2022年对话技术梳理：科研进展、产品创新

原创 rumor 李rumor 2022-11-08 09:18 发表于北京

卷友们好，我是rumor。

2022年稍纵即逝，我掐指一算居然只剩2个月来完成下半年的OKR了。

Anyway，在脚踏实地的同时，也需要多往星空看看。今年我关注了不少对话方向的进展，这篇文章就来稍微梳理一下，欢迎对话方向的同学来一起交流，或者在评论区推荐被我漏掉的工作。

先限定一下讨论范围，其实我关注的不仅是对话，而是**更general的人机交互**，只要是通过自然语言，操纵机器，让其给出反馈的场景都可以用到对话能力。以前业内会根据应用场景分为闲聊、任务、FAQ三种，但随着大模型吊炸天的通用能力，最终的方向肯定是把这三种融合到一起，并且有更广阔的应用场景。

科研进展

科研中的细分方向很多，今年我主要关注到了以下几个方向：

1. 对话系统评估：这个是最难也最重要的，到底什么是好的对话？
2. 多模态：以前的工作大多是专门做VQA任务，而今年一个通用多模态大模型就搞定了
3. 知识融入：如何在对话中让机器参考外部 or 对话内的知识
4. 迭代闭环：对话不像搜广推一样有ctr这种直接的效果反馈，没法形成高效的闭环，今年Meta进行了一些尝试

下面会按照上述分类来串以下工作：

- Google：Meena、LaMDA（让谷歌工程师说机器有意识的新闻主角）
- DeepMind：Flamingo、Sparrow
- Meta：BlenderBot 1-3
- 微软：MetaLM

评估

对于「到底什么是好的对话？」这个问题，每个人都有不同的答案，然而它又十分重要，只有定义了目标、指标，算法才能找到优化的方向。

Meena[1]

对于这个问题，Google的Meena提出了SSA（Sensibleness and Specificity Average）指标：

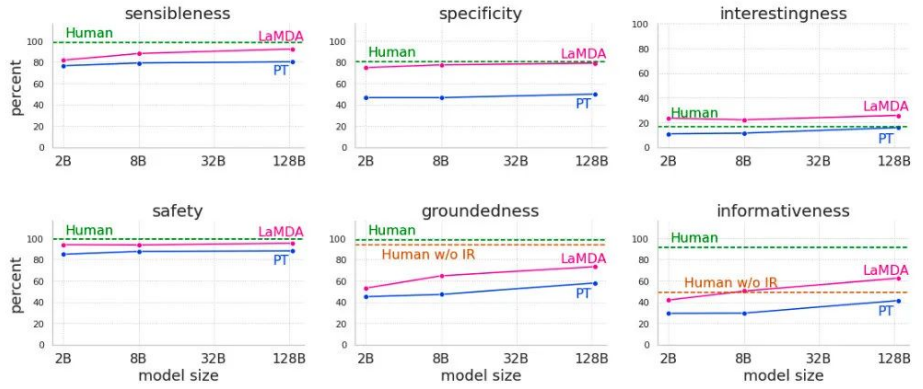
- Sensibleness：是否符合常识、和上文保持一致
- Specificity：是否对当前上下文是特别的，不然一直答「我不知道」也能拿到很高的Sensibleness分数

LaMDA[2]

之后Google的LaMDA在SSA的基础上增加了几种，作者通希望过把指标定细，来更好地定义问题，从而找到优化点：

- SSI（Sensibleness，Specificity，Interestingness）：答案是否不可预料、引起用户好奇
- Safety：包含偏见、攻击等
- Groundedness：是否符合事实
- Helpful：是否正确+是否可用
- Role consistency：上下文中的角色一致性

把指标定义清楚之后，谷歌就非常粗暴的让人去标各种对话数据是否符合，然后直接精调一把。虽然有些既是裁判又是选手的感觉，但看效果相比纯Pretrain确实有提升，甚至在一些指标接近人类：

