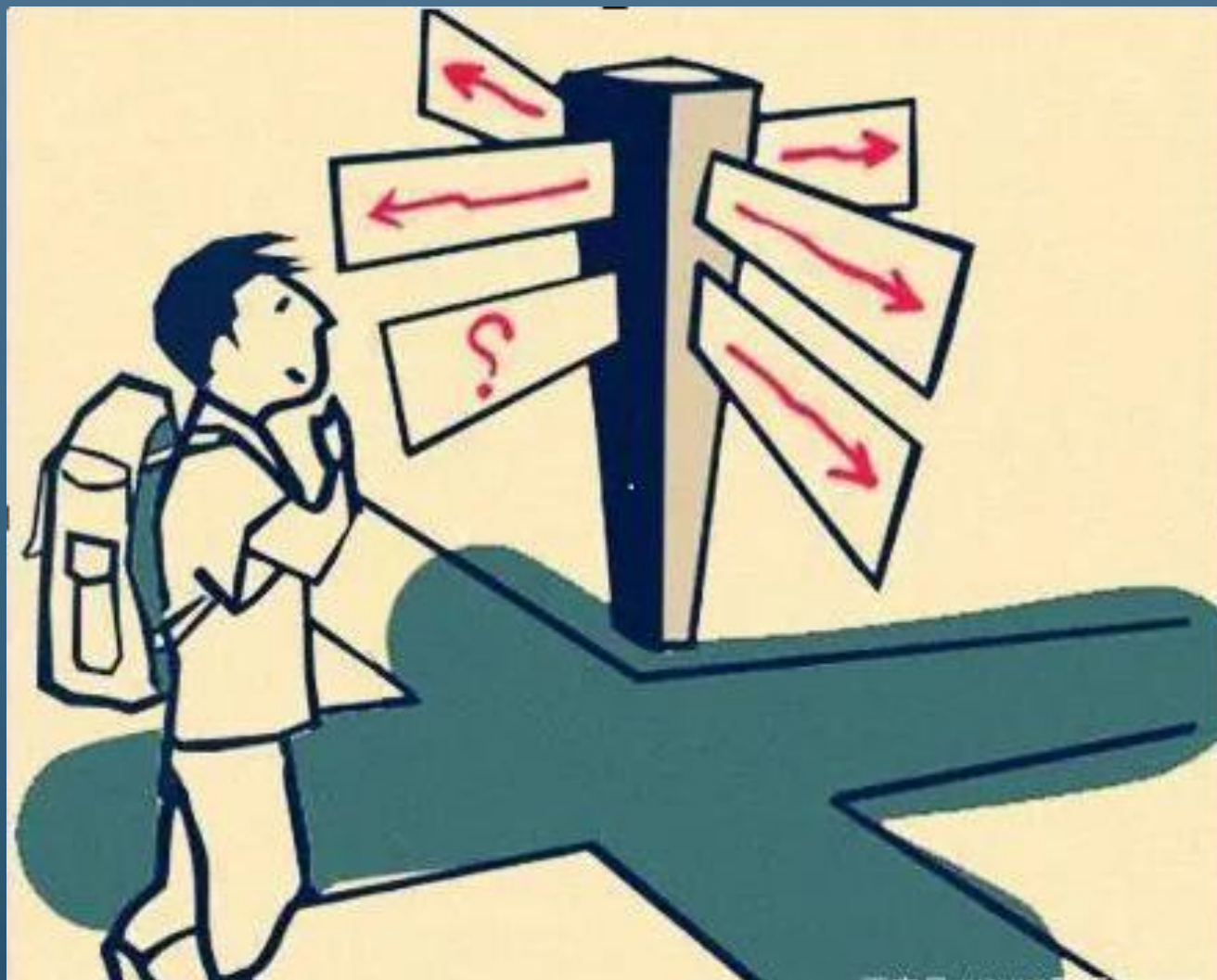
- 规范的语言表达能力
- 多风格文本写作能力
- 多语言翻译能力
- 代码生成能力
- 平面设计
- 图画/视频生成
……

