arXiv:2611.33050v1 [cs.CL] 24 Nov 2026

值得等待吗? 响应时间和流式传输如何影响人们对生成式人工智能回答的判断及聊天机器人参与度

Zihan Chen
史蒂文斯理工学院霍博肯，NJ，美国 zchen61@stevens.edu

Di Zhu史蒂文斯理工学院
霍博肯，NJ，美国
dzhu1@stevens.edu

Lei Nico Zheng马萨诸塞大学
波士顿分校波士顿，MA，美国
lei.zheng@umb.edu

摘要

基于大语言模型（LLM）构建的生成式人工智能聊天机器人如今以截然不同的速度提供回答：推理模型会在回复前进行思考，而响应缓存则可以立即返回存储的答案。用于对这些模型进行排名的盲测偏好投票，通常被视为对其文本质量的判断；如果交付方式也会产生影响，那么更快的回答理应更具优势。然而，关于劳动幻觉以及文本如何出现在屏幕上的研究表明，与同一答案一次性显示相比，逐字呈现、仿佛正在书写的答案反而更容易获得认可。我们利用自然使用场景中的三类变异来检验这些解释。首先，法国政府的聊天机器人竞技场 compar:IA 通过随机抽取，对重复提示词交替提供两种回答：一种是边生成边流式呈现的新生成的回答，另一种是来自同一模型、此前存储并一次性显示的答案。在相同提示词下，重放使回答赢得决定性盲选投票的概率降低了 25.8 个百分点。同一段存储文本以流式方式呈现时赢得了 57% 的决定性投票，而重放时仅赢得 29%；投票者还认为，重放的回答不够完整、更肤浅，也更常出现不正确的情况。其次，服务提供商拥塞会拖慢流式回答。在 324,029 段对话中，较慢的回答并未在投票中获得可检测的净认可，但首次交流较慢会降低访问者提出后续问题的可能性。第三，在日常使用 ChatGPT 产生的 4,378,514 条 WildChat 回复中，因拥塞而变慢的回复会降低用户继续对话的可能性。交付方式改变了盲测投票的结果，而且方向与竞技场设计者的预期相反；等待则会减少后续问题，而当两个答案都很慢时，答案之间的比较无法捕捉这种影响。这些结果关系到人工智能服务如何缓存和流式传输回答、如何规划延迟预算，以及偏好排行榜如何评估模型。

关键词 响应时间 · 延迟 · 生成式人工智能 · 大语言模型 · 聊天机器人 · 流式传输 · 缓存 · 劳动幻觉 · 运营透明度 · 用户参与度 · 人工智能评估 · 人机交互

1 引言

人工智能系统需要多长时间来回答，已经成为一种设计选择。推理模型的构建目的，是投入更多计算，因此也需要更多时间，然后才会给出回答（Guo 等人，2025年；Snell 等人，2025年）。与此同时，工程师们也在努力加快回答的生成：推测解码可以缩短生成时间（Leviathan 等人，2023），而响应缓存则能几乎即时地返回对重复提问存储的答案（Bang，2023）。聊天机器人提供商如今同时推出“快速”和“思考”模式。用户看到回答瞬间出现、看着回答逐字写出，还是等待系统思考，都是产品团队每天要做的决定，通常取决于成本。

盲测偏好投票是比较生成式人工智能模型的主要方式。公共竞答平台会向访问者展示针对同一提示词给出的两份匿名回答，并根据投票对模型进行排名（Chiang 等人，2024年；Termignon 等，2026年）。这些投票被视为对各模型文本的评价；如果交付方式可能产生影响，人们便会预期较快的回答更占优势。compar:IA（法国政府的竞技场）的设计者写道：“更快的回答可能会被认为更流畅或更可靠，而较慢的回答则可能在内容质量之外，独立地对用户偏好产生负面影响”（Termignon 等，2026年）。人们会认为缓慢生成的算法预测准确性较低（Efendi´c 等人，2020年）；大语言模型应用中的响应能力“普遍被认为至关重要”（Tan 等人，2026年）；而数十年来有关系统响应时间的研究表明，延迟越长，满意度越低、回访意愿越弱、使用越少（Miller，1968；Shneiderman，1984；Guynes，1988；Palmer，2002；Galletta 等人，2004；Nah，2004）。另一种解读则将时间视为审慎思考与用心程度的标志。如果人们认为某件产品投入了更多心力，就会对它评价更高（Kruger 等人，2004；Morales，2005）；而在一项近期的实验室研究中，等待聊天机器人的回答开始生成2秒的参与者，认为其回答不如等待9秒或20秒的参与者那样深思熟虑（Tan 等人，2026年）。

关于输出呈现方式的研究表明，对于即时出现而非在屏幕上逐步展开的回答，人们有理由预期其速度评价会恰恰相反。展示工作过程会提高人们对服务的重视程度，即使结果完全相同，这种效应被称为劳动幻觉（Buell 和 Norton，2011年）。当用户认为推荐代理在其推荐上投入了更多努力时，他们会给予其更高的评分（Tsekouras 等人，2022年）；而对产品制作过程中所投入努力的看法，有助于解释消费者为何较不重视人工智能制造的产品（Rix 等人，2025年）。关于等待界面的研究提供了第二个理由：等待期间运行的显示内容能够吸引用户注意力并使其沉浸其中（Lee 等人，2012年）；而逐字呈现的生成式人工智能回答比逐行呈现的回答更容易被接受，因为用户会感到更加沉浸（Jin 和 Gong，2026年）。在一项关于人工智能写作工具的实验中，Zhou 等人 （2025年）预期文本的评判质量不会因其在屏幕上的呈现方式而有所不同；但他们发现了一个小幅效应：在略快于阅读速度的节奏下，评分最高。目前尚不清楚这些效应是否会影响对模型进行排名的盲选，以及影响方向为何。

