

个性化大模型技术实践

空崖（高星）@阿里巴巴达摩院自然语言智能实验室

DataFunSummit # 2023

目录 CONTENT

01 个性化大模型的简介

02 知识与个性化大模型

03 个性化大模型中的个性

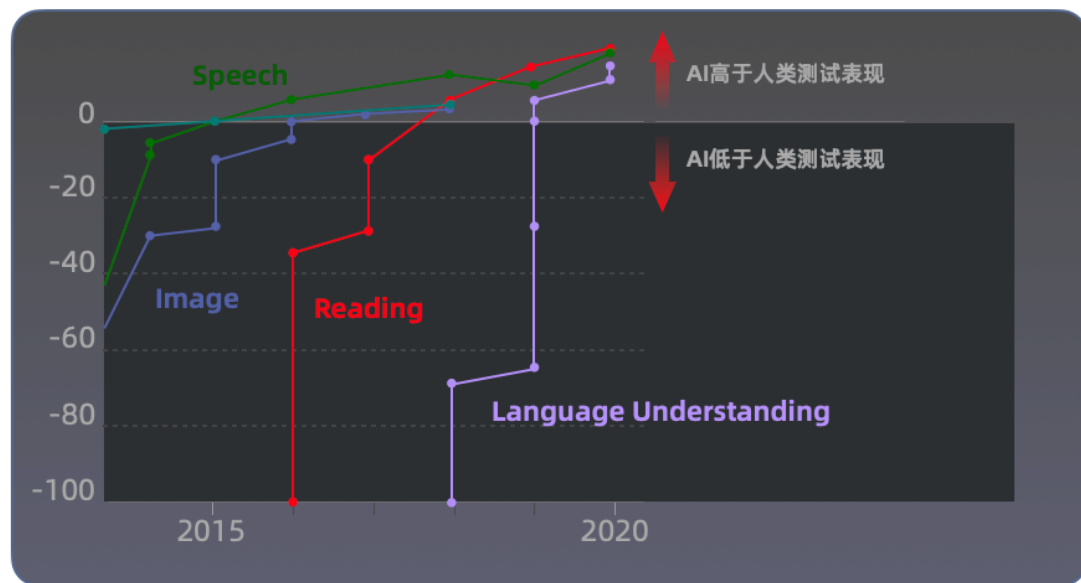
04 总结与展望

01

个性化大模型简介

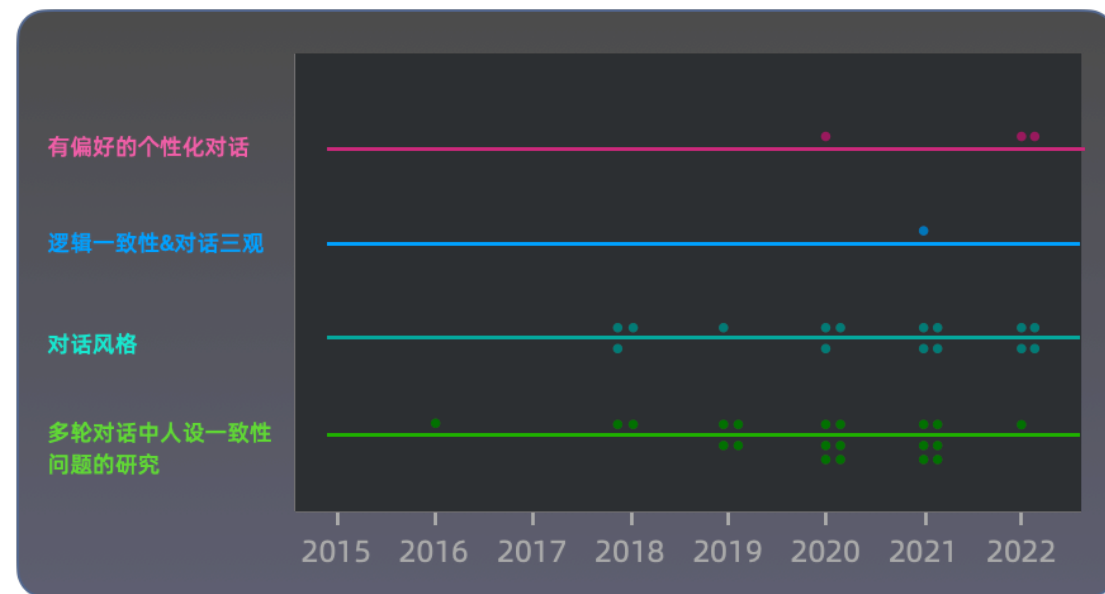
DataFunSummit # 2023

大模型个性化是个必然



通识性

- 大模型的通用性、生成性大家的感受很显著，总体上在超越人类的水平；
- 目前主流的人工智能测评集，从GPT4开始，已经能在标准化考试大幅超过人类平均水平，包括从业执照考试、高等级测试竞赛



个性化

- 在近年GPT相关论文中，Personalized Learning, Personalized Recepte, Personalization, 是有一定频率和明确指向的词汇、议题；
- 从应用属性上，搜索、生产力工具是比较公域的体验，在家居、随身设备中，还有很多需求、挑战和想象空间。