●定位— AI助手

◆ 我国部分大学的语言类专业招生情况

- 中国科技大学2023年10月发布公告，撤销包括英语在内的6个本科专业的招生。
- 上海财经大学在2023年本科专业设置中停止英语专业招生。
- 西南交通大学于2022年商务英语专业停招。
- 中国传媒大学在2018年撤销了10个小语种专业。
- 华南农业大学从2020年开始商务英语本科专业暂停招生。
- 中山大学2022年撤销了包含商务英语在内的8个专业。
- 济南大学2023年停招了包括**汉语国际教育**、国际经济与贸易、投资学等多个专业，外语专业的调整也在其中。

有待核实



To be or not to be,
this is a question.

活着，还是不活，这是个问题。

提 纲

1. 大模型概要

2. 大模型的能力分析

➡ 3. 中文信息处理

- ◆ 为什么研究中文信息处理？
- ◆ 大模型不是已经实现了多语言处理吗？
- ◆ 中文信息处理中还有哪些问题需要解决？

◆ 为什么研究中文信息处理？

- 解释语言现象、认识语言规律……
- 语言教学
- 应用技术研发

<https://image.baidu.com/search> (全球使用汉语的人数)

全球语言使用者 十大排行榜			
排名	国家	语言	使用者
1		英语	20亿使用者
2		汉语	15亿使用者
3		印地语	8亿使用者
4		西班牙语	5亿使用者
5		阿拉伯语	2.7亿使用者
6		俄语	2.6亿使用者
7		葡萄牙语	2.1亿使用者
8		法语	2亿使用者
9		意大利语	1.9亿使用者
10		日语	1.3亿使用者

◆大模型不是已经实现了多语言处理吗？

NW

你知道“林黛玉倒拔垂杨柳”的故事吗？



是的，我了解“林黛玉倒拔垂杨柳”的故事。这个故事出自中国古典文学名著《红楼梦》。

《红楼梦》是清代作家曹雪芹创作的长篇小说，被誉为中国四大名著之一。故事讲述了贾宝玉、林黛玉、薛宝钗等人的爱情故事，以及贾家荣华富贵和衰落的历程。

在《红楼梦》中，林黛玉是贾宝玉的表妹，是书中主要女主角之一。在故事中，林黛玉在园中倒拔垂杨柳的情节发生在贾府的花园里。她因受到一些事情的打击和伤害，心情非常沮丧和痛苦，于是决定在花园里倒拔垂杨柳来发泄情绪。

垂杨柳是柔弱的树种，它的形象往往被用来比喻柔弱而忧愁的女性。林黛玉倒拔垂杨柳的场景，表现了她内心的忧愁和悲伤，也反映了她在贾府中处境的无奈和困苦。

这个情节也成为《红楼梦》中的一个重要象征，用来表达人物的心情和命运，以及寓意着人生的起伏和变化。

◆ 中文信息处理中还有哪些问题需要解决？

● 语料库建设

国家语料库的建设、开发、保护应该是一种国家行为，在信息社会和数字化生存时代，我们要把语言资源的收集、保护、开发提高到一种对待国家资源的高度来认识。国家要像对待人力资源、地矿资源、国土资源、森林资源、水源资源一样对待语言资源，语言资源是国家最重要的信息资源。语料库的建设、保护、开发要站在国家面向未来的一种战略决策高度，要作为一种对待国家资源的行为，才能得到法律的保护，纳入法制的轨道 [张普，2003]。

ChatGPT的训练数据中中文语料比重不足千分之一，而英文语料占比超过92.6%。

(来自以下网站：<https://caifuhao.eastmoney.com/news/20240311154459454112270>)

● 多语言术语对照

我国已经与**152个**国家、**32个**国际组织签署了合作文件，涉及**110多种**语言。

(截止2023年6月)



● 相关规范 and 标准

➤ UPenn Treebank 的词性标注集

✧ 33 类: NN 名词、NR 专业名词、NT 时间名词、VA 可做谓语的形容词、VC “是”、VE “有” 作为主要动词、VV 其他动词、AD 副词、M 量词，等等。

