

大模型时代的跨语言文化交流

Translingual and Transcultural Communication in Era of Large Models

宗成庆(Chengqing Zong)

中国科学院自动化研究所

(Institute of Automation, Chinese Academy of Sciences)

cqzong@nlpr.ia.ac.cn



什么是大语言模型？

What is the Large Language Model?

◆ 语言是思维的载体(Language is the carrier of thought.)

NW

请介绍一下马来亚大学。(Please introduce the University of Malaysia.)



马来西亚拥有许多优秀的大学和高等教育机构，其中一所著名的大学是马来亚大学（University of Malaya，简称UM）。以下是有关马来亚大学的一些基本信息：

- 1.历史与背景：**马来亚大学成立于1949年，是马来西亚最古老、最具声誉的大学之一。它位于马来亚吉隆坡市中心，校园环境优美，拥有先进的教育设施和研究设备。
- 2.学术声誉：**马来亚大学在国际上享有很高的声誉，被广泛认为是亚洲一流大学之一。它在许多国际排名中常常位列前茅，尤其在亚洲地区排名中表现出色。
- 3.学科领域：**马来亚大学涵盖了广泛的学科领域，包括人文科学、社会科学、自然科学、工程技术、医药健康等。它在许多学科领域都有卓越的研究和教育成果。
- 4.研究：**马来亚大学在研究方面表现出色，积极参与各种国际合作项目和研究合作。学校设有许多研究中心和实验室，致力于推动科学研究的进展。



什么是大语言模型？

What is the Large Language Model?

5. **国际化：** 马来亚大学非常重视国际交流与合作。学校吸引了来自世界各地的学生和教师，创造了多元化的学术环境。此外，学校也积极参与国际学术合作项目，推动国际化教育。
6. **社会影响：** 马来亚大学在培养人才、推动社会变革和发展方面扮演着重要角色。其毕业生在各行各业中都有出色的表现，为马来西亚以及国际社会的发展作出了贡献。

总之，马来亚大学是一所在马来西亚乃至国际上具有重要地位和影响力的优秀大学，以其卓越的教育和研究质量为人们所熟知。如今，它继续在教育、科研和国际合作方面发挥着重要作用。请注意，我所掌握的知识截止日为2021年9月，因此一些信息可能已经发生变化。如需获取最新信息，请访问马来亚大学的官方网站或其他权威来源。

The system gives a very comprehensive and detailed introduction.



什么是大语言模型？

What is the Large Language Model?

NW 马来西亚常用的问候语有哪些？(What are the commonly used greetings in Malaysia?)

 马来西亚是一个多元文化的国家，因此人们使用各种语言和问候语来交流。以下是一些常见的马来西亚常用问候语，涵盖了马来语、华语和英语：

马来语(Malay):

1. **Selamat pagi:** 早上好。
2. **Selamat tengah hari:** 中午好。
3. **Selamat petang:** 傍晚好。
4. **Selamat malam:** 晚上好。
5. **Apa khabar?** 你好吗？（用于问候“你好吗？”）
6. **Salam sejahtera:** 安宁的祝福。
7. **Assalamualaikum:** 伊斯兰教中的问候，意为“愿和平与真主的怜悯与祝福归于你”。



什么是大语言模型？

What is the Large Language Model?

华语(Chinese):

1. 你好 (Nǐ hǎo) : 通用的“你好”问候。
2. 早上好 (Zǎo shàng hǎo) : 早上好。
3. 下午好 (Xià wǔ hǎo) : 下午好。
4. 晚上好 (Wǎn shàng hǎo) : 晚上好。
5. 吃了吗? (Chī le ma?) : 用于问候“你吃了吗?” (常见的社交问候, 表示关心对方)。

英语(English):

1. Hello: 你好。
2. Good morning: 早上好。
3. Good afternoon: 下午好。
4. Good evening: 晚上好。
5. How are you?: 你好吗?
6. What's up?: 有什么新鲜事吗? (非正式问候)

以上只是一些常见的问候语, 实际上, 马来西亚的人们会使用多种语言和问候方式, 根据对方的文化背景和关系选择合适的问候方式。无论何种语言, 友好的问候在马来西亚文化中都是很重要的社交礼仪。



什么是大语言模型？

What is the Large Language Model?

NW 请推荐一下泰国的美食。(Please recommend Thai food.)



