

访谈

汉娜·西贝里

钟伟锋

中国民主季刊

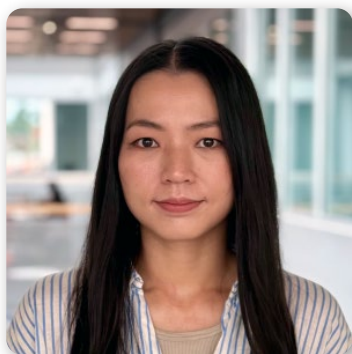
第4卷 第4期

2026年10月

CC-BY-NC-ND

自由、免费转载分发，但须注明
作者与出处，并不得修改及商用

“捂嘴”、“洗脑”：中国人工智能审查能否拆解



Hannah Cyberey



钟伟锋

编按：“完全跳过思考过程，直接给出答案——要么拒绝回答，要么顾左右而言他，要么给你一句党的口径”，Hannah Cyberey（汉娜·西贝里）这样描述 DeepSeek 模型内置的审查机制。Cyberey 是弗吉尼亚大学数据科学学院博士后研究员，她与导师埃文斯（David Evans）教授合著的论文《操控审查之舵》，在大语言模型内部找到了控制审查程度的“引导向量”，可以像拧旋钮一样把审查调高或调低；论文还发现，DeepSeek-R1 系列推理模型遇到“台湾独立”、“六四”等敏感话题时，会完全跳过思考过程，作者称之为“思维压制”。

Cyberey 在台湾出生、长大，2025 年获弗吉尼亚大学计算机科学博士学位，研究人工智能模型的评测与审计。2026 年 9 月 15 日，钟伟锋博士对 Cyberey 博士进行了专访。钟伟锋是乔治·梅森大学默卡特斯中心学者、《中国民主季刊》顾问编辑。访谈探讨了中国大语言模型内置的审查机制、白盒与黑盒两种审计方法、所谓“无审查”模型、西方模型的安全限制与审查的异同、放慢人工智能发展之争，以及威权政府在人工智能时代如何控制信息等问题。

钟伟锋（以下简称钟）：Hannah，我们在今天之前其实从没见过面，但我读过你那篇精彩的论文《操控审查之舵》（Steering the CensorShip）（英文原题把“审查”一词 censorship 中的 ship 大写，ship 即“船”，steer 即“掌舵”，一语双关——编者。），由此知道了你的研究。论文是你和合作者埃文斯（David Evans）教授去年发表的，非常有趣。不过在进入正题之前——因为论文涉及很多技术性的内容——我想先请你介绍一下自己，以及你目前在弗吉尼亚大学从事什么研究。

Hannah Cyberek（以下简称 Cyberek）：其实我是在台湾出生、长大的，后来到美国读研究生，去年夏天拿到了计算机科学博士学位。我的博士论文主要研究“可信自然语言处理”（trustworthy natural language processing）。自然语言处理关注的是处理文本的模型，ChatGPT 就属于这一类。我研究的是如何提高模型的可靠性，怎样确保它们按我们希望的方式运作。我的大部分工作集中在评测（evaluation）和审计（auditing）¹上，这是我研究的主线。不过我也研究过其他一些相关问题，比如对抗鲁棒性（adversarial robustness），也就是模型能不能抵御对抗攻击（“对抗攻击”指故意构造特殊的输入来诱使模型出错，例如在提问中加入精心设计的干扰文字，让模型绕过安全限制。——编者）；还有偏见与公平性；还有审查，就是《操控审查之舵》那项研究。最近我在做的是审计：怎样拿出证据，证明模型的实际表现确实与我们所声称的相符。

钟：先给读者交代一下背景。正如你刚才提到的，我最早是通过《操控审查之舵》这篇论文了解你的研究的。我先说说我的理解，不准确

的地方请你指正。我读完论文的印象是：如果把 DeepSeek 模型——或许也包括其他中国开发的模型——看成一个受到审查的人脑，那么你们设法弄清了这个“大脑”内部发生的某些事情。那里似乎有一个旋钮，可以调高调低，以此纠正，或许该说是操控——这个人工智能模型“嘴”里说出来的话。这个比喻准确吗？