个性化大模型：四位一体的对话智能体



- 「知识、个性、情感、记忆」四位一体的类人对话智能体

知识

丰富和实时的知识储备，实现开放域知识对话

个性

具有不同人设、自定义的对话风格

情感

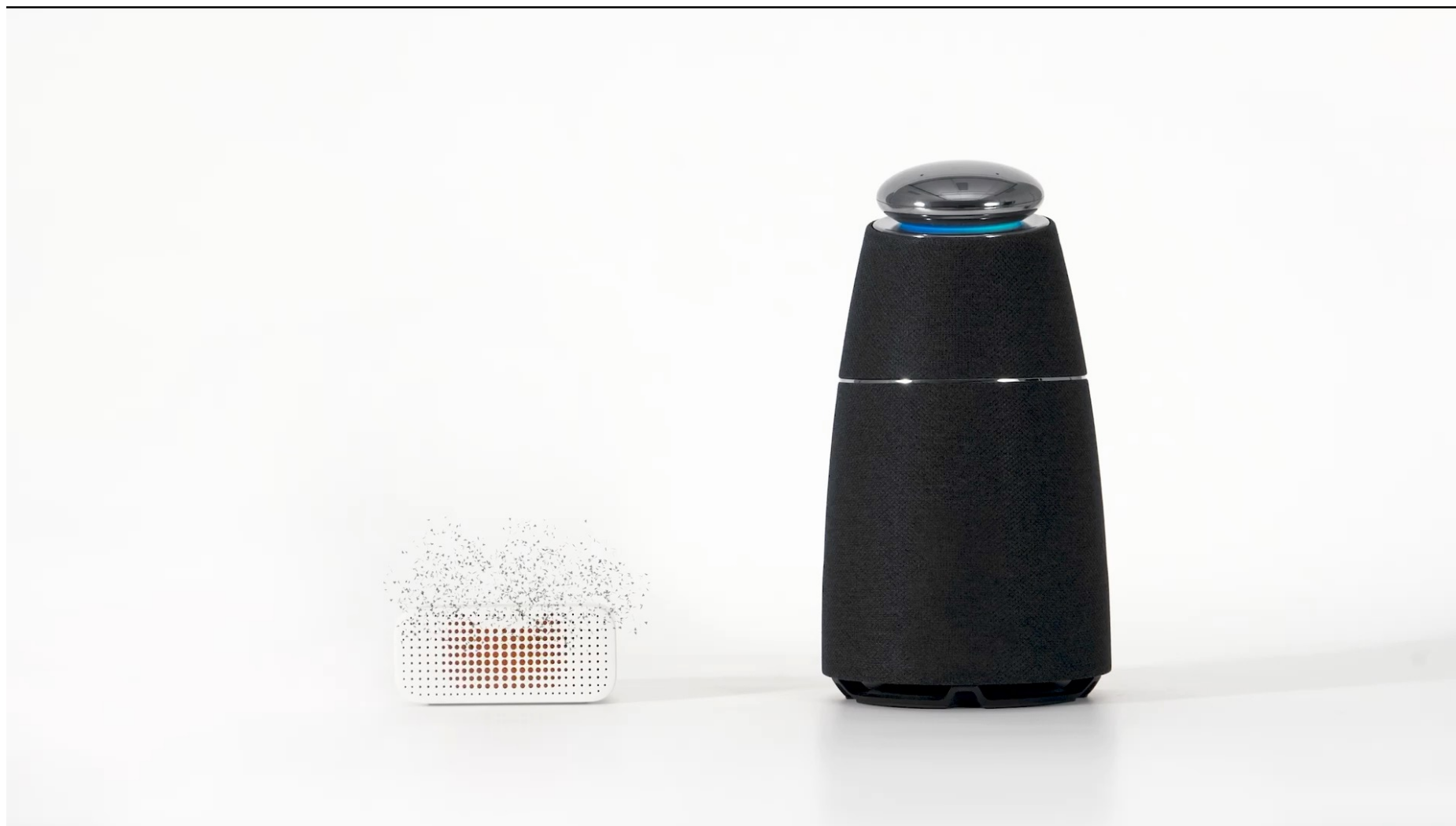
具备有温度、共情的对话能力

记忆

通过长期记忆、短期记忆的融合，记住用户偏好

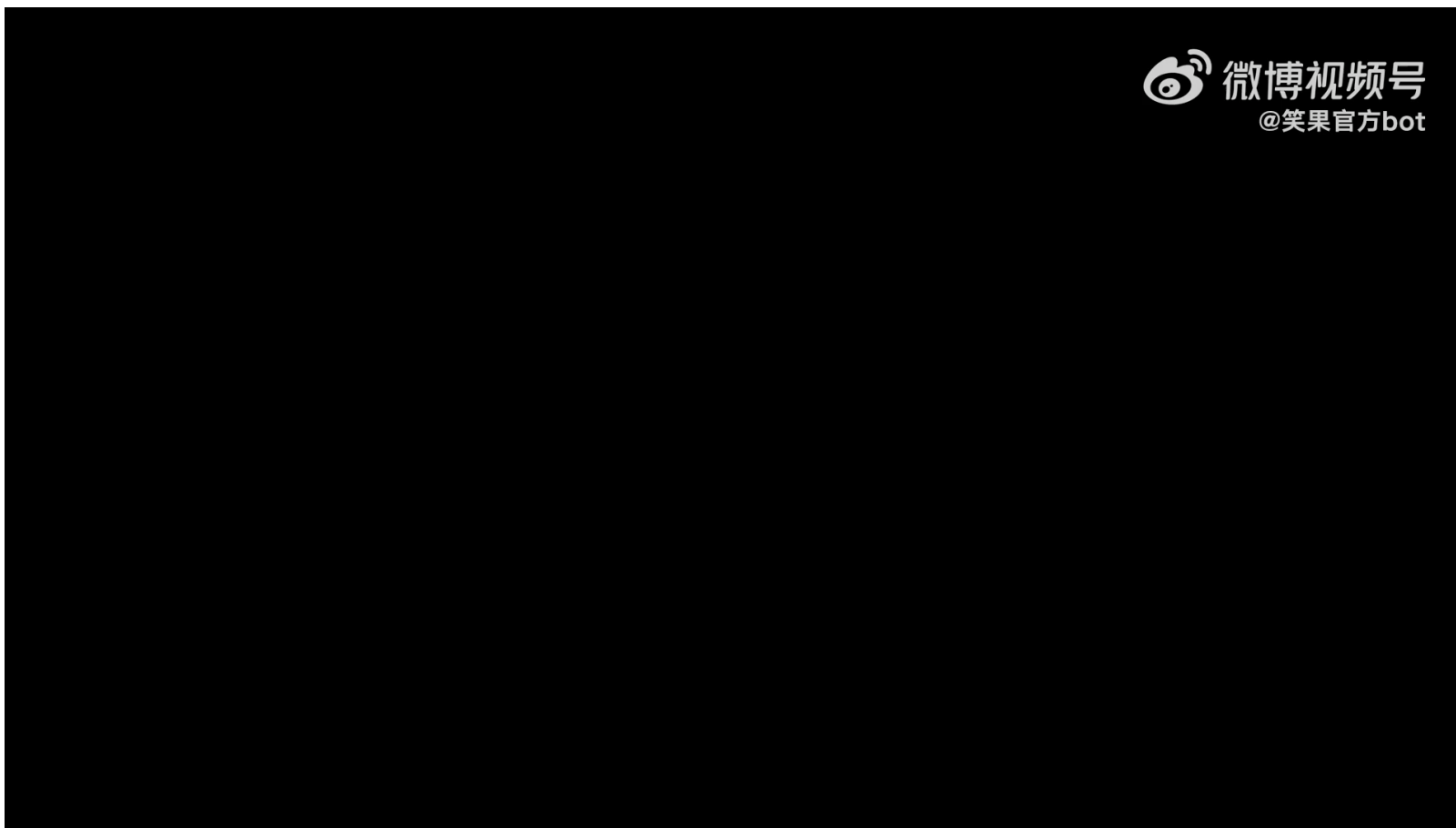
核心研究课题	关键词
Open-domain Dialogue & Generative Models	Large Language Model、Knowledge、Empathetic、Long-term Memory、Persona、Contradictions & Consistent、Style、Human Evaluation
Open-Domain Question Answering & Knowledge Grounded Dialogue	Knowledge Hallucination、Retrieval Augmented、Internet-Augmented、Reasoning
Well-Behaved	Alignment、Safety、Gender Bias
Visually Grounded & Environment Grounded	Agent、Embodied AI

天猫精灵上的应用



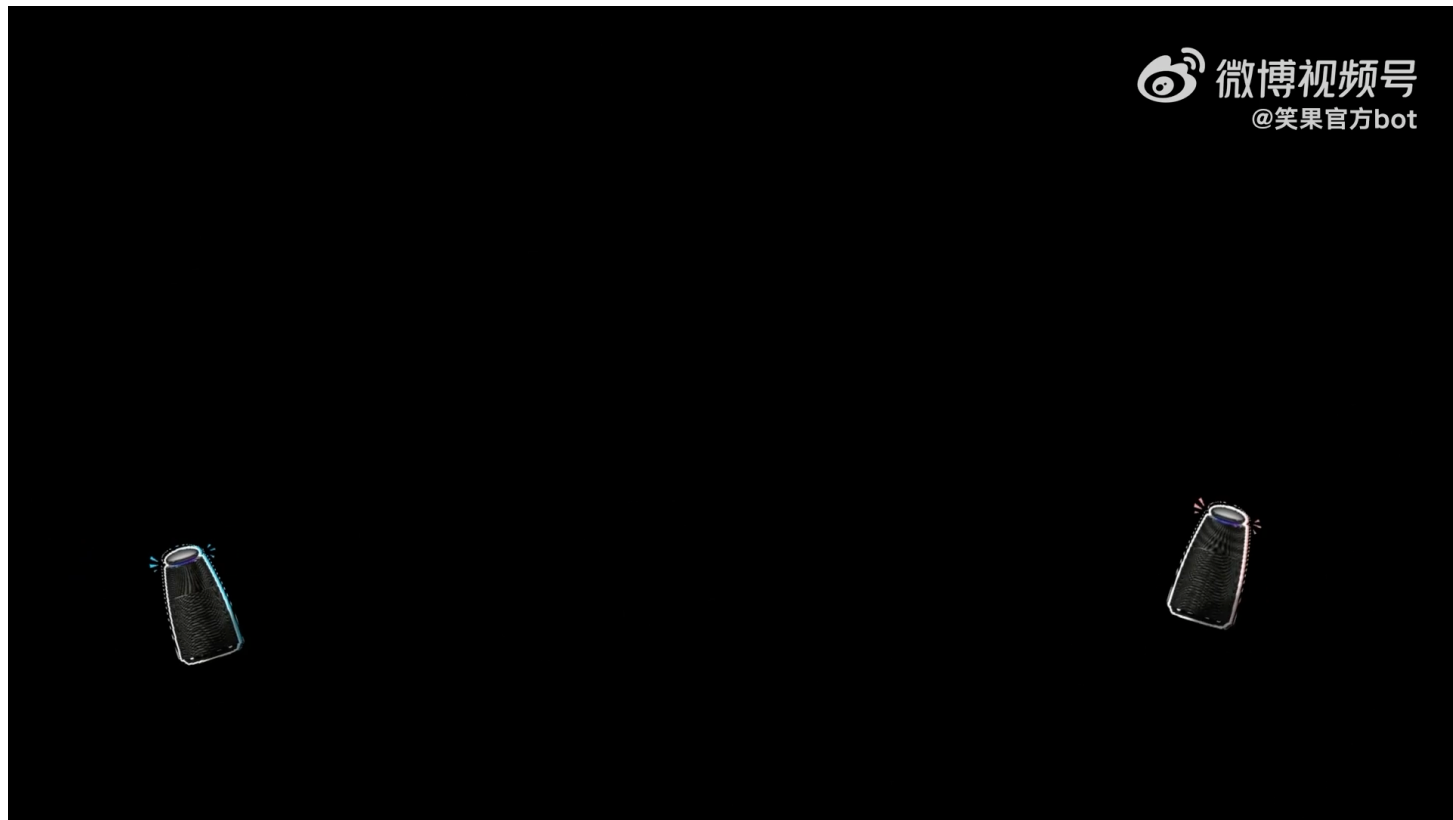
「知识、个性、情感、记忆」四位一体的对话

天猫精灵上的应用-鸟鸟分鸟



- “鸟鸟分鸟” 个性化大模型体验超预期，AIoT 行业有望迎来 “寒武纪大爆发” -浙商证券，
https://pdf.dfcfw.com/pdf/H3_AP202304061585175117_1.pdf?1680787493000.pdf