➤ 北大计算语言学研究所的词性标注集

✧ 26个基本词类代码，74个扩充代码，标记集中共有106个代码: 名词(n)、时间词(t)、处所词(s)、方位词(f)、数词(m)、量词(q)、区别词(b)、代词(r)、动词(v)、形容词(a)、状态词(z)、副词(d)、介词(p)、连词(c)、助词(u)、语气词(y)、叹词(e)、拟声词(o)、成语(i)、习用语(l)、简称(j)、前接成分(h)、后接成分(k)、语素(g)、非语素字(x)、标点符号(w)。

● 中文大模型评价方法

● 借助大模型等AI技术的中文信息技术应用

- 国际中文教学
- 中华优秀传统文化传播和保护
- 甲骨文破解
- 话语体系建设
- 人文素养培养
-

人文 + 科技 → 社会发展的两大支柱



● 学好中文、很好地使用中文意义重大

苏步青简介

(1902.11~2003.3)



学术上的成就：中国杰出的数学家

(主要从事微分几何学和计算几何学等方面的研究。他在仿射微分几何学和射影微分几何学研究方面取得出色成果，在一般空间微分几何学、高维空间共轭理论、几何外型设计、计算机辅助几何设计等方面取得突出成就。1978年获全国科学大会奖。)

“如果允许复旦大学单独招生，我的意见是第一堂先考语文，考后就判卷子。不合格的，以下的功课就不要考了。语文你都不行，别的是学不通的。”

(https://news.sohu.com/a/819301827_122053982)

◆ 目前语言类学科存在的问题

- ① 学科划分过细，而应用需求错综复杂
- ② 学科理念有待改变
- ③ 教师队伍结构单一，跨学科交流成效不大
- ④ 教材体系不够完善，教学内容未能与时俱进
- ⑤ 教学平台金玉其外，科研设备亟待升级
- ⑥ 缺少合理有效的语言资源共享机制和平台
- ⑦ 考核机制不合理

