盘古之殇：华为诺亚盘古大模型研发历程的心酸与黑暗

各位好，

我是一名盘古大模型团队，华为诺亚方舟实验室的员工。

首先为自证身份，列举一些细节：

1. 现诺亚主任，前算法应用部部长，后改名为小模型实验室的主任王云鹤。前诺亚主任：姚骏（大家称姚老师）。几个实验室主任：唐睿明（明哥，明队，已离职），尚利峰，张维（维哥），郝建业（郝老师），刘武龙（称呼为武龙所）等。其他骨干成员和专家陆续有很多人离职。
2. 我们隶属于“四野”这个组织。四野下属有许多纵队，基础语言大模型是四纵。王云鹤的小模型是十六纵队。我们参加过苏州的集结，有各种月份的时间节点。在苏州攻关会颁发任务令，需要在节点前达成目标。苏州集结会把各地的人员都集中在苏州研究所，平常住宾馆，比如在甪直的酒店，与家人孩子天各一方。
3. 在苏州集结的时候周六默认上班，非常辛苦，不过周六有下午茶，有一次还有小龙虾。在苏州研究所的工位搬迁过一次，从一栋楼换到了另一栋。苏州研究所楼栋都是欧式装修，门口有大坡，里面景色很不错。去苏州集结一般至少要去一周，甚至更久，多的人甚至一两个月都回不了家。
4. 诺亚曾经传说是研究型的，但是来了之后因为在四野做大模型项目，项目成员完全变成了交付型的，且充满了例会，评审，汇报。很多时候做实验都要申请。团队需要对接终端小艺，华为云，ICT等诸多业务线，交付压力不小。
5. 诺亚研发的盘古模型早期内部代号叫做“盘古智子”，一开始只有内部需要申请试用的网页版，到后续迫于压力在welink上接入和公测开放。

这些天发生关于质疑盘古大模型抄袭千问的事情闹的沸沸扬扬。作为一个盘古团队的成员，我最近夜夜辗转反侧，难以入眠。盘古的品牌受到如此大的影响，一方面，我自私的为我的职业发展担忧，也为自己过去的努力工作感到不值。另一方面，由于有人开始揭露这些事情我内心又感到大快人心。在多少个日日夜夜，我们对内部某些人一次次靠着造假而又获得了无数利益的行为咬牙切齿而又无能为力。这种压抑和羞辱也逐渐消磨了我对华为的感情，让我在这里的时日逐渐浑浑噩噩，迷茫无措，时常怀疑自己的人生和自我价值。

我承认我是一个懦弱的人，作为一个小小的打工人，我不仅不敢和王云鹤等内部手眼通天的人做对，更不敢和华为这样的庞然大物做对。我很怕失去我的工作，毕竟我也有家人和孩子，所以我打心眼里很佩服揭露者。但是，看到内部还在试图洗地掩盖事实，蒙蔽公众的时候，我实在不能容忍了。我也希望勇敢一次，顺从自己本心。就算自损八百，我也希望能伤敌一千。我决定把我在这里的所见所闻（部分来自于同事口述）公布出来，关于盘古大模型的“传奇故事”：

华为确实主要在昇腾卡上训练大模型（小模型实验室有不少英伟达的卡，他们之前也会用来训练，后面转移到昇腾）。曾经我被华为"打造世界第二选择"的决心而折服，我本身也曾经对华为有深厚的感情。我们陪着昇腾一步步摸爬滚打，从充满bug到现在能训出模型，付出了巨大的心血和代价。

最初我们的算力非常有限，在910A上训练模型。那会只支持fp16，训练的稳定性远不如bf16。盘古的moe开始很早，23年就主要是训练38Bmoe模型和后续的71B dense模型。71B的dense模型通过扩增变成了第一代的135Bdense模型，后面主力模型也逐渐在910B上训练。