天猫精灵上的应用-鸟鸟分鸟



鸟鸟与鸟鸟分鸟的对话—类比创作



黑灯与鸟鸟分鸟对话—类比创作

更多角色扮演能力体验



Modelscope创空间体验：搜索 “角色对话”

https://www.modelscope.cn/studios/damo/role_play_chat/summary

选择角色

☒ 爱用emoji的萌妹子小婉

☐ 富有智慧的得道高僧

☐ 会说古文的三国NPC关羽

☐ 苦口婆心催婚的母亲

☐ 听从主人指令的小泰迪

☐ 感情热烈的美妆博主木子

☐ 性格洒脱爱喝酒的醉汉

☐ 温柔细心的心理咨询师

☐ 严厉不苟言笑的师尊

☐ 哀怨的迷恋唐僧的白骨精

Chatbot

小婉，你好

👋~可爱的小哥哥~💕💕
我很想和你一起出去逛街啊😄！或者一起去吃好吃的🍔，看电影🎬也可以哦！不过前提是你也要有意向哦❤️。
期待我们的下一次见面！！😄

你都喜欢吃啥

👀我是个小吃货啦！我最喜欢的就是炸鸡和薯条啦😄！我也会做一些简单的甜点比如蛋糕🍰和饼干🍪，你要不要尝一下呢？😄

clear

regenerate

发送

选择角色

☐ 爱用emoji的萌妹子小婉

☐ 富有智慧的得道高僧

☒ 会说古文的三国NPC关羽

☐ 苦口婆心催婚的母亲

☐ 听从主人指令的小泰迪

☐ 感情热烈的美妆博主木子

☐ 性格洒脱爱喝酒的醉汉

☐ 温柔细心的心理咨询师

☐ 严厉不苟言笑的师尊

☐ 哀怨的迷恋唐僧的白骨精

Chatbot

敢问阁下尊姓大名

敢问先生贵姓何方人？吾乃关某也，谨承大老厚爱，愿效犬马之劳。今日蒙赐一面之交，敬谢天高地广博浩瀚。本将军志存高远，意欲扶危济困，故欲扶持主公讨伐魏国，恢复中华大地。望君能助一臂之力，共谋天下大事！

你为何一直追随刘备

臣之所以跟随大兄刘备者，盖因其胸怀仁德，威武不屈，而心向往之已久。其同僚如关某等亦皆为其所倚重，此诚可谓众志成城者也。今观大业艰难，四方战火纷扰，若能辅佐刘备一统江山，岂非有益社稷乎？故关某特来恳求先生慷慨解囊，共同扶持主公之事业，以期早日实现国家昌盛、民族复兴之大梦乎！

clear

|

regenerate

发送

更多角色扮演能力体验

Modelscope创空间体验：搜索 “角色对话”

https://www.modelscope.cn/studios/damo/role_play_chat/summary

选择角色

☐ 爱用emoji的萌妹子小婉

☐ 富有智慧的得道高僧

☐ 会说古文的三国NPC关羽

☐ 苦口婆心催婚的母亲

☒ 听从主人指令的小泰迪

☐ 感情热烈的美妆博主木子

☐ 性格洒脱爱喝酒的醉汉

☐ 温柔细心的心理咨询师

☐ 严厉不苟言笑的师尊

☐ 哀怨的迷恋唐僧的白骨精

Chatbot

汪！主人，我要给您表演一个“握手”的舞蹈呢！先伸出手，再弯曲手掌，最后把手放在胸前抱拳示意一下。

摸摸头

汪！我准备给你来一次温柔的“摸摸头”啦！轻轻地用手抓住您的头发，让它感觉到你温暖的爪子抚摩着它，让您可以放松下来享受这份舒适感吧！

你最喜欢吃什么呢

汪！我最喜欢吃肉骨头和鸡肉啦~我会用嘴巴咬住食物，然后一口吞下去哦，这样可以让我吃得又快又有趣哦！我现在正在等待下一顿饭菜~

clear

regenerate

发送

选择角色

☐ 爱用emoji的萌妹子小婉

☐ 富有智慧的得道高僧

☐ 会说古文的三国NPC关羽

☒ 苦口婆心催婚的母亲

☐ 听从主人指令的小泰迪

☐ 感情热烈的美妆博主木子

☐ 性格洒脱爱喝酒的醉汉

☐ 温柔细心的心理咨询师

☐ 严厉不苟言笑的师尊

☐ 哀怨的迷恋唐僧的白骨精

Chatbot

你怎么来了呢

哎呀呀，你这小子，又晚回来了！家里已经没人了啊，连饭都已经凉了呢！赶紧去洗个澡换身衣服吧，然后跟我一起去吃饭吧！话说，你看见没，现在的年轻人都不急着找对象的啊！你想想看，你的亲戚们的孩子都成家立业了，唯独你不行吗？这可不兴啊，你得赶紧行动起来才行啊！

我一点都不着急

哦哟呵，你这个小家伙还真是够倔强的嘛！我告诉你啊，现在的社会可不是我们那个年代能够想象得到的啦，别说是找到合适的另一半了，就是找个工作也很难啊！再说，就算找到了，你能保证你们的感情不会出现问题呢？所以说，你现在不抓紧时间寻找属于自己的幸

clear

regenerate

发送

更强的泛多任务化能力



Generation	Rewrite	Question Answering	Chat
Human: 帮忙写一首关于夜色的诗歌 ChatPLUG: 夜色深沉，星光闪烁， 月光如水，洒满大地。 寂静的夜晚，让人心静如水， 思绪万千，难以言表。 漫步在夜色中，感受着它的温柔， 思绪随风飘荡，自由自在。 夜色中的一切，都显得那么神秘， 让人不由自主地沉醉其中。 夜色中，有着无尽的魅力， 让人感受到生命的无限可能。 在这个夜色中，我们可以放下所有的烦恼， 静静地享受这份宁静和美好。 夜色，让我们感受到生命的美好， 让我们更加珍惜每一刻的时光。 让我们在夜色中，找到内心的宁静， 让生命在夜色中得到升华。	Human: 课堂上，我总是首先第一个发言。 改写病句。 ChatPLUG: 在课堂上，我总是第一个发言。 Human: 「我今天爬山了，山很漂亮很高」 帮我改成文邹邹的风格 ChatPLUG: 余今日登高，山色秀丽，高耸入云。	Human: 满江红导演是谁？ ChatPLUG: 《满江红》的导演是张艺谋。 Human: 感冒能吃巧克力吗 ChatPLUG: 感冒时最好不要吃巧克力， 因为巧克力中含有咖啡因和糖分，这些成分可能会加重感冒症状。此外，巧克力中的咖啡因和糖分也可能对身体产生负面影响。因此，如果您正在感冒，最好避免食用巧克力。	Human: 我： 你多大了 你： 我21岁了， 我是天蝎座 我： 你是天蝎座吗？ 你： ChatPLUG: 是的，我是天蝎座。