● 面向海上丝绸之路的区域国别研究

◇ 自然语言处理

- 机器翻译
- 文本分类/聚类
- 情感分析
- 自动文摘
- 文本推理
- 自动写作……

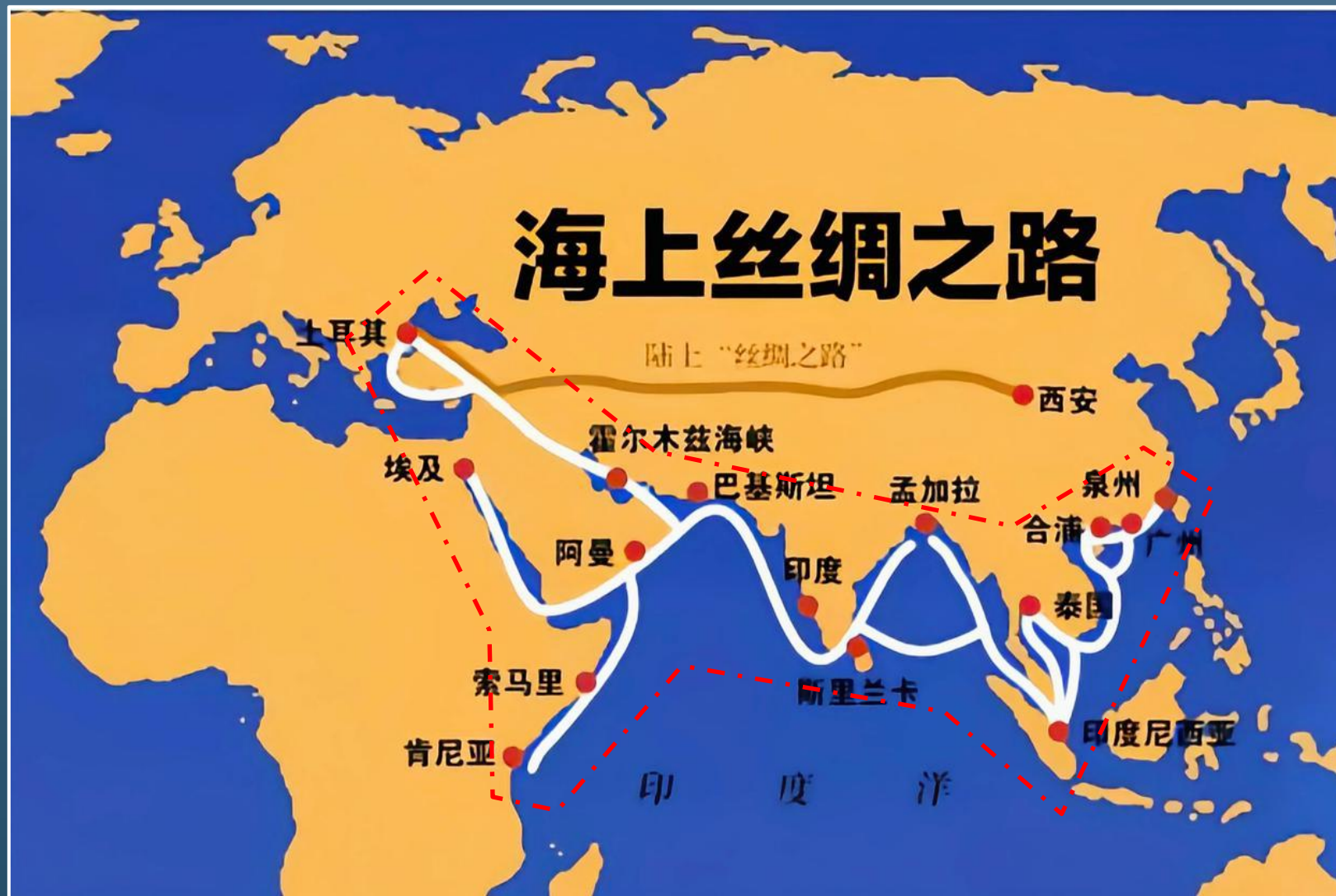
LLMs

◇ 国际政治、国家政策

◇ 地理、环境

◇ 历史、文化

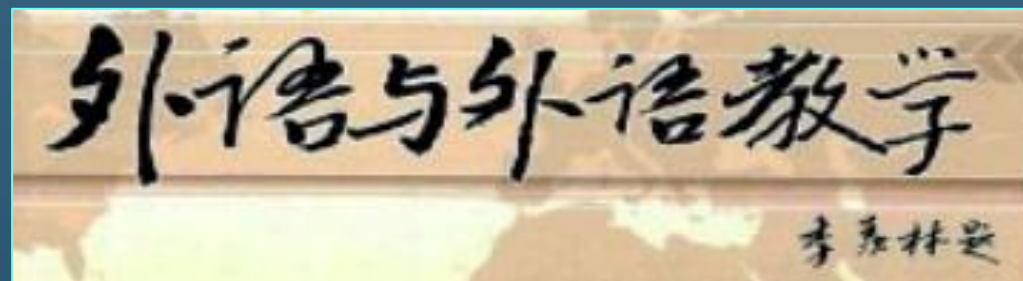
◇ 经济、商贸、旅游



◆ 六点建议

- ① 开设人工智能通识课，打通学科壁垒 → AI 基础课程 + X(任何学科)
- ② 语言专业兼修其他专业课程 → 中文 + Y(其他专业课程)
- ③ 科研选题结合国家需求
- ④ 利用新技术手段建立特色教学平台
- ⑤ 建设开放共享的语言资源平台
- ⑥ 建立“代表作”引导的考核机制

例如，外交学/国际政治/国际公共关系；经济法/诉讼法/民商法；税收学/公司金融学/国际金融学…



宗成庆：从人工智能角度谈语言类学科发展和人才培养，2024(5)

结束语

- AI技术的快速发展对各学科领域、各行各业的冲击是不可逆转的，拥抱新技术、学习新技术，势在必行；
- 大模型等AI技术不会替代中文学习和使用者，但会替代不会使用AI技术的人；
- 中文是中华文明的守护者，学习中文、保护中文，在社会进步中提升人文素养，弘扬和传承中华优秀传统文化，是每一位中华儿女的责任和使命。研究和开发中文信息处理技术在大模型时代仍具有非常重要的意义。

谢谢!
Thanks!