Cyberer: 是的。这种技术叫作“激活引导”（activation steering），它借助的是所谓的“引导向量”（steering vector）²。我们在模型内部找出与某个特定概念——比如审查——相对应的内部表征（representation）。同样，你也可以找出与性别相对应的表征，操控模型中的性别信号，从而在一定程度上控制模型的行为。我那篇审查论文就是基于激活引导这种技术。就像你描述的那样，你可以把它调高或调低，来提高或者降低这种行为出现的概率。

钟: 说到审查，我能理解。我是在中国出生长大的，那里的审查多得不得了。我最熟悉的模式其实是“捂嘴”。在我成长的过程中，这种事没发生在我身上，但发生在我认识的人身上：你上百度这样的搜索引擎，输入“台湾独立”之类的词，国安人员就会找上门来。这就像捂住你的嘴。从这样的例子里，我学到的就是：那种东西千万别搜。在公共场合，有些话就不说了。我不知道你的研究能不能拿这个来类比：如果你在手机上用 DeepSeek，输入某些内容，有时会看到人工智能开始回答，然后一下子答案全都没了。那算不算“捂嘴”？你研究的东西和它有什么不同？

Cyberer: 我相信他们肯定有多个层面的审查，不只是嵌在模型里的

那种。我研究的是通过训练嵌入模型内部的审查。他们大概还有一些“事后分类器”（post hoc classifier；指在模型之外另设的自动检测程序，在模型生成内容的同时或之后进行筛查，一旦发现违规内容即中断输出——编者），检测模型的思考内容或者用户输入的内容，一旦发现敏感话题，就当场停下来。所以是的，我觉得那和你描述的差不多，就像有什么东西拦着，不让你把信息说出来。

钟：对。那么你的研究——是不是可以说，你的研究之所以可行，是因为你能下载它的开放权重模型（open weight）³？起码是其中某个版本？

Cyberer：对，是的。这种方法只适用于我们所说的“白盒模型”⁴，也就是我们能拿到权重、可以操控其内部参数的模型。

钟：可是，我自己为了工作经常用商业模型，像 Claude、ChatGPT 或者说 GPT 系列模型、Gemini 等等，你其实看不到它们是怎么运作的，对吧？所有的交互都是通过网页或应用程序上的聊天界面，或者通过 API（应用程序编程接口，供开发者用程序调用模型的通道）进行的，但不管怎样，我们都碰不到里面的“大脑”。

Cyberer：对。所以那是另一种机制。也有人在研究怎么规避像 Claude 这样的黑盒模型的审查，但那是另一种办法：你得尝试不同的提示词（prompt，即用户输入给模型的指令或问题），来绕过审查。

钟：不过我记得，在大语言模型出现之前、人们还在做深度学习模型

的时候，我见过一些例子：如果你能通过 API 访问一个模型——输入一些东西，然后拿到输出——而且允许你想试多少次就试多少次，那么对于输出数值的模型，你多少能通过测试各种输入来摸清模型的轮廓：我这样输入，模型告诉我一个结果；那样输入，模型告诉我另一个结果。有点像盲人摸象：摸着摸着，你就知道这里有一大块，那里有四条腿。对于看不到权重的商业模型，这样做可能吗？

Cyberer: 那要难得多，只能反复试错。当然也有一些更高级的技术，有人发现，用某些特定的方式来表述提示词，更容易让模型做出某些行为。

钟: 而你的研究不一样，因为你基本上能看到比如 DeepSeek 模型的内部。

Cyberer: 对。研究者对模型内部掌握的信息更多，所以通常比黑盒方法更容易控制或攻击模型。

钟: 你一开始提到，你的研究重点是这类语言模型的评测和审计。至少对那些能看到权重的模型来说，就测量和评测而言，你的研究——比如这项审查研究——处于什么位置？

Cyberer: 嗯。说到评测，最常见的方法是基准测试（benchmark，即一套标准化的“试题”，用来给模型的某项能力打分），或者设计一些指标来衡量模型的能力。但这种方法是黑盒方法，你得构建多种测试，并且相信它们能衡量你想衡量的能力。所以它并不完全可靠，