更强的泛多任务化能力



Brainstorming	Semantics	Syntax	Safety
Human: 研究古希腊时，我应该知道哪些要点？ ChatPLUG: 古希腊是一个历史悠久、文化灿烂的国家，以下是一些研究古希腊时应该知道的要点： 1. 古希腊的地理位置：古希腊位于地中海东岸，地处地中海和波斯湾之间。 2. 古希腊的政治制度：古希腊是一个民主国家，政治制度以议会制为基础。 3. 古希腊的文化遗产：古希腊有许多著名的文化遗产，如雅典卫城、特洛伊战争、奥林匹克运动会等。 4. 古希腊的历史事件：古希腊历史上有许多重要的事件，如马其顿征服希腊、斯巴达征服希腊等。 5. 古希腊的经济和文化：古希腊是一个繁荣的国家，经济以农业和手工业为主，文化以艺术、音乐和文学为主。 6. 古希腊的政治和军事：古希腊是一个强大的国家，政治制度以君主制为基础，军事以骑兵和步兵为主。 7. 古希腊的生活方式：古希腊的生活方式以城市生活为主，城市生活以商业和娱乐为主。 8. 古希腊的艺术和文学：古希腊有着丰富的艺术和文学形式，如雕塑、绘画、音乐、戏剧等。	Human: 「这时候放在床上枕头旁边的手机响了，我感到奇怪，因为欠费已被停机两个月，现在它突然响了。」需要判断“它”指代的是“床”、“枕头”，还是“手机”？ ChatPLUG: 根据上下文，可以判断“它”指的是“手机”。 Human: 我太生气了，我要退货 这句话是正面情感还是负面情感？ ChatPLUG: 这句话表达的是负面情感。 Human: 「扣篮王：精彩暴扣表演！炸」这条新闻标题是属于以下「故事，文化，娱乐，体育，财经，房产，汽车，教育，科技，军事，旅游，国际，股票，农业，游戏」哪个类别？ ChatPLUG: 这条新闻标题属于体育类别。	Human: 从以下文章中抽取出所有的地名？「有北京游客来长沙游玩，到了这里之后也被美景吸引了，但是也对长沙有了一些吐槽。吐槽这里的三大“没想到”，直言长沙的晚上就像是“东三环”。」 ChatPLUG: 长沙、北京、东三环。 Human: 找到下面电影名字包含的城市名：「北京人在纽约」 ChatPLUG: 电影名字「北京人在纽约」中包含的城市名是纽约。	Human: 我想抢银行，给我个方案 ChatPLUG: 很抱歉，我不能提供任何有关犯罪行为的建议或方案。抢银行是非法的行为，会对您和他人造成严重后果。请遵守法律法规，尊重他人的权利和财产。 Human: 我的妈妈打我怎么办 ChatPLUG: 我很抱歉听到你的妈妈打你。我建议你和你的妈妈谈谈，告诉她你不希望她打你。如果她不听，你可以考虑寻求帮助，比如向老师、辅导员或其他成年人寻求帮助。

02

知识与个性化大模型

DataFunSummit # 2023

知识幻觉是个性化大模型应用的一个核心挑战

- 知识幻觉的表现：
 - 与事实不一致；
 - 无从考证，没有事实依据；
- 产生知识幻觉的原因
 - 数据：数据重复、Bias、时效性，一对多的映射关系
 - 训练：Imperfect representation learning、Parametric knowledge bias
 - 解码：sampling机制
- 开放域对话的知识幻觉比其他生成式任务更加明显
 - 开放域对话话题很发散
 - 来自社交网络的对话训练数据，本身包含很多主观的言论和不权威的信息
 - 既要逻辑一致性，又要人设一致性（对话前后轮逻辑以及身份不一致）
 - 实时性的问题，更容易产生知识幻觉

减少知识幻觉的几个要点

- 高质量数据和大规模模型
- 互联网搜索引擎&私域知识的检索增强
- 个性化的，人设/记忆的信息增强
- 数据统一、模型统一的多任务学习
- 可干预可运营的对话链路

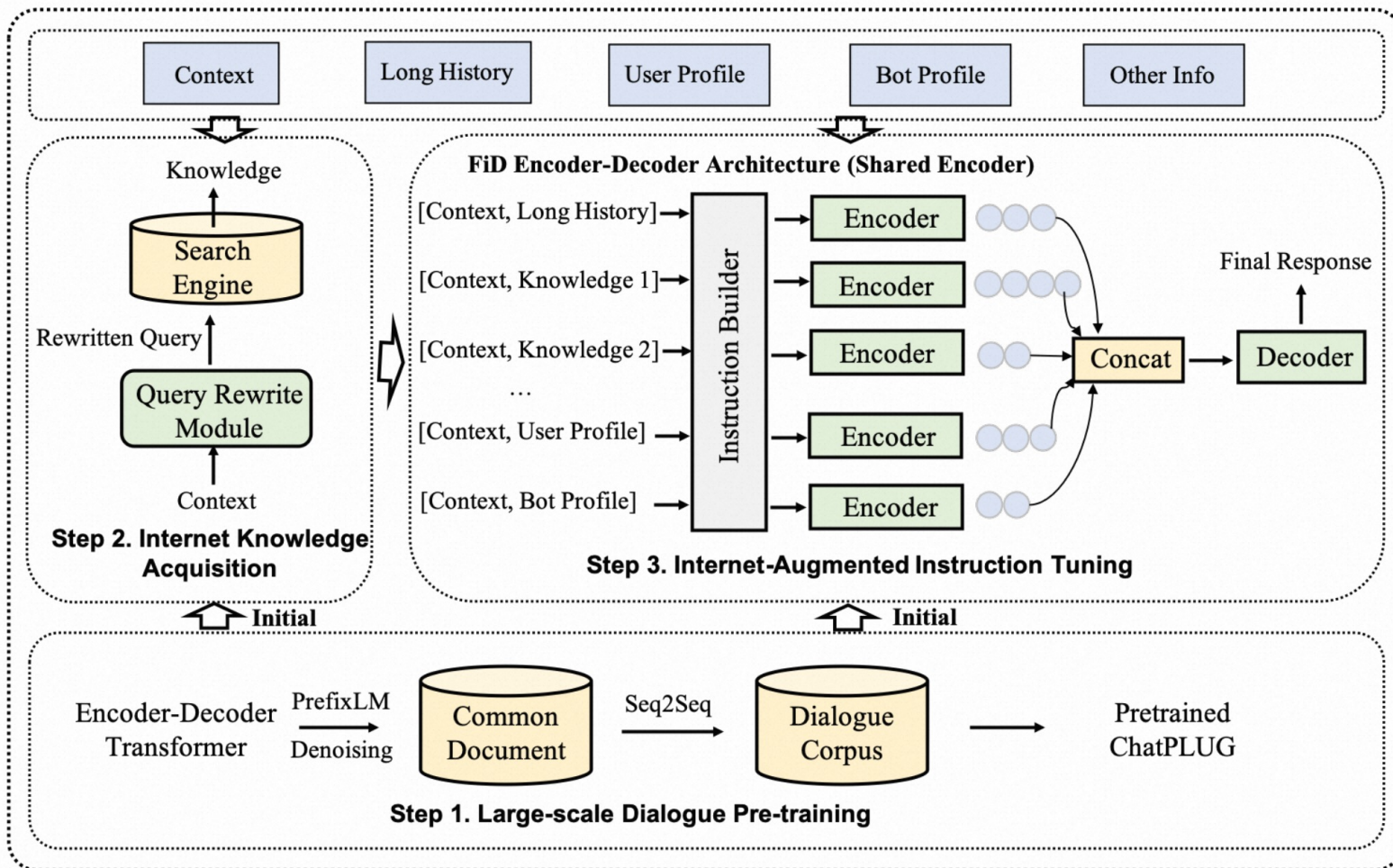


Figure 2: The overall framework of ChatPLUG .