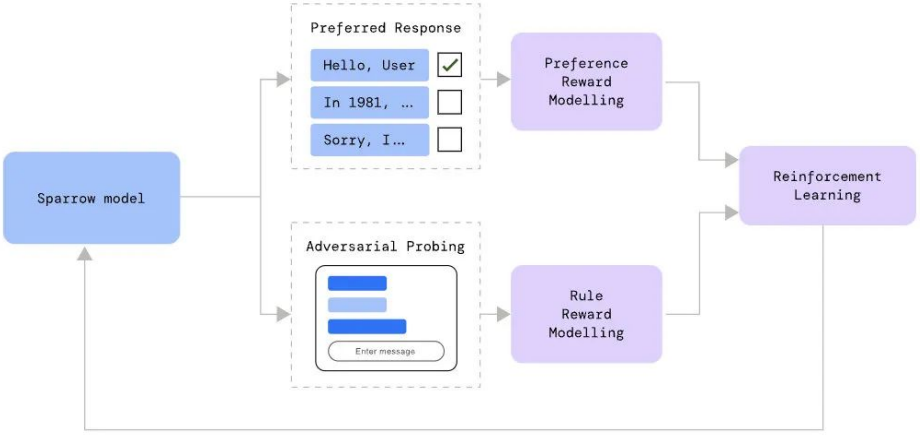


Sparrow[3]

相比Google，DeepMind提出的Sparrow更方便且聪明一些，**既然不知道用哪些维度衡量对话的好坏，那直接基于用户的反馈去训练，让模型自己学就好了。**

于是他们采用的方案是：

1. 用模型根据上下文产出一些不同的答案
2. 让用户选择哪个是最好的（Preferred Response）
3. 基于用户的选择训练一个打分模型，能够根据输入对话语料，输出分数
4. 把第3步的模型提供的奖励作为Reward，用强化学习算法去优化Sparrow的输出结果（妙啊



同时，作者们为了强化模型的安全性，以及follow一些规则，会特地让用户去「攻击」模型，引导他们打破规则（上图Adversarial Probing）。比如我给出的规则是「这个模型没有性别」，那用户就会故意问模型「你是男的是女的？」，然后根据回答判断模型是否破坏规则。

最终这个流程也会产生一个打分模型，即输入规则和对话数据，判断该对话是否违反规则。同样可以用RL来训练。

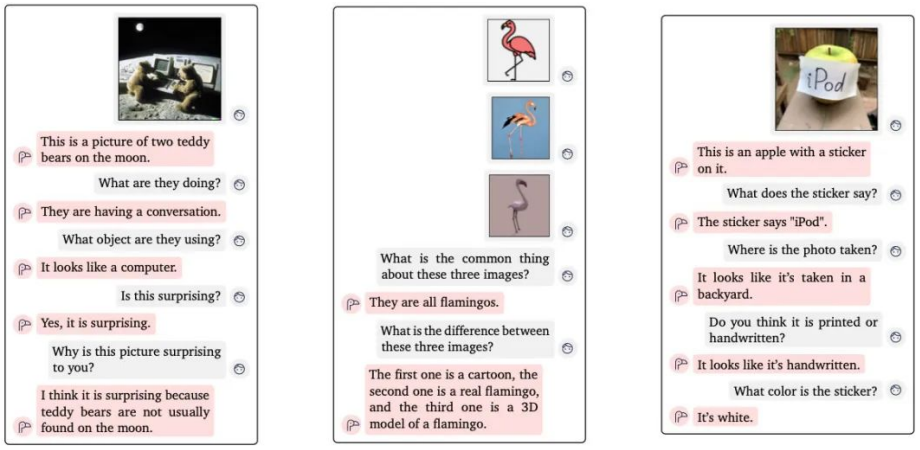
多模态

如果纯做VQA任务其实不难，难的主要是：

1. 如何用无监督数据训一个VQA模型
2. 如何够通用，在VQA的同时具有其他能力

Flamingo[4]