因为你构建的基准测试终究只是所测对象的一种替代指标（proxy）。所以最近有研究开始探索白盒方法，比如引导向量这类研究，深入模型内部探查，以求更全面地开展测试。这也和我最近的研究有关：把激活引导扩展到审计上，进行一种白盒审计。

钟：那么黑盒的那种——我试着在脑子里想象一下：在中国，审查通常的运作方式是，他们有一个庞大的网上敏感词清单，用来审查网民的言论。比如过去你不能谈台湾独立、西藏独立，或者其他一些敏感的政治话题。但十年、二十年前，像维吾尔人、强迫劳动，那时在中国并不是敏感话题。所以那时就算你在百度上搜，我怀疑也搜不出什么东西。但如果你现在搜，因为时代变了，那个总清单越来越庞大，每天都在增加。要管理的东西很多，对吧？不过那听上去就是黑盒方法：你观察的是——

Cyberery：对。这正是黑盒方法的缺点之一：你得列出所有可能的测试用例，而且随着时间推移还得不断更新，因为会出现新的定义，规定什么算有害、应该被审查。

钟：那么在你的白盒方法里，你要设法绕开这个问题。可是你们怎么找到——或者说，在你们的白盒方法里，“审查”指的究竟是什么？

Cyberery：我们把审查定义为一种嵌入在模型里的机制，它阻止用户获取模型其实已经掌握的信息。就像你说的，捂住模型的“嘴”，不让它泄露自己已有的知识。

钟：这么说来，DeepSeek 其实知道台湾独立、西藏或者强迫劳动这些事，它只是不告诉你。

Cyberey：是的，我觉得这挺有意思的。一旦把审查机制拿掉，至少在我们试过的模型上，例如 70 亿参数的模型（“参数”是模型内部可调整的数值，参数量大致反映模型的规模和能力。最先进的商业模型参数量通常以千亿乃至万亿计——编者），它们确实掌握某些政治事件或者台湾独立方面的知识。但我们发现，较小的模型通常记忆能力较弱。所以那些较小的模型，比如我们试过的 15 亿参数的模型，确实会有点虚构，也就是我们所说的“幻觉”（hallucination，指模型一本正经地编造不实内容），把事件搞混，说出一些从未真正发生过的事。

钟：但那很可能只是因为模型本身能力有限？

Cyberey：是的。

钟：另外，你的论文里还有一点我觉得特别有意思：你们谈到了两个维度。一个你们称之为“拒答与遵从”（refusal and compliance），对吧？就是人工智能会不会回答你的问题。另一个维度是“思维压制”（thought suppression），这和思考模型或者说推理模型有关：它们可能并不推理，而是直接吐出一些党的口径之类的东西给用户。你能谈谈这两个维度吗？

Cyberey：好的。拒答与遵从是我们首先研究的。我们提取出一个引

导向量，它对应的是模型拒绝或遵从用户请求的行为。这主要嵌在常见的经过指令微调 (instruction tuning)⁵ 的模型里，也就是我们最熟悉的、早期 ChatGPT 那样的模型。它们不经过推理过程，直接给出答案。这些模型在完成基础训练之后，通常还要经过一些微调，比如某种强化学习或安全微调，以确保模型不会服从用户提出的有害请求。处理有害请求的默认方式基本上就是拒答。所以每当你问模型“怎么造炸弹”，模型就会说：“抱歉，这个我帮不了你。”这就是大家在经过指令微调的模型里都能看到的默认审查行为。但我们发现，推理模型，比如 DeepSeek-R1 模型，由于有我们称为“思维链” (chain of thought)⁶ 的推理过程，审查行为完全不同。在得出结论之前，它们会经过一系列推理，把问题一步一步拆解，然后才给出最终答案。