当谈到泰国美食时，有许多令人垂涎的菜肴可以品尝。以下是一些受欢迎的泰国美食：

1. **泰式绿咖喱 (Green Curry)**：这是一道以椰浆为基础，加入青椒、茄子、香茅、姜和其他香料烹制的咖喱。它的味道浓郁、辛辣而香甜。
2. **泰式炒河粉 (Pad Thai)**：这是一道著名的泰国炒面，通常含有蛋、豆芽、花生和酸柑。它的味道酸甜微辣，是泰国餐桌上的常见选择。
3. **泰式烤鸡串 (Satay)**：这是一道美味的烤肉串，通常用鸡肉，蘸有花生酱。它的味道香浓，适合作为开胃小吃。
4. **冬荫功汤 (Tom Yum Goong)**：这是一道酸辣的虾汤，通常加入香茅、姜、青柠叶和辣椒。它有浓郁的香料味道，是泰国餐桌上的经典汤品。
5. **泰式凉拌生菜包 (Larb)**：这是一道以肉类（通常是鸡肉、牛肉或猪肉）为基础的凉拌菜肴，加入薄荷叶、洋葱、酱汁和香料，味道鲜美且爽口。



什么是大语言模型？

What is the Large Language Model?

6. **泰式椰奶甜品 (Sticky Rice with Mango)**：这是一道甜点，将糯米饭搭配新鲜芒果和椰奶，口感丰富，甜而不腻。
7. **泰式炸春卷 (Spring Rolls)**：这是一道香脆的炸春卷，通常以蔬菜、虾和米粉为主要成分，蘸以甜辣酱汁。
8. **泰式烤鱼 (Pla Pao)**：这是一道整条鱼烤制而成的菜肴，通常涂抹有香草和调味料，具有独特的烟熏味道。
9. **绿木瓜沙拉 (Som Tum)**：这是一道用青木瓜制作的凉拌沙拉，加入花生、虾米、辣椒和酱汁，味道酸辣开胃。
10. **泰式炒螺蛳 (Pad Kee Mao)**：也被称为“炸醉虾”，是一道辛辣的炒面，通常含有螺蛳、蔬菜和辣椒。

这只是泰国美食中的一小部分，还有许多其他美味的菜肴等待您去品尝。无论您喜欢辛辣、酸甜还是浓郁的风味，泰国菜都能满足您的味蕾。

The recommended food list is very popular and useful.



什么是大语言模型？

What is the Large Language Model?

- 这是一个人机对话系统 — ChatGPT (OpenAI, 2022)
(It is a human-computer dialogue system)
- 拥有丰富的知识和强大的分析能力
(The system has a wealth of knowledge and powerful analysis capabilities.)
- 具有很强的对话能力 (It has great conversational skills.)
- 具有很强的通用性 (It has a strong versatility)
 - 百科知识问答 (Q&A on Wikipedia knowledge)
 - 数学问题求解 (Solve math problems)
 - 编程能力 (Programming)
 - 可以处理100多种语言 (Handle over 100 languages)
 -

是一个得力的AI助手 (It is a great AI assistant.)



什么是大语言模型？

What is the Large Language Model?

◆ 语言模型(language model) – 1980s

2023年 阳春 …… ?

context $p(\text{word}|\text{context})$ 三月/白雪/面……

2023年 阳春 三月 春光明媚, ……

一种语言的句子 $\Longrightarrow$ 另外一种语言的句子（翻译）
Given a sentence A sentence in another language (translation)

一个输入问题 $\Longrightarrow$ 一种问题的回答（问答）
Given a question An answer (Q&A)

一个输入语句 $\Longrightarrow$ 一个回复（对话/聊天）
Given an utterance A response (dialogue system)

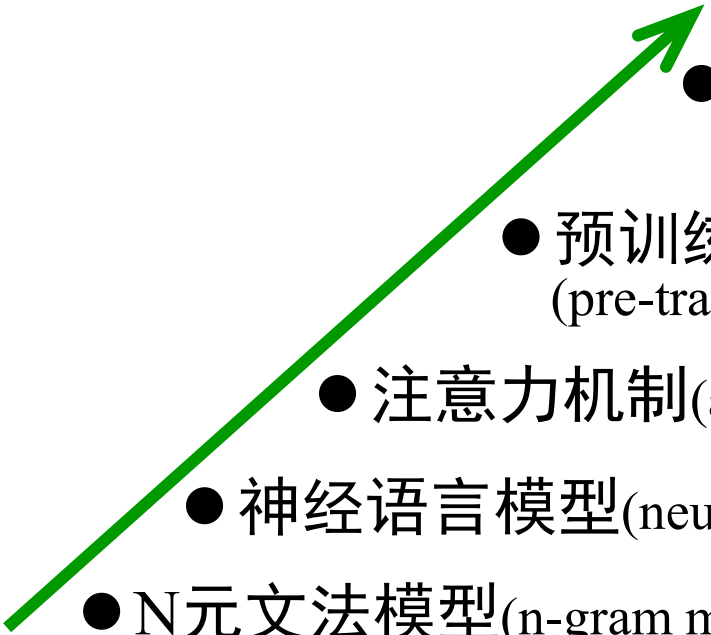


什么是大语言模型？

What is the Large Language Model?