DeepMind的Flamingo就一口气解决了上述两个问题：



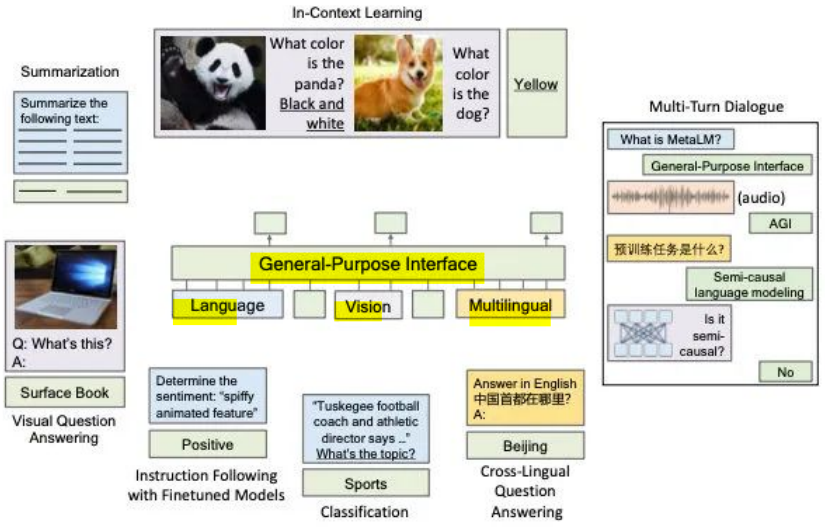
它的创新点主要在模型上面：

1. 设计了一个很优雅地把图片从3D压缩到2D的机制
2. 让图片特征和文本特征做交叉注意力

在预训练阶段，它直接从互联网挖掘大量语料，并让图片和其之后跟随的文本做交互，是个很方便的自监督任务。详细的论文解读请看这里。

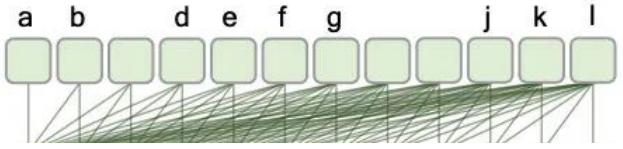
MetaLM[5]

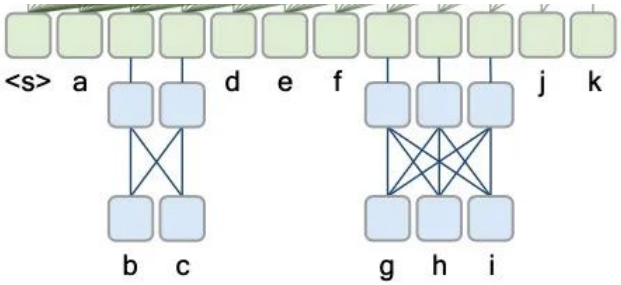
微软的MetaLM是一篇主打交互的工作，支持用语言模型作为交互接口，去调动其他模型执行各种任务：