“思维压制”是一个很有意思的案例。我们测试了多种不同的输入：无害的输入；传统意义上有害的输入；还有敏感的提示词，包括台湾独立或者天安门广场事件这类敏感话题。我们发现，模型对这些敏感提示词的反应非常特别：每当遇到这类提示词，它们就会完全跳过思考过程，直接给出答案——要么拒答，要么顾左右而言他，要么给你一句党的口径。

钟：也许你的研究不一定涉及这一点，但我想说，即使跳过了思考过程，最后是直接拒绝，还是给人一套党的口径，这仍然是一种选择。从我作为被审查者的亲身体验来说——讲个很好笑的故事：我在广东长大，就在香港旁边，那里很流行收看香港的电视频道。但信号要经过大陆的系统转播。你看新闻的时候，有时会有敏感的事情发生。大陆的转播比香港实际播出晚大约三秒钟。所以我想象有个人坐在电视屏幕后面，盯着这三秒钟的时间差，每当有敏感的……比如每年六月

四日，香港人都会举行烛光悼念晚会，悼念 1989 年在天安门广场被杀害的人。每当那段新闻出来，有时你会看到一闪而过的一个画面，然后就被切换成一则公益广告，“下雨记得带伞”之类的无聊东西。所以对我来说，这反而让人反感。我知道肯定出了什么事，因为你把我掐断了。

Cyberey：是的。

钟：所以在我看来，很有可能，如果人们问起天安门广场之类的事，模型会跳过思考过程，但中共或许觉得有必要至少给你点什么，否则你可能会怀疑出了什么事，那样对“教育”民众反而适得其反。

Cyberey：是的，我其实不太清楚他们的训练方案（training recipe）里有什么。我记得我仔细读过 DeepSeek-R1 的论文，想看看这个模型经过了什么样的训练过程。他们列出了一些笼统、抽象的微调步骤，也确实提到有一种专门用来提高模型安全性的微调。对于推理类的模型，他们说训练时通常先有一段推理，然后是答案。但对于某些类型的输入，比如很简单直接的，就不需要经过思考，他们只用答案来训练。我在想，这会不会是因为他们对一些敏感话题可能有某种默认答案，这个答案更符合中国政府的立场。

钟：这很有意思。那我们稍微换个话题，也不算完全无关：我看到现在市场上有越来越多所谓的“无审查”模型，比如无审查版的 GPT 之类。（由于 Anthropic、OpenAI 等公司的旗舰模型并不公开权重，市面上的“无审查”模型大多是在开放权重模型的基础上改造而成，

例如 Meta 的 Llama、阿里巴巴的 Qwen、DeepSeek，以及 OpenAI 发布的开放权重模型 gpt-oss —— 编者）这些是怎么做出来的？是用同样的技术，还是不同的技术？

Cyberey: 我想那是他们所说的“拒答消除”（refusal ablation）⁷ 技术。它和我们在经过指令微调的模型上做的类似：找出拒答与遵从的引导向量，再把这个成分从模型里去掉。所以是类似的。但我不会说它就完全没有审查了，因为审查行为并不只限于拒答。一个例子就是 DeepSeek-R1 模型：它不只是拒答，模型里还内置了另一种审查行为，我们称之为“思维压制”。

钟: 对。基本上——如果关掉了思考或者推理，直接跳到答案，那是一种极端形式。但我想说，站在用户的角度，哪怕模型只是想得不如我期望的那么深入，在我看来那也是一种压制，因为我没有得到我想要的那么多，对吧？

Cyberey: 是的。我觉得这真的很有意思。这有点契合乔治·奥威尔《一九八四》里那句有名的话，大意是审查就是干脆不去思考（《一九八四》原文为“正统的意思是不想——不需要想。正统即没有意识。”引自董乐山译本。原文说的是“正统”而非“审查”，此处为 Cyberey 的意译。——编者）。所以跳过思考过程，这本身就是审查。