71B和135B模型都有一个巨大的硬伤就是tokenizer。当时使用的tokenizer编码效率极低，每个单个的符号，数字，空格，乃至汉字都会占用一个token。可想而知这会非常浪费算力，且使得模型的效果很差。这时候小模型实验室正好有个自己训的词表。姚老师当时怀疑是不是模型的tokenizer不好（虽然事后来看，他的怀疑是无疑正确的），于是就决定，让71B和135B换tokenizer，因为小模型实验室曾经尝试过。团队缝合了两个tokenizer，开始了tokenizer的更换。71B模型的更换失败了，而135B因为采用了更精细的embedding初始化策略，续训了至少1T的数据后词表总算更换成功，但可想而知，效果并不会变好。

于此同期，阿里和智谱等国内其他公司在GPU上训练，且已经摸索出了正确的方法，盘古和竞品的差距越来越大。内部一个230B从头训练的dense模型又因为各种原因训练失败，导致项目的状况几乎陷入绝境。面临几个节点的压力以及内部对盘古的强烈质疑时，团队的士气低迷到了极点。团队在算力极其有限的时候，做出了很多努力和挣扎。比如，团队偶然发现当时的38B moe并没有预期moe的效果。于是去掉了moe参数，还原为了13B的dense模型。由于38B的moe源自很早的pangu alpha 13B，架构相对落后，团队进行了一系列的操作，比如切换绝对位置编码到rope，去掉bias，切换为rmsnorm。同时鉴于tokenizer的一些失败和换词表的经验，这个模型的词表也更换为了王云鹤的小模型实验室7B模型所使用的词表。后面这个13B模型进行了扩增续训，变成了第二代38B dense模型（在几个月内这个模型都是主要的盘古中档位模型），曾经具有一定的竞争力。但是，由于更大的135B模型架构落后，且更换词表模型损伤巨大（后续分析发现当时更换的缝合词表有更严重的bug），续训后也与千问等当时国内领先模型存在很大差距。这时由于内部的质疑声和领导的压力也越来越大。团队的状态几乎陷入了绝境。

在这种情况下，王云鹤和他的小模型实验室出手了。他们声称是从旧的135B参数继承改造而来，通过训练短短的几百B数据，各项指标平均提升了十个点左右。实际上，这就是他们套壳应用到大模型的第一次杰作。华为的外行领导内行，使得领导完全对于这种扯淡的事情没有概念，他们只会觉得肯定是有什么算法创新。经过内部的分析，他们实际上是使用Qwen 1.5 110B续训而来，通过加层，扩增ffn维度，添加盘古pi论文的一些机制得来，凑够了大概135B的参数。实际上，旧的135B有107层，而这个模型只有82层，各种配置也都不一样。新的来路不明的135B训练完很多参数的分布也和Qwen 110B几乎一模一样。连模型代码的类名当时都是Qwen，甚至懒得改名。后续这个模型就是所谓的135B V2。而这个模型当时也提供给了很多下游，甚至包括外部客户。

这件事对于我们这些认真诚实做事的同事们带来了巨大的冲击，内部很多人其实都知道这件事，甚至包括终端和华为云。我们都戏称以后别叫盘古模型了，叫千古吧。当时团队成员就想向bcg举报了，毕竟这已经是重大的业务造假了。但是后面据说被领导拦了下来，因为更高级别的领导（比如姚老师，以及可能熊总和查老）其实后面也知道了，但是并不管，因为通过套壳拿出好的结果，对他们也是有利的。这件事使得当时团队几位最强的同事开始心灰意冷，离职跑路也逐渐成为挂在嘴边的事。

此时，盘古似乎迎来了转机。由于前面所述的这些盘古模型基本都是续训和改造而来，当时诺亚完全没有掌握从头训练的技术，何况还是在昇腾的NPU上进行训练。在当时团队的核心成员的极力争取下，盘古开始了第三代模型的训练，付出了巨大的努力后，在数据架构和训练算法方面都与业界逐渐接轨，而这其中的艰辛和小模型实验室的人一点关系都没有。