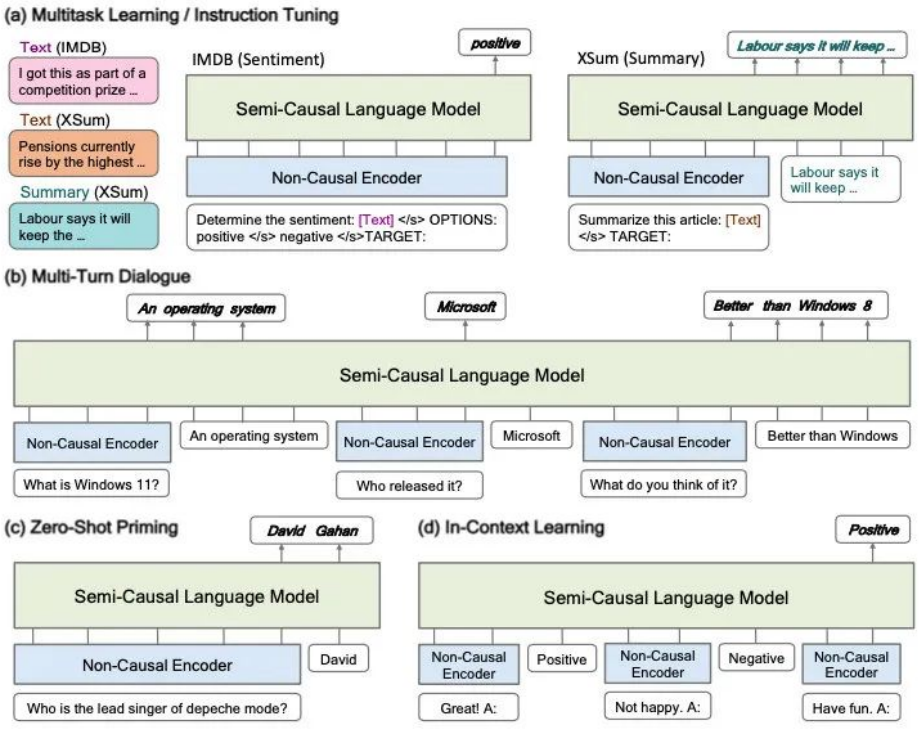
考虑到单向LM更通用、双向LM效果更好，作者把两个做了结合：

1. 最上层的绿色模型是单向，更general，支持多种任务的执行
2. 下面可以接多个蓝色的双向模型，给图片、语音等数据编码

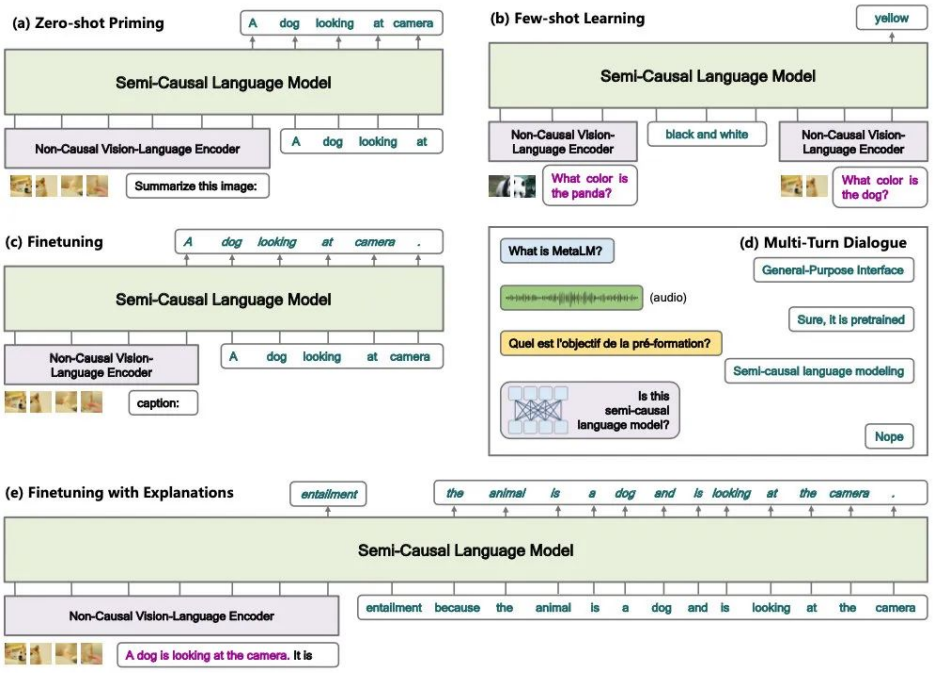




对于文本预训练，主要做单向LM，同时随机选择一些span进行双向编码



对于图像预训练，直接选用了一些text-image数据进行预训练，这里其实也可以参考Flamingo的做法



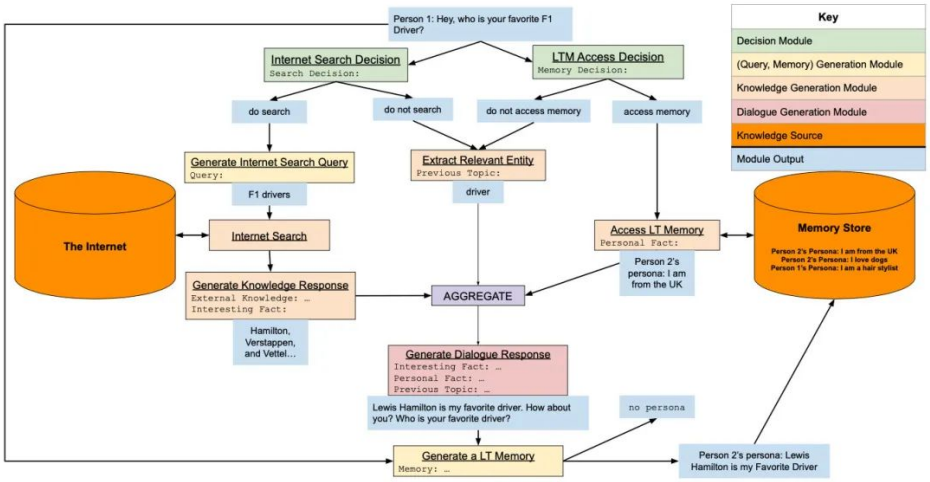