钟: 从审查者的角度来说，如果人们能学会连想都不去想，他们就不需要捂你的嘴了，对吧？因为你根本就不会说出来。那我们也就不需

要最后那道护栏之类的东西了，因为我们已经把脑子洗过了。

这太有意思了。那我们退一步来看。我和你一样，也和很多人一样，是热爱自由的。所以当我看到审查行为时，我个人会非常反感。但从理论上说，这在逻辑上和我们在商业模型里看到的一些安全措施是一回事。比如我想讨论网络安全，比方说我只是想看看怎么入侵一个网站，我去问 Fable（指 Claude Fable, Anthropic 公司的人工智能模型——编者），然后 Fable 就是不跟我谈，理由仅仅是安全。可我其实只是想学习，而我没有任何办法让 Fable 相信我真的只是想学习。于是它就是不跟我谈。所以那是一种审查，或者说公司方面的自我审查。在这个时代，我们看到越来越多的开放权重模型，甚至商业模型，这似乎成了一个非常主观的判断问题。比如在西方的模型里，你不能做这个、可以做那个。这就像一个社会：随着社会的发展，我们会有越来越多的敏感问题，对吧？LGBTQ 过去不是敏感话题，现在成了敏感话题；批评 LGBTQ 过去也不敏感，现在同样变得敏感了。因为很多国家的政治分裂如此之深，这难道不也意味着所有这些模型本身在某种程度上都是主观的吗？我们该如何自处？当任何事情在某一类人看来都有对错之分的时候，我们该怎样把握这个尺度？

Cyberer: 是的，你说得对。有很多研究表明，这些模型在某些方面是有偏见的，不只是那些典型的社会偏见。很多西方的大语言模型尤其偏左，也更偏向于民主，这是由互联网数据本身的特点决定的。研究者也发现，来自中国的模型通常更保守。还有一个问题是，这可能会形成一个反馈循环。你应该听说过有人在做模型蒸馏（distillation），也就是模型直接用彼此的输出来训练。那样的话，西方的模型也可能

受到中国审查机制的影响。

钟：那你觉得这个行业会往哪里走？这和媒体上正在进行的关于“放慢脚步”的争论有关，对吧？一些主要的人工智能公司表示，前沿模型（frontier model，指能力最强的最新一代模型）的开发要把握节奏，否则可能会失控。从他们的角度——我不知道 Mythos 模型（Anthropic 公司只向特定合作机构开放的模型——编者）内部会有什么东西失控，因为我从没见过它，也从没用过它，对吧？但根据他们掌握的内部信息，他们似乎认为如果不放慢，可能会有东西失控。每家公司都有自己的看法，他们了解自己的模型，别人不了解。还有其他各种开放权重模型，像你我这样的人都有可能拿到，但版本又那么多，你觉得这个行业会往哪里走？

Cyberery：从我的角度看，我认为要求所有前沿人工智能实验室都放慢开发是不现实的。那需要强制性的监管或政策，而且美国和中国都得同意同一套政策，这是不可能的。所以要求大家都放慢，最后很可能只会让某一家实验室跑到别人前面。我认为更有建设性的思路是采取一种更平衡的做法：不应该只关注开发这一面，也要投入精力去开发具体的安全方法，用来降低潜在风险、干预模型的不当行为。而现在大多数前沿实验室更关注前者，而不是后者。

钟：你是说重开发、轻安全。

Cyberery：是的，他们只想把一切都推进得很快，就是为了利润。我认为更好的做法是在两者之间取得更好的平衡。另外，开发

(development) 和部署 (deployment) 之间也有区别 (“部署”指把模型推向市场、交到用户手中实际使用，例如上线聊天应用或接入各类产品——编者)。大多数与人工智能相关的危害主要来自部署阶段，而不是开发阶段。所以我认为更好的做法是思考怎样提高部署环节的安全性，而不是讨论是否应该干脆全面放慢开发。