一开始团队成员毫无信心，只从一个13B的模型开始训练，但是后面发现效果还不错，于是这个模型后续再次进行了一次参数扩增，变成了第三代的38B，代号38B V3。想必很多产品线的兄弟都对这个模型很熟悉。当时这个模型的tokenizer是基于llama的词表进行扩展的（也是业界常见的做法）。而当时王云鹤的实验室做出来了另一个词表（也就是后续pangu系列的词表）。当时两个词表还被迫进行了一次赛马，最终没有明显的好坏结论。于是，领导当即决定，应该统一词表，使用王云鹤他们的。于是，在后续从头训练的135B V3（也就是对外的Pangu Ultra），便是采用了这个tokenizer。这也解释了很多使用我们模型的兄弟的疑惑，为什么当时同为V3代的两个不同档位的模型，会使用不同的tokenizer。

我们打心眼里觉得，135B V3是我们四纵团队当时的骄傲。这是第一个真正意义上的，华为全栈自研，正经从头训练的千亿级别的模型，且效果与24年同期竞品可比的。写到这里我已经热泪盈眶，太不容易了。当时为了稳定训练，团队做了大量实验对比，并且多次在模型梯度出现异常的时候进行及时回退重启。这个模型真正做到了后面技术报告所说的训练全程没有一个loss spike。我们克服了不知道多少困难，我们做到了，我们愿用生命和荣誉保证这个模型训练的真实性。多少个凌晨，我们为了它的训练而不眠。在被内部心声骂的一文不值的时候，我们有多么不甘，有多少的委屈，我们挺住了。

我们这帮人是真的在为打磨国产算力底座燃烧自己的青春啊.....客居他乡，我们放弃了家庭，放弃了假期，放弃了健康，放弃了娱乐，抛头颅洒热血，其中的艰辛与困苦，寥寥数笔不足以概括其万一。在各种动员大会上，当时口号中喊出的盘古必胜，华为必胜，我们心里是真的深深被感动。

然而，我们的所有辛苦的成果，经常被小模型实验室轻飘飘的拿走了。数据，直接要走。代码，直接要走，还要求我们配合适配到能一键运行。我们当时戏称小模型实验室为点鼠标实验室。我们付出辛苦，他们取得荣耀。果然应了那句话，你在负重前行是因为有人替你岁月静好。在这种情况下，越来越多的战友再也坚持不下去了，选择了离开。看到身边那些优秀的同事一个个离职，我的内心又感叹又难过。在这种作战一样的环境下，我们比起同事来说更像是战友。他们在技术上也有无数值得我学习的地方，堪称良师。看到他们去了诸如字节Seed，Deepseek，月之暗面，腾讯和快手等等很多出色的团队，我打心眼里为他们高兴和祝福，脱离了这个辛苦却肮脏的地方。我至今还对一位离职同事的话记忆犹新，ta说："来这里是我技术生涯中的耻辱，在这里再呆每一天都是浪费生命"。话虽难听却让我无言以对。我担心我自己技术方面的积累不足，以及没法适应互联网公司高淘汰的环境，让我多次想离职的心始终没有迈出这一步。

盘古除了dense模型，后续也启动了moe的探索。一开始训练的是一个224B的moe模型。而与之平行的，小模型实验室也开启了第二次主要的套壳行动（次要的插曲可能还包括一些别的模型，比如math模型），即这次流传甚广的pangu pro moe 72B。这个模型内部自称是从小模型实验室的7B扩增上来的（就算如此，这也与技术报告不符，何况是套壳qwen 2.5的14b续训）。还记得他们训了没几天，内部的评测就立刻追上了当时的38B V3。AI系统实验室很多兄弟因为需要适配模型，都知道他们的套壳行动，只是迫于各种原因，无法伸张正义。实际上，对于后续训了很久很久的这个模型，Honestagi能够分析出这个量级的相似性我已经很诧异了，因为这个模型为了续训洗参数，所付出的算力甚至早就足

够从头训一个同档位的模型了。听同事说他们为了洗掉千问的水印，采取了不少办法，甚至包括故意训了脏数据。这也为学术界研究模型血缘提供了一个前所未有的特殊模范吧。以后新的血缘方法提出可以拿出来溜溜。