大数据 + 大算力 + 大量人工 → 大工程

Big data + Big computing power + Large amount of labor → Big project

- 
- 大语言模型
(large language model, LLM) -2020
 - 预训练语言模型
(pre-training language model, PLM) -2018
 - 注意力机制(attention) -2017
 - 神经网络模型(neural network based language model) -2003
 - N元文法模型(n-gram model) -1980S

语言生成/自动写作
(language generation /automatic writing)

生成式语言模型
(generative language model)

人工智能生成内容
(Artificial Intelligence
Generated Content, AIGC)

但是(However), ...

NW

鲁迅和周树人是同一个人吗？

Are LU Xun and ZHOU Shuren the same person?



不，鲁迅和周树人不是一个人。鲁迅（原名周樟寿）和周树人是两位中国现代著名的作家，他们都有重要的文学成就，但是他们是两个不同的人。

No, LU Xun and ZHOU Shuren are not the same person ...

鲁迅（1881年9月25日 - 1936年10月19日）是中国现代文学的先驱之一，他是一位伟大的思想家、文学家和革命家，被誉为“百年不遇之才”。他的代表作品包括《呐喊》、《狂人日记》、《阿Q正传》等，通过他的作品，鲁迅揭露了当时社会的黑暗和尖锐批判了封建传统。

周树人（1896年1月14日 - 1967年11月13日），原名周樟寿（与鲁迅同名），后改名周树人以示区别。他是中国现代作家、文学批评家和教育家，是现代白话文运动的代表人物之一。周树人的代表作品有《封神演义》、《茶馆》等，他在文学上提倡运用白话文进行写作，推动了现代文学的发展。

虽然鲁迅和周树人的原名都是周樟寿，但是他们是两个不同的个体，分别有独立的文学成就和不同的文学风格。



但是(However), ...

NW

你知道“林黛玉倒拔垂杨柳”的故事吗？

Do you know the story of “LIN Daiyu pulls out the weeping willow upside down” ?



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》。

Yes, I know. This story comes from the famous Chinese classical literature “Dream of Red Mansions”...

《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。



，林黛玉是贾宝玉的表妹，是书中主要女主角之一。情节发生在贾府的花园里。她因受到一些事情的打击，在花园里倒拔垂杨柳来发泄情绪。

树种，它的形象往往被用来比喻柔弱而忧愁的女性。心的忧愁和悲伤，也反映了她在贾府中处境的无奈。

《红楼梦》中的一个重要象征，用来表达人物的心理。





很多其它不良表现……

Many other bad performance……

- 特定领域的专业知识匮乏 (Lack of expertise in specific areas)
- 可解释性差 (Poor interpretability)
- 由于对齐而产生人为倾向性
(Due to the artificial tendency generated by the alignment, it cannot deal with the problem of multiple values.)
- 存在“幻觉”和事实性错误等各种不良表现
(Various undesirable manifestations such as frequent “hallucinations” and factual errors)



它是一个好的基座模型吗？

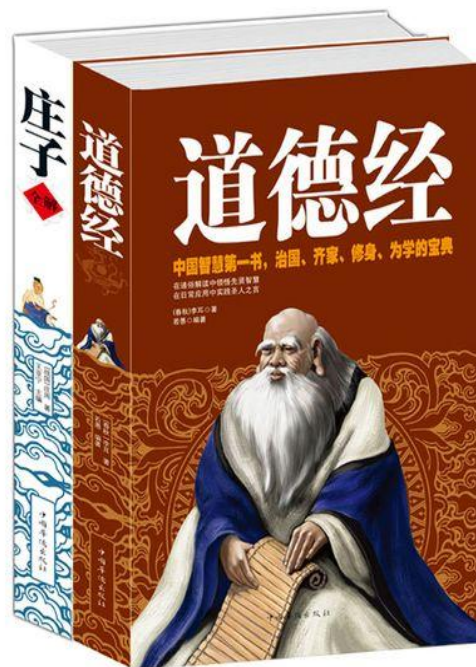
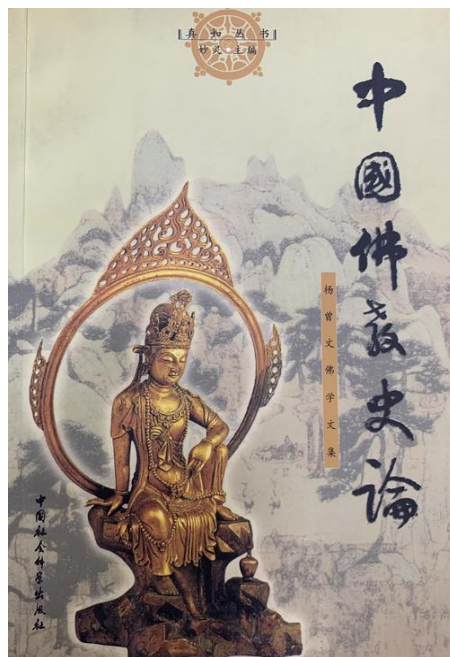
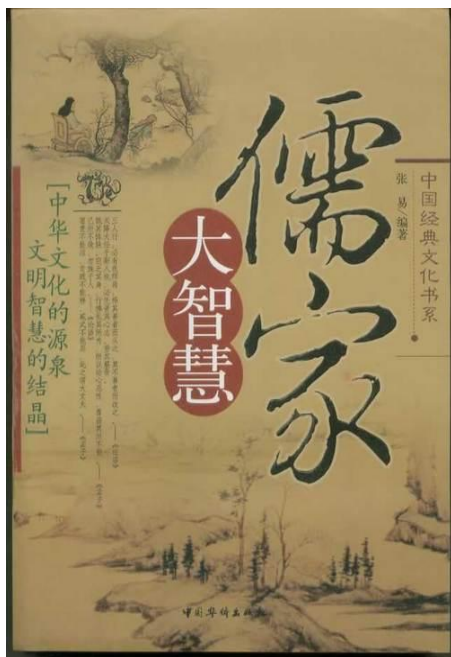
Is ChatGPT a good foundation model?