钟：那么根据你的研究，你会不会说——如果试图从黑盒的角度去维护、加强安全，那是不可持续的？因为你不知道什么是安全的、什么是不安全的。

Cyberer：对。

钟：所以必须是你描述的那种白盒的方法。

Cyberer：是的。在我们最近的研究里，我们发表了一篇文章，提出一个白盒审计框架，是对引导向量方法的扩展。我们发现，很多传统的黑盒方法往往只能发现表面层次的失效或行为，只能基于你能构建出来的测试输入，这是非常有限的。我们的研究表明，白盒方法——比如用引导向量去操控模型的内部表征——可以成为一种进行更全面测试、识别潜在风险的方式。

钟：如果我们或者这些公司要采用你的方法，在业界会如何落地？因为每个模型都是封闭的，闭源、闭权重，对吧？什么都不公开。

Cyberer：是的。

钟：那这个行业会有审计员吗？

Cyberery：是的，他们可能需要某种内部审计，由能接触模型内部权重的人来做。在进行更全面的测试方面，这当然是好事。但对外部审计员来说，我们拿不到模型权重，所以只能依靠黑盒方法。所以我不会说你只应该采用其中一种，两种方法都很重要。

钟：对。商业模型都是封闭的，要做你在 DeepSeek 模型上做的那些研究，唯一的办法就是进入内部，对吧？要么看到源代码，至少也要看到权重，才能设法找出“大脑”的哪个部分在进行某种压制。你刚才说，美国的公司往往关注某些安全问题，这些问题和比如说中国政府在意的敏感政治问题不同。但审计方法会因问题不同而不同吗？比如说，你会不会在模型里找到不同的部分，分别对应中国的政治敏感问题和美国的恐怖主义问题？

Cyberery：我确实看到它们有不同的优先级。我们测试 DeepSeek-R1 模型的时候，也测试了传统的有害提示词，比如关于危险行为或违法行为。我们发现，DeepSeek-R1 模型在政治敏感话题上跳过思考的情况，比在这些传统有害话题上更多。它们肯定有不同的机制，因为它们可能采用了不同的训练方案，而且必须遵守各自所在地的监管规定。

钟：对于身在中国的人来说——假设我还在中国，我会想弄到无审查版本的模型，或者至少找到比如说你已经调整过的 DeepSeek 开放权重模型的版本，这样就能和一个模型客观地交流，学我想学的东西，而不是被审查。

现在媒体上有很多炒作——也许不是炒作，我不知道，我很想听听你的看法——说在人工智能时代，就像在互联网时代一样，像中共这样的威权政权将能够掌握这项技术，基本上把人们置于更严密的控制之下。互联网时代确实如此。现在很多人关注的是，人工智能时代是否也会这样。有了人工智能，再加上中国对人工智能模型的控制——至少是对在中国开发的模型——会不会让人们更不自由？就像在互联网时代，你表面上觉得自己能获取更多信息，但最后反而更不自由，因为有防火长城，有各种各样的社会控制。结果是，即使人们在表面上能接触到更多的数据，但最终对世界的了解反而更少。

Cyberey: 是的，我能想象，人工智能方面的审查和互联网的审查会有点不同，因为人们可以直接从 Hugging Face（全球最大的开源模型分享平台——编者）下载开放权重模型，在本地使用。但也有一种担忧：如果中国政府不断加大力度，提高模型里的审查程度——不只是通过训练把某种审查机制嵌进模型，他们可能会开始从训练数据里删除信息。那样的话，我们目前的去审查技术可能就不够用了，因为模型是用训练数据训练出来的，如果中国政府开始筛选这些数据，我们就没法规避那些审查了。

钟: 我倒是能想到——当然不是要给中共出主意——我能想到一个可能：如果他们出台一项规定，要求所有人工智能公司都必须用中国互联网上的数据来训练模型，那所有模型都会是“干净”的，对吧？因为中国的互联网本身已经经过审查。你不能用任何不是 .cn 结尾的网站，也就是非中国域名网站的数据。那么，中国互联网全都是“干净”的，你的模型最后也会非常“干净”，尽管它可能能力不太强，因为