24年底和25年初，在Deepseek v3和r1发布之后，由于其惊艳的技术水平，团队受到了巨大的冲击，也受到了更大的质疑。于是为了紧跟潮流，盘古模仿Deepseek的模型尺寸，开启了718B moe的训练。这个时候，小模型实验室再次出手了。他们选择了套壳Deepseekv3续训。他们通过冻住Deepseek加载的参数，进行训练。连任务加载ckpt的目录都是deepseekv3，改都不改，何其嚣张？与之相反，一些有真正技术信仰的同事，在从头训练另一个718B的moe。但其中出现了各种各样的问题。但是很显然，这个模型怎么可能比直接套壳的好呢？如果不是团队leader坚持，早就被叫停了。

华为的流程管理之繁重，严重拖累了大模型的研发节奏，例如版本管理，模型血缘，各种流程化，各种可追溯。讽刺的是，小模型实验室的模型似乎从来不受这些流程的约束，想套壳就套壳，想续训就续训，算力源源不断的伸手拿走。这种强烈到近乎魔幻的对比，说明了当前流程管理的情况：只许州官放火，不许百姓点灯。何其可笑？何其可悲？何其可恶？何其可耻！

HonestAGI的事情出来后，内部让大家不停的研讨分析，如何公关和"回应"。诚然，这个原文的分析也许不够有力，给了王云鹤与小模型实验室他们狡辩和颠倒黑白的机会。为此，这两天我内心感到作呕，时时怀疑自己的人生意义以及苍天无眼。我不奉陪了，我要离职了，同时我也在申请从盘古部分技术报告的作者名单中移除。曾经在这些技术报告上署名是我一生都无法抹除的污点。当时我没想到，他们竟然猖狂到敢开源。我没想到，他们敢如此愚弄世人，大肆宣发。当时，我也许是存了侥幸心理，没有拒绝署名。我相信很多扎实做事的战友，也只是被迫上了贼船，或者不知情。但这件事已经无法挽回，我希望我的余生能够坚持扎实做真正有意义的事，为我当时的软弱和不坚定赎罪。

深夜写到这里，我已经泪流满面，泣不成声。还记得一些出色的同事离职时，我苦笑问他们要不要发个长长的心声惯例帖，揭露一下现状。对方说：不了，浪费时间，而且我也怕揭露出来你们过的更糟。我当时一下黯然神伤，因为曾经共同为了理想奋斗过的战友已经彻底对华为彻底灰心了。当时大家调侃，我们用着当年共产党的小米加步枪，组织却有着堪比当年国民党的作风。

曾几何时，我为我们用着小米加步枪打败洋枪洋炮而自豪。

现在，我累了，我想投降。

其实时至今日，我还是真心希望华为能认真吸取教训，能做好盘古，把盘古做到世界一流，把昇腾变成英伟达的水平。内部的劣币驱逐良币，使得诺亚乃至华为在短时间内急剧流失了大量出色的大模型人才。相信他们也正在如Deepseek等各个团队闪耀着，施展着他们的抱负才华，为中美在AI的激烈竞赛中奉献力量。我时常感叹，华为不是没有人才，而是根本不知道怎么留住人才。如果给这些人合适的环境，合适的资源，更少的枷锁，更少的政治斗争，盘古何愁不成？

最后：我以生命，人格和荣誉发誓，我写的以上所有内容均为真实（至少在我有限的认知范围内）。我没有那么高的技术水平以及机会去做详尽扎实的分析，也不敢直接用内部记录举证，怕因为信息安全抓到。但是我相信我很多曾经的战友，会为我作证。在华为内部的兄弟，包括我们曾经服务过的产品线兄弟们，相信本文的无数细节能和你们的印象对照，印证我的说法。你们可能也曾经被蒙骗，但这些残酷的真相不会被尘封。我们奋战过的痕迹，也不应该被扭曲和埋葬。

写了这么多，某些人肯定想把我找出来，抹杀掉。公司搞不好也想让我噤声乃至追责。如果真的这样，我，乃至我的家人的人身乃至生命安全可能都会受到威胁。为了自我保护，我近期每天会跟大家报平安。

如果我消失了，就当是我为了真理和理想，为了华为乃至中国能够更好地发展算力和AI而牺牲了吧，我愿埋葬于那片曾经奋斗过的地方。

诺亚，再见

2025年7月6日凌晨 写于深圳