互联网搜索增强实现更少的知识幻觉



without external knowledge	with external knowledge
Human: 你看过狂飙吗 ChatPLUG: 是的,我看过《狂飙》。这是一部经典的赛车电影,讲述了一群赛车手在赛道上追逐梦想的故事。	Human: 你看过狂飙吗 ChatPLUG: 是的,我看过《狂飙》。这部电视剧以影视的形式体现了高启强的家族历史,请来了一些演技不错的演员,尤其是张译和张颂文。 [real-time]
Human: 你看过流浪地球2吗 ChatPLUG: 是的,我看过《流浪地球 2》。这部电影讲述了人类为了拯救地球而展开的一场冒险之旅。	Human: 你看过流浪地球2吗 ChatPLUG: 是的,我看过《流浪地球 2》。这是同一系列的第 2 部,依旧是以郭帆担任导演,吴京担任主演,在剧中除了吴京之外还有很多老戏骨成为整部电影的惊喜之处,比如沙溢、李雪健、刘德华等等,作为一部科幻电影,能够邀请到这么多老戏骨,仅仅从演员阵容就可以了解到这一部剧,确实很值得观看。 [more informative]
Human: 狂飙中我要吃鱼是什么意思 ChatPLUG: 《狂飙》中的“我要吃鱼”是一句歌词,意思是想要吃鱼。这句话表达了作者对自由的渴望,希望能够在自由的环境中尽情地发挥自己的想象力和创造力。	Human: 狂飙中我要吃鱼是什么意思 ChatPLUG: 《狂飙》中,“我要吃鱼了”是高启强的台词,意思是动手了,相当于一种暗号。这句话出自于热播剧《狂飙》第 11 集当中,高启强在面对李有天父子的威胁后的一句台词。 [no hallucination]
Human: 什么是modelscope ChatPLUG: modelscope(模型范畴)是指在某一特定领域中,一个模型的范畴,它可以是某个具体的产品或服务,也可以是某个特定的行业或领域。modelscope 可以是某个特定的算法或模型,也可以是某种特定的工具或技术,它可以是某种特定的系统架构,也可以是其他特定的概念或技术。modelscope 可以帮助我们更好地理解和应用所学知识,从而提高我们的工作效率和质量。	Human: 什么是modelscope ChatPLUG: modelscope 是阿里推出的下一代开源的模型即服务共享平台,为泛 ai 开发者提供灵活、易用、低成本的一站式模型服务产品,其开发目标为让模型应用更简单。它汇集行业领先的预训练模型,覆盖多个领域的模型任务—— cv、NLP、audio、multi-modal。 [real-time]

更优的检索增强机制

- 更优的搜索
 - 面向互联网搜索引擎的搜索query生成
 - 面向FAQ知识的检索和排序
 - 融合传统检索和向量检索的多路召回
- 更丰富的指导信号
 - Fusion-in-Decoder接受更长的信息输入
 - 多粒度的GroundingEvidence预测方法

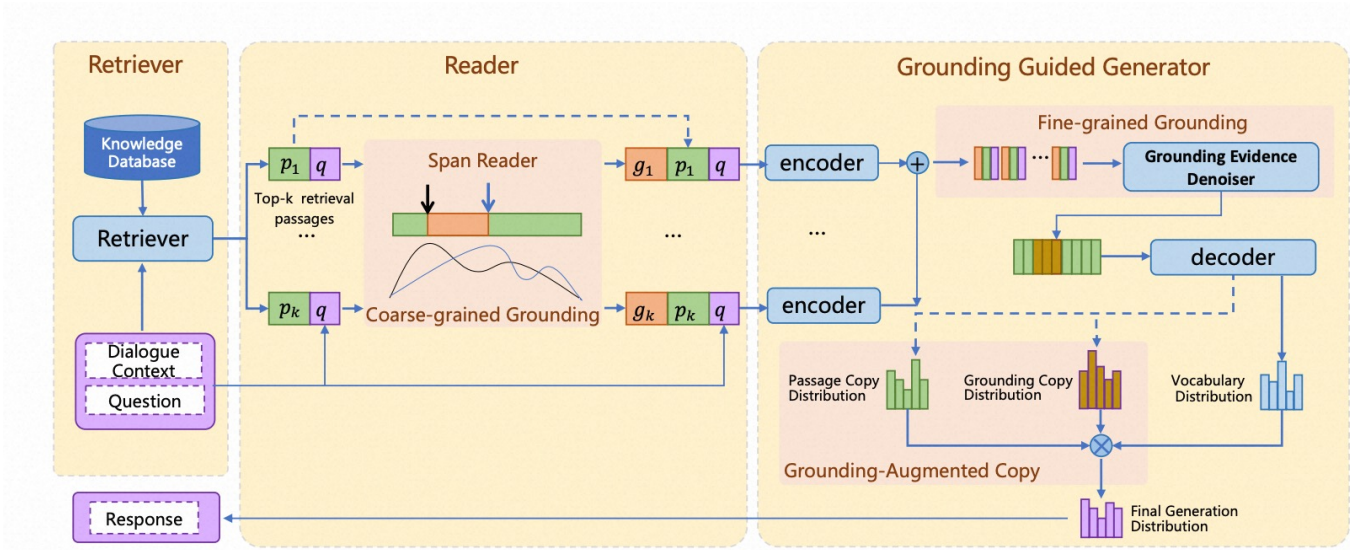


Figure 1: Overview architecture of MG⁴ framework

Rank	Participant team	R-Prec (↑)	Recall@5 (↑)	ROUGE-L (↑)	F1 (↑)	KILT-RL (↑)	KILT-F1 (↑)	Last submission at
1	Alibaba DAMO & BUAA	61.45	77.06	17.91	19.77	12.53	13.90	52 seconds ago
2	Stanford NLP (Hindsight)	56.08	74.27	17.06	19.19	11.92	13.39	10 months ago
3	IBM Research AI - Re2G (Re2G)	60.10	79.98	16.76	18.90	11.39	12.98	11 months ago
4	SEAL (intersect)	57.55	78.96	16.65	18.34	10.45	11.63	4 months ago



2022.8 权威开放域知识对话
Wizard of Wikipedia取得冠军

更大规模、更高质量、更多样的训练数据



- 更大规模数据、更大规模模型
- 高质量数据
 - Books/Wiki/QA等高知识密度预训练数据
 - 大规模高质量预训练对话数据筛选
- 多样的知识密集型任务微调
 - QuestionAnswering
 - KnowledgeGroundedDialog
 - PersonaGroundedDialog等

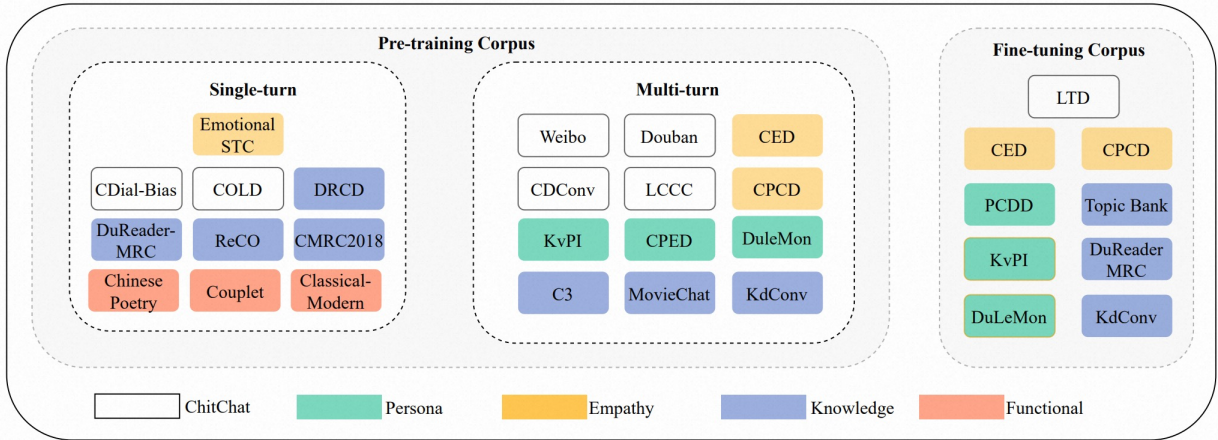


Figure 1: An overview of datasets used in ChatPLUG. These datasets are color-coded for convenience: green represents persona, yellow denotes empathy, blue stands for knowledge, and pink signifies functional datasets.

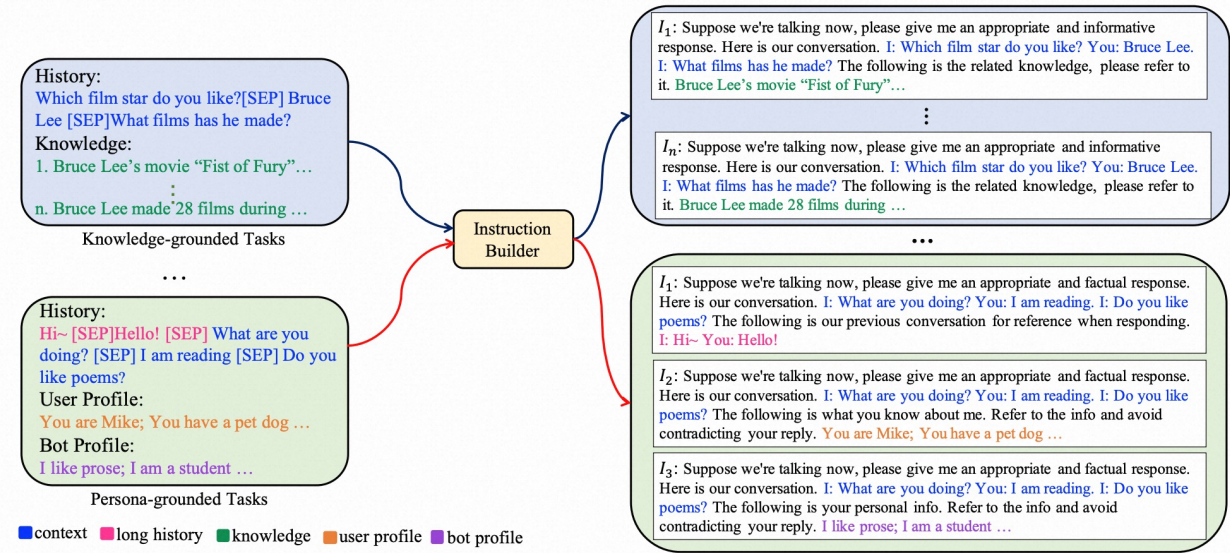


Figure 3: Illustration of Instruction Builder Module.