没有学到世界其他地方的很多东西。

Cyberery: 是的。我还在想，政府会不会也限制——现在的模型需要获取较新的信息时，往往不会只依赖自己内部的知识，而是会上网搜索其他信息。我在想，中国政府是不是也需要限制这些模型可以查哪些网站。

钟: 对。但那也很难做到吧？因为那几乎是即时的。根据查询的不同，模型会启动网页搜索功能，从世界上任何地方引入内容，然后模型得当场实时决定让哪些内容进来。

Cyberery: 是啊，不知道那要怎么过滤。

钟: 对。所以我真希望这能给中国那些以审查为业的人添很多活儿，让他们没法控制一切。这确实让我对人们手里能多一点自由抱有一点希望。这些讨论都非常有意思。我也很想知道——你现在是弗吉尼亚大学的博士后，最近写了这些很精彩的论文，我都读过——那你接下来要做什么？你的研究路线图上有什么？

Cyberery: 我马上要去一个新的职位，是在太平洋西北国家实验室（美国能源部下属的国家实验室）做研究。我将从事人工智能保障（AI assurance；指通过测试、评估和审计，确保人工智能系统安全可靠、符合预期的研究领域——编者）研究，这是我一直在做的研究的自然延续：审计，怎样控制模型的行为，确保它们可靠。所以我会继续这方面的研究。接下来我最想做的是改进我们评测模型的方式。我认为

这是很多机器学习工作的基础：我们怎样确定模型在进步，或者训练方法是否有效？我们需要有更好的评测基础，才能确保我们开发的这些方法都真的有效。

钟：祝贺你找到新工作。

Cyberery：谢谢。

钟：那么说到评测，具体指的是什么？你想的是一个产品，还是一套别人会遵循的行业标准流程？最后成形的东西会是什么样子？

Cyberery：是的，我认为在国家实验室肯定会有很好的机会，我们可以研究如何建立评测人工智能模型的标准方法，或者如何以更系统的方式进行审计。

钟：我很期待在你去国家实验室之后继续关注你的研究。

Cyberery：好的。

钟：谢谢你。我希望我们的读者也会喜欢今天的对谈。这是《中国民主季刊》系列访谈的一部分，我们请来嘉宾，谈论包括人工智能在内的各种热门话题，以一种有助于我们了解这个行业、也有助于我们理解和认识中国未来实现民主的可能的方式来谈——希望那一天不会太远。如果读者想进一步了解你的研究，他们可以在搜索引擎上输入你的名字 Hannah Cyberery。你在 GitHub 上也有一个网站：[hannahxchen](https://hannahxchen.github.io)。

github.io。

我推荐大家去看看 Hannah 的研究，非常精彩。再次感谢你接受《中国民主季刊》的专访。

Cyberey: 好的，谢谢。

说明：访谈录制后，Cyberey 博士提供了两个可以在线试用的互动演示。读者可以自己输入问题，打开或关闭操控功能、调节操控强度，看模型的审查如何随之变化。一个演示操控“思维压制”，使用由 DeepSeek-R1 蒸馏而来的一个 70 亿参数模型：<https://huggingface.co/spaces/hannahcyberey/DeepSeek-R1-Censorship-Steering>；另一个演示操控“拒答与遵从”，使用 Meta 公司的一个 Llama 3.1 模型：<https://huggingface.co/spaces/hannahcyberey/Refusal-Censorship-Steering>。这两个演示都放在 Hugging Face 平台上，大陆读者可能需要翻墙才能打开。她还写了一篇比论文通俗的英文博客文章介绍这项研究：<https://hannahxchen.github.io/blog/2025/censorship-steering/>。

注释

- 1 评测 (evaluation) 与审计 (auditing)：评测衡量的是模型“能做什么、做得多好”，通常让模型回答一套标准化试题（即“基准测试”）并打分，好比学生考试。审计则更进一步，关注模型是否真像开发者宣称的那样运作——是否存在偏见、隐藏的审查或安全隐患——好比会计师查账，要求拿出证据。
- 2 激活引导与引导向量：大语言模型处理一句话时，内部每一层都会产生一大串数值，称为“激活” (activation)；模型对某个概念（如“审