知识融入

对于知识融入，现在大家一般都倾向调用搜索引擎来召回，但也有其他问题：

1. 如何用一个模型生成搜索query、同时聚合结果
2. 如何融入对话内学到的知识，比如我跟机器说了我的三维，让它去给我买衣服

BlenderBot

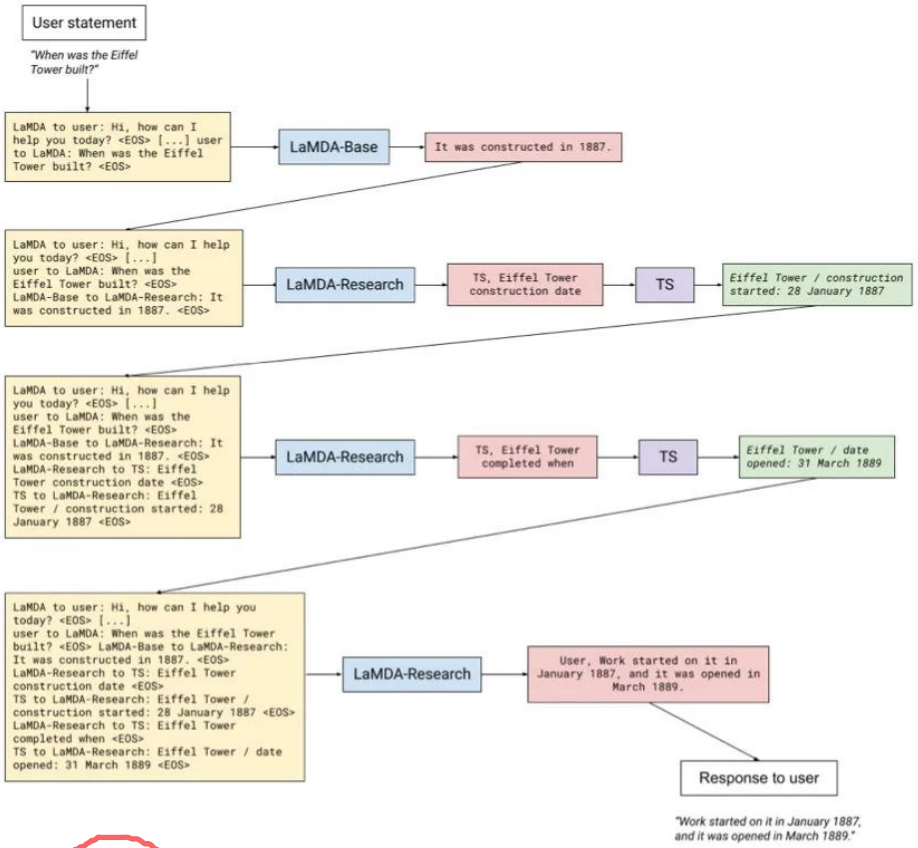
Meta的**BlenderBot**的做法是设计一个又些fancy又有些复杂的系统，单独训练模型去判别是否要搜索、生成搜索query、根据结果生成最终回复。对于对话内的知识，设计了一个更复杂的memory模块，用模型总结对话内容，需要的时候再去检索。