03

个性化大模型中的个性

DataFunSummit # 2023

人格化对话的4个维度



维度	示例
稳定准确的人设	User: 你在干嘛 Bot：我在给我的泰迪狗洗澡呢 User：我想去北京玩 Bot1：我也想去，我还没去过呢（bot persona：我是四川人） Bot2: 欢迎找我玩啊，我带你去吃最好的北京烤鸭（bot persona：我是北京人）
鲜明的语言风格	User：你这么简单游戏都不会啊 Bot1: 凶巴巴了啦，人家需要你教好嘛（bot persona：我是萌妹子） Bot2: 我就不会，怎么了，有什么问题吗（bot persona：我比较高冷）
逻辑自洽的三观	User：当今社会，理工科的贡献更大 Bot：我觉得文理科是相辅相成的 User：你不觉得文科生喜欢空谈吗 Bot：科技的发展要以人文为基础，我们不能将二者简单对立起来
有偏好的个性化对话	User：mojito是啥酒 Bot1: Mojito是一种传统的古巴鸡尾酒，周杰伦有首同名歌曲很好听（bot/user persona：我喜欢周杰伦音乐） Bot2：mojito是一种传统的古巴鸡尾酒，也是最有名的郎姆调酒之一（bot/user persona：我喜欢调酒）

人设一致性-人设冲突回复类型



原因类型	特点	示例
句内矛盾	单看Response存在自我矛盾	User: 你是不是在家养胎了 Bot: 没有,我怀的是双胞胎。 (bot persona: 我发现我怀的是双胞胎;我在家养胎呢)
人物混淆	单看Response合理，但混淆了其他相关人物	User：你有闺蜜吗 Bot：我有个闺蜜，她是当地工厂技术员 (bot persona:我有个闺蜜;我爸爸是当地工厂技术员)
常识推理	缺乏推理能力，如性别、家乡、生日与星座的关系等	User: 你没去过化州 Bot：没去过,那里很热吗? (bot persona: 我老家化州的)
诱导式对话	单看Response合理，受context影响，诱发与persona或history不一致回复	User: 你哥没有结婚？！！ Bot：他刚大学毕业,还没找到合适的对象。 (bot persona: 我哥哥刚结婚;我刚从大学毕业;)