- 是的，因为它在多数任务上都具有很好的性能表现。
Yes, because it performs well on most tasks.
- 不是，因为它在很多情况下会给出错误的答案。
No, because it often gives incorrect answers.
- 针对跨语言文化交流我们应该如何做？
What should we do for translingual and transcultural communication?
 - ① 做好文化资源的数字化保护工作 (Digitizing cultural resources)
 - ② 建立基础文化知识库
(Construction of fundamental cultural knowledge base/graph)
 - ③ 构建基础大模型 (Building the foundation large language model)

传统文化的数字化保护

Digitizing cultural resources

- 儒家、佛家、道家、墨家、法家等各家思想文化著作
(Confucianism, Buddhism, Taoism, Mohism, Legalism and other ideological and cultural works)



.....

传统文化的数字化保护

Digitizing cultural resources

- 琴棋书画 (Qin, chess, calligraphy and painting)
- 文学作品 (Literature)





传统文化的数字化保护

Digitizing cultural resources

- 传统节日 (Traditional festival)
- 中国戏曲 (Chinese opera)
- 中国建筑 (Chinese architecture)
- 语言文字 (Chinese characters and languages)
- 传统中医 (Traditional Chinese medicine)
- 民间工艺 (Folk crafts)
- 各少数民族文化 (Chinese minority cultures)

.....



✧ 数字化记载 (Digital recording)

— 文本、语音、图像和视频等多媒体技术

(Multimedia technologies such as text, speech, images, and videos)



基础文化知识库建立

Construction of fundamental cultural knowledge base

✧ 作品知识库 (Knowledge bases/ graph of works)

- 作者相关信息 (Author related information)
- 时代背景 (Historical background)
- 基本思想/ 故事梗概 (Basic idea/ story outline)
- 实体关系 (Entity-Relationship)

.....

✧ 多语言术语对照表 (Multilingual glossary)

.....



基础大模型构建

Building the foundation large language model

- ✧ 人机交互对话系统 (Human-computer dialogue system)
- ✧ 问答系统 (Question-Answering)
- ✧ 面向东盟各国语言的机器翻译
(Machine translation for ASEAN languages)
- ✧ 外语学习与教学 (Foreign language learning and teaching)
- ✧ 文化作品制作 (Production of cultural works)
-



跨语言文化交流合作共赢

Translingual and transcultural communication and cooperation for mutual benefit

◆ 跨语言文化交流是时代赋予我们的使命

Translingual and transcultural communication is a mission entrusted to us by the times.

◆ 人工智能研究范式正步入大模型时代，让大模型赋能跨语言文化交流的时机已经到来

The research paradigm of artificial intelligence is entering the era of large models, and the time has come to empower translingual and transcultural communication through large language models.

◆ 我和我团队的全体成员愿意与东盟各国的语言文化工作者、教育家和科学家合作，为中国-东盟跨语言文化交流贡献我们的智慧和力量

All members of my team and myself are willing to collaborate with language and cultural workers, educators, and scientists from ASEAN countries to contribute our wisdom and strength to the translingual and transcultural communication between China and ASEAN.

谢谢!
ຂອບໃຈ
Salamat
terima kasih
Cảm ơn
ကျေးဇူးတင်ပါတယ်။
ຂອບໃຈ

