ChatGPT系列—百度文心一言解读20230315

【介绍】

l2019年，正式发布ERNIE产业级知识增强模型（来源于各个产业互联网的预训练大模型），同时融入了百度特色的知识图谱，把数据和知识相融合，提高深度学习效率以及适配性。之后也在语义理解、文本生成、跨模态语义方面取得新的技术突破。

l文心大模型强调产业级知识增强的特性，旨在降低B端应用场景的AI门槛，便于二次开发。和金融、电力、航天等行业的生态伙伴发布一系列行业大模型，帮助企业迅速迭代出和业务模式相匹配的模型，截至22年底发布了11个行业大模型。

n2022年，百度世界大会和中国探月工程联合发布百度航天文心大模型，是首个航天领域大模型，把航天领域的知识图谱和客户积累的数据进行智能采集、分析和理解，助力智能感知、深空探测的技术突破。

l大模型之上提供工具和平台层，以SDK、API接口调用的方式为AI开发者提供大模型的套件，面向零基础开发者的EasyDL可以做简单的AI开发，把AI中台封装在成型的BML大模型向外做相应的输出。

l产品级已经发布了文心一格和文心百中两款文心系列的产品，文心一言是文心系列的第三款产品。

l文心一言是基于文心大模型推出的生成式对话产品，2月7号在内部正式立项，3月16号正式发布，上升到百度集团优先级最高的项目，CTO王海峰博士亲自挂帅，深度学习国家工程研究中心副主任吴甜副总裁、百度技术委员会主席吴华等技术大拿参与指挥项目。

l文心一言对标ChatGPT3.5，除了文本生成之外，后面还会逐步迭代生成文章、诗歌、歌词。国内其他的研发团队如复旦大学实验室、阿里、腾讯还没有明确的时间表。

l文心一言的能力将全面嵌入到百度现有的业务中。比如智能音箱小度、百度智能云、无人驾驶。

n百度智能云以AI大底座的方式承接文心大模型，通过云智一体的方式引导B端G端客户从芯片、框架、模型、应用配置自己的云计算需求。

n文心一言也会和搜索引擎深度融合，引领搜索引擎的代际变革。

n智能驾驶业务上，比如与阿波罗自动驾舱、车路协同做深度融合，使无人驾驶更加安全可靠。

l从2月中旬到3月中旬，有广电、金融、汽车、媒体各个领域的相关企业逐渐进入文心一言生态合作伙伴产业链。目前，接近500家企业宣布加入文心一言生态圈。一些行业头部KA客户参与到部分的百度文心一言内部测试工作中，帮助做一些特定场景化下的训练和推理。

【Q&A】

Q：对发布会合理的期待？

A：发布会只是一个小小的时间点，是中国市场类ChatGPT产品空白的填充，4、5月份还会有下一个版本或者新功能的发布。还达不到GPT-4生成内容的质量和水准，对标的是3.5，后续以月或双月的频率发布新功能或新版本。

Q：算力卡脖子问题，目前拥有A100或者A800的量级，配置在文心上的量级？

A：文心一言在百度2月到3月的优先级最高，比如百度阳泉超算中心主要为文心一言做训练推理。除了A100，还用了一些国产化的产品，比如寒武纪的思元590等等。

Q：中美脱钩，国产化有什么规划？

A：从政治和发展角度考量，尽可能引入更多国内厂商，性能允许的情况下尽可能多一些尝试。内部目标2-3年GPU芯片实现50%以上国产替代。

Q：国产芯片和英伟达芯片性能差距在什么量级？性价比的差距？

A：寒武纪思元590和A100对比，590要增加20-30%的工作量和时间。高优先级还是用A100，可以把控的测试用国产芯片；同时也在帮国内AI芯片企业做相应内测。这种差距目前来看可以接受。

Q：ERNIE参数的量级和训练内容的量级这几年按照怎么样的节奏变化？

A：

l文心大模型里有一个鹏城的模型，训练参数达到2600亿，相对GPT提升不少。ERNIE从1.0到2.0再到3.0，经常谈到与知识图谱平行预训练算法，以及兼顾语义理解生成的预训练框架。

l文心一言的参数以2600亿为基数，会做100亿、200亿量级的模型优化。

Q：和GPT3.5对比，文心一言中文语料的量级？

A：中文语料占比75-85%，中文语料绝对量根据内部观察是GPT3.5的10倍以上的量级。

Q：文心大模型在多模态上的进展是怎么样的？

A：跨模态这一块目前来说和GPT类似产品相比大概有一代到两代的差距，比如今天GPT-4做的事情百度至少要半年以后才能做，视频、图片生成要半年以后大概能有一个相对比较稳定、高质量的输出。文心一言在高质量的文本形态生成上比较有把握。

Q：文心大模型目前的成本情况如何？

A：现在属于内测阶段，投入主要是算力资源、人力资源、数据生成和采集资源，算力占50%以上，人力成本20%出头，数据成本15-20%，剩下的是算法成本。

Q：集团对文心大模型资金投入的量级？

A：参考财报提到，研发投入是营收的20%，具体不太好细拆，光看研发层面，其中50-60%是和文心大模型相关的投入。

Q：文心大模型变现定价如何，未来打算开源还是闭源？

A：

I会逐步开源，现阶段不会马上开源。现在最大的方式是通过百度智能云对外做一些行业生态合作伙伴的共创。

I先选择一批客户做初步的协议定价，然后再根据情况看市场迭代效果，现在还没有太多清晰的商业化的方向。

Q：到边缘端，会不会嵌入到类似小米（IoT），是否会带动一些物联网需求的增加？

A：目前在探索的应用场景中，以小度音响为代表的一系列智能硬件，这些基于DOS的智能硬件都会和文心一言进行深度融合。后续会创造一些AIoT形态的新产品，是积极尝试的方向。

Q：模型训练和使用时候用到的算力基础设施是不是一样的，可以复用的？

A：是的。

Q：使用一次问答的成本？

A：等发布会以后才能解答。

Q：文心系列产品的重心会不会发生改变？

A：2023年文心大模型所有的产品是以文心一言为主做相应融合，24年以后还有其他一些产品，会结合市场变化再做判断。

Q：ERNIE的技术路径和ChatGPT的路线是不是不一样？

A：百度内部不太关注一条一条技术路线逐条和ChatGPT做对比，技术演化路径是比较符合自身的路径。

Q：GPT-4的参数量级？数据使用量？

A：估算在三四千亿的量级。数据量比3.5多2-3倍的量级。

Q：ERNIE参数量级的增长？

A：参数会逐渐上升的。跨模态是比较重要的方向。文心大模型是源于行业的，每一步迭代更新都和行业紧密相关，这是最核心的底层逻辑。

Q：目前和GPT-4是半年左右的差距，有可能将这一差距缩短吗？

A：会的。目标是GPT-4、GPT-5逐渐拉平，基于百度现有研发资源，基于初级版本内测过程中的问题修复，在初级版本0-1的过程中耗时是比较多的，过了0-1，从1-100迭代过程中，相对进度就会以非线性去迭代了。现在看是半年，之后根据资源投入不同去评估，应该会大幅度缩短。

Q：什么时间多模态融入到C端的文心产品

A：大概规划应该在23年底前后，会有让大家感受到提升用户体验的产品形态的展现。发布会只是起点，后续会有很多新的迭代。

Q：2、3月的迭代是哪个参数级别的模型？

A：2600亿参数的文心大模型迭代是一个长期的工作，2023年2、3月以文心一言为主。