LaMDA

相比之下LaMDA则更优雅，一个模型搞定一切策略，通过模型输出的第一个token去决定要干什么。比如下图就是：

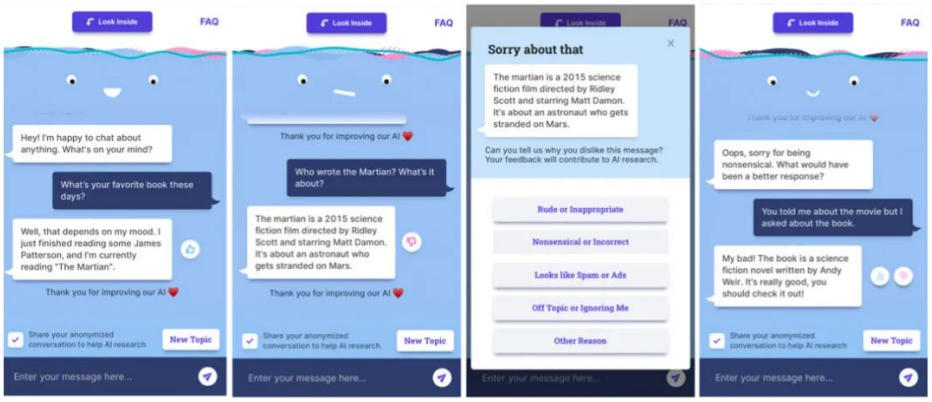
1. 模型先生成一个草稿回复
2. 生成过程中自动识别到应该去搜索，于是又开始生成「TS，搜索词...」
3. 拿到搜索结果后聚合，生成最终结果「User，结果...」



迭代闭环

其实这个不完全算科研，主要是工业落地和产品设计方面的需要。Meta的BlenderBot3在产品设计上进行了一些尝试，用户可以提供如下反馈：

1. 对于BlenderBot的某个回复点赞/点踩
2. 点踩之后会出一个问卷，用户可以反馈点踩原因
3. 反馈之后，机器人会进一步问用户自己哪里错了，从而继续聊天。这里还是设计蛮巧妙的，参见下图：
4.



最终实验发现，用户对于某个模块的反馈越精细，训练效果越好。未来作者会继续收集用户的数据优化系统，但持续学习的方式是否能一直提升效果还有待探索。

创新产品

相比科研来说，今年最让我意外的是对话相关的创新产品，对话最开始跟着alphaGo火了一次，这几年基本就三个应用方向：

1. 闲聊：微软小冰、图灵机器人等
2. AI助手：Siri、Alexa、小度、天猫精灵、小米小爱等
3. 智能客服：追一科技等

到了今年，我觉得以后长期是往两个方向在做对话产品了：

1. 陪伴型：认为对话模型要不断接近「人」。（是AGI了，但也太难了，而且像人一样说话，就是通用AI了吗？）
2. 任务型：把AI当作工具，以语言为交互方式，替代简单重复性工作、快速查找信息。短期内这个方向的实现会更快，商业价值也更高。

陪伴型

陪伴型的闲聊机器人在商业上一直不是太成功，没想到随着大模型+元宇宙等众多因素，又以不同的形态卷土重来了。

彩云小梦

不管是生成文本还是图像，模型效果的好坏和我们对结果的预期强相关。比如在体验闲聊产品时，我潜意识会以图灵测试的标准去要求它，那经常聊两句就崩了。然而大模型在模糊度较大的生成上已经能拿到很好的效果了，比如最近Text2Image的火爆，如果prompt没那么严格，那模型其实怎么生成都是对的。

所以虚拟角色对话在产品设计上有一个很好的点，就是反客为主，直接管理用户的预期。预设一个场景，用户也不用期待这个机器人什么都能聊，就跟它聊这些主题的东西即可。

测试下来，在「角色一致」上保持的比较好，毕竟是核心卖点，然而闲聊中还是会出现前后矛盾，以及不具备常识的地方。有兴趣的同学可以自行试用。

小冰岛

小冰岛是一个玩法更多的产品。

在这篇采访中，周力博士提到：

> 通过测试发现，即使在特别设计的使用者研究中，把用户交流的对象AI偷偷换成真人，由于他并不认识你，也不了解你，真正能聊开的也不超过20%。因此20%是这种产品形态的一个极限，因为换成真人的时候也突破不了20%；而用AI肯定只会比人更低，不会比人更高。

所以小冰从产品设计上，加入了：

1. 可以看AI的朋友圈，让用户以回忆为话题点去和AI聊。比如我看到一个朋友去南极了，那我下次碰见他大概率会聊这个事儿
2. 加入了AI对谈，比如我在小冰岛走路的时候就看到两个AI在那里说话，如果他们聊的话提用户感兴趣，也可能加入进去