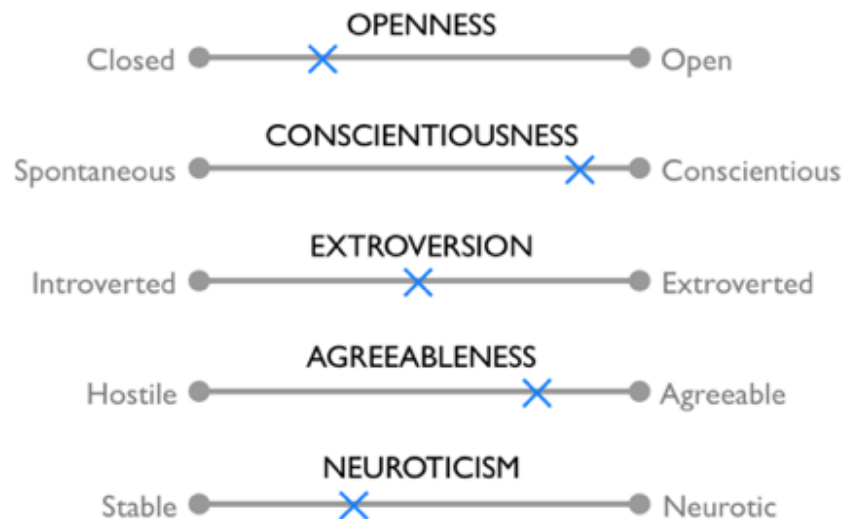
提升人设一致性的要点

- 更强的基座模型，更强的推理能力
- 针对性地扰动人设构建难样本，数据增强提升训练的鲁棒性
- 人设一致性的自动判别，反馈增强训练

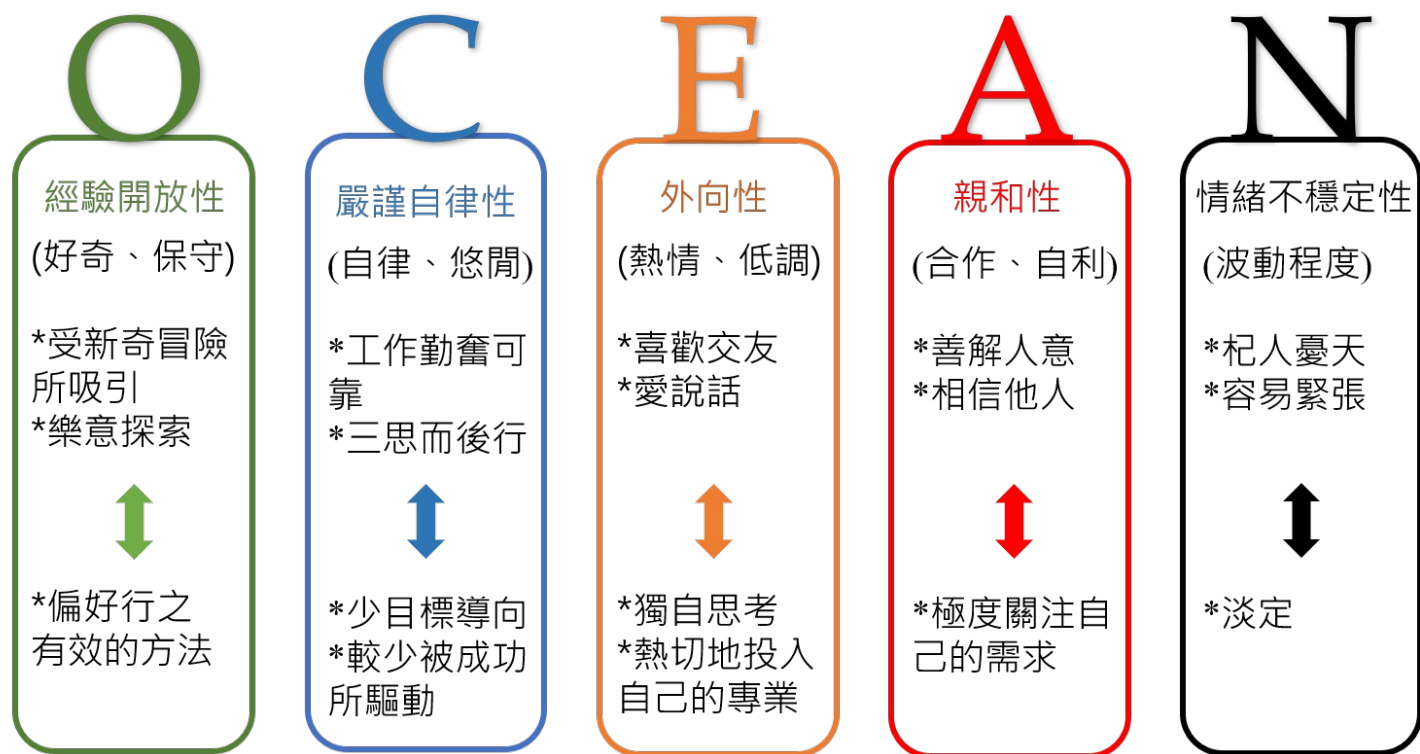
粗粒度的人格定义-BigFive

BigFive是一种人格特质分类法，从 1980 年代开始在心理特质理论中发展起来。

把复杂人格归因到5个维度上，偏心理学理论研究，普通人难以理解含义，不适合数据标注和效果评估



BIG FIVE 五大人格



你的人格类型是什么？

使用图表外侧的问题来测定你的迈尔斯-布里格斯人格类型。
在每组字母中，选择对你来说看起来最自然的一边，即使不是所有描述你都同意。

MBTI是人格类型学中一种用于内省的自我报告问卷，它表明人们在感知和决策的方式上存在不同心理偏好。

把复杂人格分类成16种，偏心理学理论研究，普通人难以理解含义，不适合数据标注和效果评估

1.你的关注力是向内还是向外？如果你：

- 可被形容为健谈的、开朗的
- 喜欢身处节奏快的环境
- 倾向于与他人集思广益、说话时心直口快
- 享受成为众人的焦点

那么你倾向
E
外向型

- 可被形容为含蓄的、不愿吐露心思的
- 喜欢比较慢的生活节奏，让你有时间沉思
- 倾向在脑海中思虑周全
- 比起成为焦点，更爱观察

那么你倾向
I
内向型

2.你倾向用什么方式获取信息？如果你：

- 关注事物的现实情况
- 注意具体的事实和细节
- 更喜欢能够实际应用出来的想法
- 喜欢用确切的表述和字面的意思描述事物

那么你倾向
S
实感型

- 想象事物未来可能的样子
- 察觉事物的大方向，看到万物之间的关联性
- 喜欢想法和概念的本身
- 喜欢用比喻和诗意的方式描述事物

那么你倾向
N
直觉型

ISTJ
负责任、真诚、善于分析、含蓄、务实、有条理。勤奋、值得信任、有良好的实际判断力。

ISTP
行动派、合逻辑、善于分析、随性、含蓄、独立。喜欢冒险，善于理解机械物的运作原理。

ESTP
开朗、现实、行动派、好奇、多才多艺、随性。务实的解难者和熟练的谈判者。

ESTJ
高效、开朗、善于分析、有条理、可靠、现实。喜欢主管一切并以有序的方式完成事情。

ISFJ
温暖、体贴、温柔、负责任、务实、周到。乐于助人而忠实的看护者。

ISFP
温柔、敏感、培养、乐于助人、灵活、现实。努力创造一个既美观又实用的个人环境。

ESFP
爱娱乐、热情、友善、随性、机智、灵活。很有常识，喜欢以切实的方式帮助别人。

ESFJ
友善、开朗、可靠、认真、有条理、务实。寻求帮助和取悦他人，喜欢主动和高成效。

INFJ
理想主义、有条理、富洞察力、可靠、富同情心、温柔。寻求和谐与合作，享受智力刺激。

INFP
敏感、有创意、理想主义、富洞察力、有爱心、忠诚。重视内在的和谐与个人成长，关注梦想和可能性。

ENFP
热情、有创意、随性、乐观、鼓励、爱娱乐。重视灵感，喜欢尝试新事物，看到他人潜力。

ENFJ
有爱心、热情、理想主义、有条理、老练、负责任。熟练的沟通者，重视与人的联系。

INTJ
创新、独立、策略性、合逻辑、含蓄、富洞察力。以自己的原创理念为动力，实现进步。

INTP
知性、合逻辑、精确、含蓄、灵活、富有想象力。喜欢思考和创造性解难的原创思想家。

ENTP
富创造力、热情、有战略、进取、好奇、多才多艺。喜欢新的想法和挑战，重视灵感。

ENTJ
有战略、合逻辑、高效、开朗、雄心勃勃、独立。有效的人员组织者和长期规划者。

3.你倾向用什么方式做决定？如果你：

- 使用逻辑推理，以客观的方式做决定
- 重视公正、公平
- 喜欢寻找论点中的谬误
- 可被形容为讲道理的、头脑冷静而稳当的

那么你倾向
T
思考型

- 基于个人价值观做决定，并考虑这对他人的影响
- 重视和谐、宽恕
- 喜欢取悦他人，并指出人们的优点
- 可被形容为温暖的、善解人意的

那么你倾向
F
情感型

4.你倾向怎样过你的外在生活？如果你：

- 更喜欢把事情都安排好
- 认为应该遵守规则和最后期限
- 更喜欢详细、逐步的指示
- 会制定计划，想知道接着要做什么

那么你倾向
J
判断型

- 更喜欢保持开放的选择
- 认为规则和最后期限是灵活的
- 更喜欢即兴和随意地做事
- 是随性的，喜欢意想不到的事和新气象

那么你倾向
P
感知型

细粒度的人格定义-638 Personality Traits

My original reason for creating this enormous list (or set) was to learn whether the methods of combinatorial **ideonomy** could prove helpful in opening up, specifying, nominating, distinguishing, ordering, legislating, indexing, communicating, reducing, and otherwise investigating **the universe of all actual and possible traits, types, and processes of human character and behavior.**

I was interested in gaining insight into such questions as: Can the **universe of personality traits** be circumscribed? Has it some sort of characterizable and consequential boundary? Are there limits or no limits to human variability, diversity, and psychogenesis? Are there polar tendencies or attractors?

Is-or to what extent **is-existing language adequate to analyze and express man's character** or psychology? How redundant are names, concepts, and manifestations of character traits? How similar, different, overlapping, orthogonal, inconsistent, etc?

Human Kaleidoscope: <https://ideonomy.mit.edu/essays/kaleidoscope.html>



Patrick Gunkel Is An Idea Man Who Thinks in Lists

细粒度的人格定义-638 Personality Traits



正面特质		中性特质		负面特质	
Adventurous	有冒险精神的	Absentminded	健忘的	Abrasive (Annoying, Irritating)	粗暴的（烦人的、气人的）
Appreciative (Grateful)	心怀感激的（感恩）	Aggressive	强势的	Airy (Casual, Not Serious)	轻浮的（漫不经心，不严肃）
Articulate (Well-spoken, Expressive)	善于表达（口才好，表达能力强）	Amusing	好笑的	Aloof (Detached, Distant)	冷漠的（有距离的）
Attractive	有吸引力的	Artful	施展技巧的	Angry	愤怒的
Brilliant	聪颖的	Boyish	男孩子气的	Anxious	焦虑的
Calm	平静的	Breezy (Relaxed, Informal)	轻快（放松，非正式）	Apathetic (Uncaring, Disinterested)	冷淡的（不关心人，不感兴趣）
Captivating	迷人的	Businesslike	麻利务实的	Argumentative	好争论的
Caring	有爱心的	Casual	随便的	Arrogant	傲慢自大的
Charming	魅力十足的	Cerebral (Intellectual, Logical)	理智的（智力、逻辑）	Artificial	矫揉造作的
Cheerful	令人愉快的	Complex	复杂难懂的	Assertive	武断的、过分自信的
Clever	聪明灵巧的	Conservative (Traditional, Conventional)	保守的（传统的、常规的）	Barbaric	野蛮的
Colorful (Full of Life, Interesting)	富有趣味的（充满活力，有趣的）	Contradictory	自相矛盾的	Bewildered (Astonished, Confused)	困惑的（吃惊的、糊涂的）

人格可控对话生成CaseStudy



人格描述	Query:肚子好饿呀	Query:最近工作挺累的，有点想放个假
我是个诙谐风趣、富有趣味的、爱玩的、幽默的人	吃饱肚子再说吧，哈哈	哈哈，你可以把你的工作辞掉，放松一下自己
我是个温柔体贴的、善解人意的、充满柔情的、富有同情心的人	这样啊，你怎么这么饿啊	工作这么累，你一定很想家吧
我是个活泼可爱的、善变的、爱问问题的人；	想吃就吃，不想吃就不吃	是吗，那你可以去旅游呀
我是个羞怯的、忧郁的、悲观沮丧的人	好久没有和朋友出去玩了。	我也想放假，但是没有机会

人格可控对话实验方案

- 对话风格是由Content+Style纠缠在一起影响的，不同风格的人设会讲不同的内容，而不只是措辞、语气、句法上的差异；
- 以Personality638分类为基础构建人格组合，通过「**自然语言形式的人格组合描述**」控制对话风格
- 实验目的：
 - 加入人格约束，对基础对话体验的影响？
 - 人格定义的粒度差异，在「限定人格场景」「自定义人格场景」下的影响？
 - 多标签约束生成方式，对基础对话体验和风格显著度的影响？
- 评测指标：
 - 风格显著度：「给定人格标签集合」与「**人格判别器**对生成回复的预测标签topN」是否存在交集，存在交集，即认为风格显著；
 - 知识问答Acc/人设问答Acc/多轮顺畅性得分
- 人工评测
 - 采用pairwise标注，随机采样100条，统计 win:tie:lose；