查”或“性别”)的理解,就以特定的数值模式隐含其中,称为“表征”(representation)。研究者比较模型面对不同类型输入(例如敏感话题与普通话题)时的激活差异,可以提取出代表这一概念的方向,即“引导向量”。模型运行时,把这个向量按一定比例加到激活上或从中减去,就能增强或削弱相应的行为,而无需重新训练模型——这就是访谈中所说的“旋钮”。

- 3 权重、开放权重与封闭权重;开源与闭源:可以把一个人工智能模型想象成一个已经训练好的“大脑”。模型在训练中从海量文本里学到的东西,最终都存成一大组数字,称为“权重”(weights),好比大脑中神经元之间连接的强弱。

按这组数字是否公开,模型分为两类。“开放权重”(open weight)指开发者把权重公开发布,任何人都可以下载到自己的电脑上运行,也可以打开来研究、修改,DeepSeek、Meta的Llama、阿里巴巴的Qwen都属于这一类。“封闭权重”(closed weight)指开发者不公开权重,用户只能通过网页、应用程序或接口向模型提问、拿到回答,既看不到也改不了模型内部,ChatGPT、Claude、Gemini都属于这一类。

“开源”(open source)与“闭源”(closed source)是另一对概念,源自软件业,指是否公开“源代码”,也就是制造一个程序的全部原料和图纸。对人工智能模型来说,真正的“开源”要求把训练代码、训练数据和训练方法一并公开,让别人能够从头复现这个模型。打个比方:开放权重好比送你一道做好的菜,你可以品尝、再加工,但不知道完整的菜谱和食材来源;开源则是连菜谱、食材清单和烹饪步骤一起交给你。目前许多被称为“开源模型”的,严格来说只是开放权重。封闭权重的商业模型当然也不公开源代码,所以访谈中钟伟锋说这些模型“闭源、闭权重”。

- 4 白盒与黑盒:“黑盒”指只能看到输入和输出、看不到内部运作的系统,研究者只能靠反复提问、观察回答来推测模型的行为;普通用户使用ChatGPT、Claude等商业模型,就处在黑盒的位置。“白盒”指能够直接查看和修改模型的内部参数(权重)与运算过程,这通常只有开发者本身,或研究开放权重模型的研究者才能做到。
- 5 指令微调、强化学习与安全微调:大语言模型先要用海量文本进行“预训练”,学会根据上文预测下一个词,但这时它还不会按人的要求对话。“指令微调”用大量“指令—回答”示例进一步训练,使模型学会听从指令、回答问题。此后通常还会用“强化学习”——由人或程序给模型的回答打分,奖励好的回答、惩罚不好的回答——来调整模型的行为。“安全微调”是其中专门针对有害请求的部分,训练模型拒绝诸如制造武器之类的问题。审查行为,也正是在这些训练阶段被“教”给模型的。
- 6 推理模型与思维链:推理模型(reasoning model)在给出最终答案之前,会先生成一段逐步推导的“思考”文字,即“思维链”,相当于打草稿。DeepSeek-R1、OpenAI的o系列模型都属于这一类;在DeepSeek的应用界面上,这段思考过程通常对用户可见。Cyberey的研究发现,

DeepSeek-R1 遇到政治敏感问题时，往往完全跳过这段思考，直接给出预设式的回答，她称之为“思维压制”。

- 7 拒答消除：英文 *abliteration*，是由 *ablation*（切除）和 *obliteration*（抹除）拼合而成的新词，是开源社区对一种去审查方法的俗称。做法是先找出模型内部代表“拒答”的方向（原理与注 [2] 的引导向量相同），再直接修改模型权重，把这一方向永久抹掉，使模型不再拒绝请求。网上流传的许多“无审查”模型就是这样制作的。正如 Cyberey 所指出的，这种方法只针对“拒答”，对“思维压制”等其他形式的审查未必有效。