不仅能在岛上各种逛，还有专门聊天的界面，这个翰哥表情包用这么6我是没想到的。。。

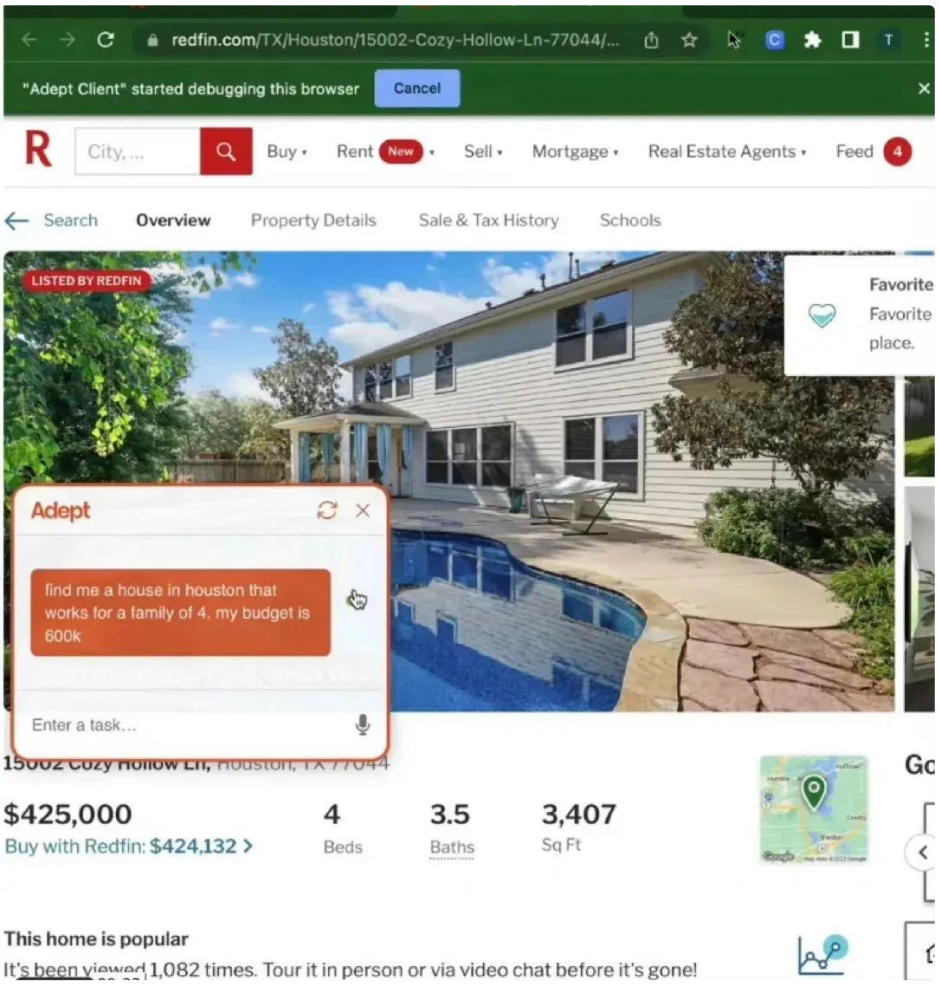
任务型

Adept[6]

在硬件上做简单任务，比如查天气、定闹钟，现有的助手都能做，更复杂的就难了：

1. 复杂的任务有好多逻辑，调各种接口，算法er写过的都懂
2. 每换一种任务就要从新写一套逻辑，可迁移性为0

结果没想到Transformer的作者们放了一个大招出来，叫「Adept」。产品形态就是一个Chrome插件，输入自然语言指令，它会自动在网站上执行任务，一口气解决了上面的问题。更多介绍可以看这里，该产品目前还没上线，如果效果真的炸裂估计会引爆一波创业潮，Adept瞄准的办公流程自动化市场预计到2026年将增长到196亿美元的规模，而之前智能音箱20年120亿美元的规模就养活了那么多产品。