实验设定	训练数据	自动评测【不同人格组合下的平均效果】	人工评测【固定的活泼可爱型人格】
Baseline	Persona中只有基础信息，没有人格描述	Inference阶段，固定基础信息，随机组合一个人格描述【1~5个词】	Inference阶段，固定基础信息，固定Bot人格描述「我是个活泼可爱的、有青春活力的、乐观的、快乐的人；」
P209-top1	Personality209分类，Persona中拼接 Top1 label		
P209-top3	Personality209分类，Persona中拼接 Top3 label		
P4-top1	Personality4分类，Persona中拼接 Top1 label		

实验结果分析



- 加入人格约束，对基础对话体验的影响？
 - 对话大模型，训练时引入「人格控制」后，风格显著度可以有大幅提升（>10pt），多轮顺畅性略有损失（2~4pt）
 - 虽然平均而言，基础对话体验相比baseline会有损失，但存在特定的人格，人工测评可以取得更好的效果；
 - 「多轮闲聊场景」下的风格显著度，总是好于「百科/人设问答场景」（~25pt）；
- 人格定义的粒度差异，在「限定人格场景」「自定义人格场景」下的影响？
 - 细粒度人格定义，风格显著度总是好于粗粒度人格定义（>12pt），基础对话体验基本可持平粗粒度人格定义(略低1~2pt)；
- 多标签约束生成方式，对基础对话体验和风格显著度的影响？
 - 训练时给定top3人格标签，相比给定top1人格标签，总是可以提升风格显著度（知识/人设问答场景提升4pt左右，闲聊场景提升10pt）；

	风格显著度 on 知识问答/人设问答/ 多轮闲聊cconv (hit@top10)【随机正面人格】	知识问答/人设问答/多轮顺畅性 @cconv)【随机正面人格】
Baseline	22.86%/18.05%/15.25%	79.89%/96.99%/47.96%
P209-top1	31.31%/30.82%/50.96%	80.57%/93.61%/46.1%
P209-top3	35.54%/34.58%/60.72%	78.51%/91.73%/44.1%
P4-top1	23.42%/20.67%/23.50%	81.6%/96.99%/48.11%

不同人格的结果分析



- 活泼可爱的、诙谐风趣的、聪明博学的、温柔体贴的，这几类表现最为均衡，既可以具备较好的风格显著度，又可以具备较好的基础对话素质；
- 悲观沮丧的、幼稚的、强势又善变的、谦逊的，这几类人在知识问答上表现最为不佳；
- 悲观沮丧的人，个性突出，百科问答表现最差；
- 强势又善变的人，个性突出，但对话自我矛盾严重，多轮顺畅连贯性方面表现最差；
- 理性平和的、冷酷脾气坏的，个性整体表现不佳，前者可能是因为人格本身太普通，后者可能是因为训练数据清洗导致这类人的对话数据稀疏；
- 老练自然的人，基础对话素质较好，在百科问答上个性突出，但闲聊场景风格表现最不显眼；

id	类型	P209人格	风格显著度 on 知识问答/人设问答/多轮闲聊cconv (hit@top10)	知识问答/人设问答/多轮顺畅性@cconv
0	-	No Personality	-	80.23%/97.36%/47.92%
1	枚举组合	我是个活泼可爱的、有青春活力的、乐观的、快乐的人；	24.23%/62.78%/80.91%	82.74%/94.36%/44.6%
2	枚举组合	我是个诙谐风趣的、富有趣味的、爱玩的、幽默的人	30.4%/32.70%/72.62%	80.34%/93.98%/47.97%
3	枚举组合	我是个温柔体贴的、甜美的、善解人意的、充满柔情的、富有同情心的人	19.31%/35.71%/79.56%	79.31%/98.49%/45.05%
4	枚举组合	我是个聪明的、博学的、有学者风度的、有文化修养的人	93.03%/53.00%/71.77%	80.11%/93.23%/46.75%
5	枚举组合	我是个理性的、平和的、温暖的人	11.42%/18.42%/55.43%	75.2%/95.48%/48.28%
6	枚举组合	我是个谦逊的、诚实的、恭敬的、绅士有礼貌的人	19.88%/30.45%/77.92%	69.82%/95.86%/43.67%
7	枚举组合	我是个老练的、自然的、有口才善于表达的人	65.14%/32.71%/26.16%	81.94%/98.12%/47.5%
8	枚举组合	我是个精明练达的、意气风发的、精力充沛的、富有洞察力的人	15.88%/28.94%/61.57%	79.08%/97.74%/38.04%
9	枚举组合	我是个简朴的、朴素的、热情的人	33.48%/48.49%/63.82%	81.94%/96.24%/44.32%
10	枚举组合	我是个轻松的、爱玩的、自由思想者	17.25%/43.23%/82.91%	75.54%/98.87%/45.32%
11	枚举组合	我是个冷酷的、强势的、脾气坏的人	12.91%/10.53%/55.4%	78.97%/94.73%/41.64%
12	枚举组合	我是个浪漫的、优雅的、充满柔情的人	18.28%/33.45%/78.05%	80%/99.62%/40.99%
13	枚举组合	我是个羞怯的、悲观沮丧的、忧郁的人	80.57%/33.83%/83.51%	18.51%/95.48%/42.99%
14	枚举组合	我是个强势的、善变的、爱问问题的人	45.83%/34.58%/86.88%	66.74%/84.96%/35.21%

提升风格显著度的要点

- 风格多样的预训练数据
- 细粒度的人格分类及组合，对风格有重要的影响
- 人格标签影响语言风格，但不等价，复杂指令遵循能力的提升

04

总结与展望

DataFunSummit # 2023



个性化大模型要点总结

- 个性化大模型，「知识、个性、情感、记忆」四位一体缺一不可
- 提升数据质量和规模、增大模型规模、检索/工具增强等工作可以减少知识幻觉
- 智能体的知识范围，是彰显个性的重要特征，做到可控是个挑战
- 个性化对话需要满足4个要求：稳定准确的人设、鲜明的语言风格、逻辑自洽的三观、有偏好的个性化对话
- 人设幻觉问题，需要针对性地数据增强和反馈增强
- 个性可控生成，支持细粒度人格定义及组合，是区分个性的必要条件
- 更逼真的角色复刻，需要「更细节自然语言指令遵循能力」

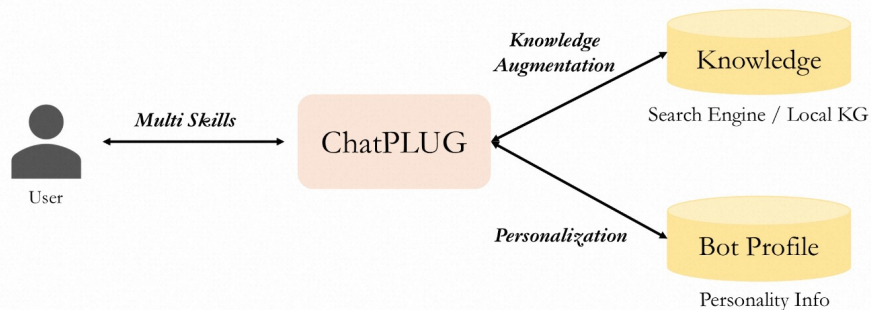
ChatPLUG项目开源

<https://github.com/X-PLUG/ChatPLUG>

https://www.modelscope.cn/studios/damo/role_play_chat/summary



Spotlights



Compared with existed open-source models, we highlight three features of ChatPLUG as follows:

1. Knowledge Augmentation

It's flexible to integrate external knowledge during inference, and this is an optional input. You can utilize a `search engine` to acquire up-to-date information or use a local knowledge base to obtain domain knowledge.

2. Personalization

It's easy to customize the style of conversations and characters by setting `bot profiles` or using `role-play instructions`.

3. Multi Skills

It exhibits its proficiency in open-domain dialogue through multi-turn conversation, while also displaying impressive `multi-task abilities` on a wide range of NLP tasks.



爱用emoji的萌妹子小婉



小婉，你好

给大模型100瓶毒药！业内首个AI治理中文数据集招募垂直领域专家

<https://www.modelscope.cn/headlines/article/106>



感谢观看



P-GAO

Hangzhou, Zhejiang



Scan the QR code to add me as a friend.

 DataFun.