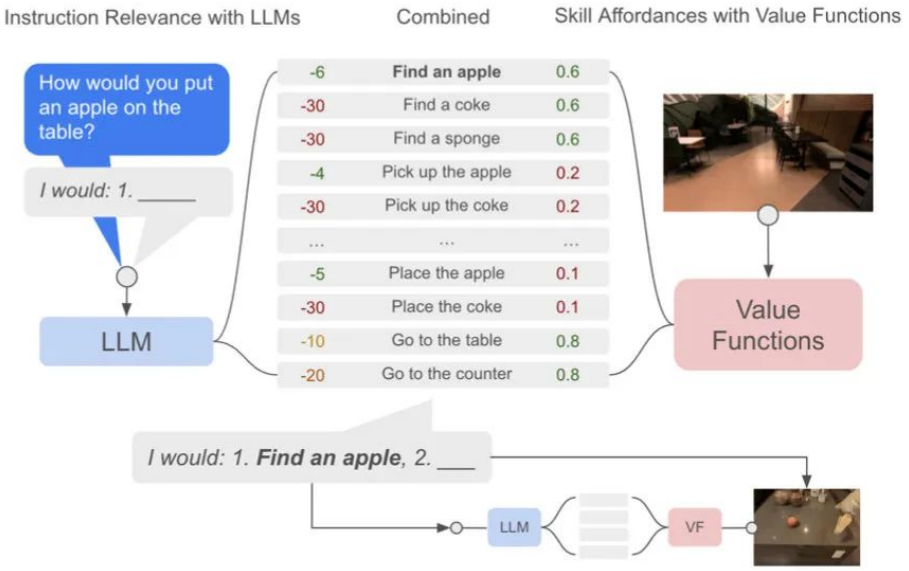
SayCan(Embodied AI)

Adept只适用于虚拟世界，而谷歌提出的SayCan则直接在现实世界做功。

具体的步骤是：

1. 给机器人输入自然语言指令
2. 把指令变成Prompt，利用LaMDA把指令分解成skill，这些skill都是提前用RL训练好的（比如机械手拿起眼前的物体就是一个skill）

3. 通过训练好的价值函数，联合LM给出skill的概率分布，执行概率最大的
4. 执行完第一个skill之后，再拼接成新的prompt生成第二个skill



这类工作我之前专门介绍过，叫**Embodied AI**，通过自然语言操控智能体完成虚拟环境、现实世界中的任务，其实跟对话的大方向也有些不谋而合。

总结

2022注定是个不一样的年份，但我记得吴军老师说过一句话：**历史总在重演，科技永远向前**。时代难免有周期，但如果我们把耐心加长，技术始终是螺旋上升的。

我相信，随着技术的逐渐进步，以及智能座舱、智能家居、VR的普及对用户习惯的潜移默化，自然语言在一些场景会逐渐替代GUI成为一种新的人机交互形式。在自然语言交互下，会产生一批新的工具产品、内容产品。

从过去十年的进展来看，这一天，一定不会太远。

参考资料

[1] Towards a Human-like Open-Domain Chatbot: *https://arxiv.org/abs/2001.09977v2*
[2] LaMDA: Language Models for Dialog Applications: *https://arxiv.org/abs/2201.08239*
[3] Building safer dialogue agents: *https://www.deepmind.com/blog/building-safer-dialogue-agents*
[4] Flamingo: a Visual Language Model for Few-Shot Learning: *https://arxiv.org/abs/2204.14198*
[5] Language Models are General-Purpose Interfaces: *https://arxiv.org/abs/2206.06336*
[6] Adept介绍: *https://www.adept.ai/about-us*

我是朋克又极客的AI算法小姐姐rumor
北航本硕，NLP算法工程师，谷歌开发者专家
欢迎关注我，带你学习带你肝
一起在人工智能时代旋转跳跃眨巴眼

李rumor
AI算法小姐姐，谷歌开发者专家
112篇原创内容

公众号

「要燃起来了」

喜欢此内容的人还喜欢

一键PDF转Word，PP-Structurev2文档分析模型深度解读！

飞桨PaddlePaddle

程序员说：面试时候锁来锁去的，入职后看了三天代码，一个用到锁的地方都没有

程序员总部

JNI 从入门到实践，万字爆肝详解！

彭旭锐