生成式人工智能让这个问题成为一项日常运营决策。在确立劳动幻觉的实验中，工作过程的展示是一段动画，在用户等待结果时播放，例如旅游网站准备票价时，显示正在搜索的航空公司列表（Buell 和 Norton，2011年）。在聊天机器人中，展示的就是回答本身：流式传输会随着文本生成而逐步显示内容，几乎所有回答都默认采用这种方式。运营者无需改动任何字词就能取消这种展示。响应缓存可以返回几乎无需成本便能生成的存储答案，而且可以立即显示，也可以按任意速度显示。平台已经开始试验如何呈现生成的输出（Jiangetal.， 2026年）。如果交付方式会影响盲投票，那么运营者出于成本原因作出的决策（例如提供存储的答案）也会改变用户对回答质量的评价，并由此影响用于模型排名的偏好数据。本文探讨同一回答的交付方式以及回答所需的时间如何改变用于对人工智能模型排名的投票，以及用户是否会继续提问。

在真实场景中厘清这些问题很困难，因为响应时间与内容息息相关。长答案需要更长时间生成，难题会引出长答案，而推理模型往往既缓慢又强大。等待时间与满意度之间的相关性，几乎无法说明等待会带来什么影响。实验室研究打破了这种关联，但它们会分配参与者必须完成的任务，因此参与者是否继续提问，既取决于体验，也取决于任务本身。

本文利用自然使用场景中的三种变异来源，将回答所需的时间与回答的内容区分开来。第一种来源是 compar:IA 中内置的一种抽签机制。compar:IA 是由法国文化部运营的公共聊天机器人竞技场，访问者可以输入一个提示词，让两个匿名模型分别作答，再将答案并排比较（Termignon 等人，2026年）。自 2026年3月起，如果某个模型此前已经回答过相同的初始提示，平台就会抽取一个随机数。根据抽取结果，平台要么生成一个新答案，供访问者观看其逐字写出；要么重放该模型先前针对相同提示生成并存储的答案，后者会在不到一秒的时间内填满屏幕。重放会逐字再现先前新生成的答案；许多存储文本在数据中既以新生成的回答出现，也以重放的形式出现，因此还可以针对相同文本进行比较。第二种来源是拥堵。在我们的研究时段内，平台服务器会等待各模型提供商的响应；这种机制使响应缓慢的提供商能够拖延其他请求，而且提供商的负载每天都有变化。因此，同一模型生成的长度相同的答案，会因访问者无法控制的原因而以不同速度到达。第三种来源是 WildChat，这是一个包含数百万条 ChatGPT 对话的语料库，其日志记录了每条回复由提供商生成的时间，以及完整接收该回复的时间（Zhao 等人，2024年）。在这个语料库中，提供商拥堵同样会改变内容相似的回复所需的等待时间。

由此得出三项发现。首先，同时展示的回答失去了盲投票。与新生成的回答相比，重放的回答赢得了22%的决定性投票，即那些选择一个回答而非判定平局的投票。在比较针对同一提示词的回答时，重放使赢得决定性投票的概率降低了25.8个百分点（标准误为4.4）；在所有符合条件的回答中，重放降低了赢得对话的占比，并提高了输掉对话的占比。即使将相同的存储文本与其自身进行比较，这种惩罚仍然存在。在265条文本的平均值中，相同文本新鲜展示时赢得了57%的决定性投票，重放时则赢得了29%；选民还认为重放的文本完整性较低，也更常被评为肤浅和不正确。其次，对于在屏幕上逐步呈现的回答，等待并未带来任何

可检测到的净收益。对于长度相同但耗时长于通常水平的回答，其获胜频率并未更高；估算结果排除了等待时间翻倍可带来超过约1.4个百分点收益的可能。第三，等待会产生行为成本。首次交互较慢，会使访问者提出第二个问题的概率每增加一个对数单位的延迟就降低1.8个百分点（约为基准率的8%）；等待时间每翻一倍，则降低1.3个百分点。大部分流失发生在等待时间最慢的五分之一中。在 WildChat 中，拥堵导致回复变慢，使用户更不愿继续；即使将比较范围限定为同一时钟小时内使用的模型，这一结论也成立。

这些结果与人们对盲投票的两种预期都相悖。交付方式影响了投票结果，而且方向与竞技场设计者预期的相反。对于相同的文本，与新鲜生成的回答流相比，快速重放使决定性胜出的比例大约减半，而选民解释自己选择时使用的是关于内容实质的说法。回答耗时并未对投票产生可检测的影响，反而体现在用户是否继续提问上。由于投票是在两个回答之间进行比较，如果延迟同时拖慢两者，就不会反映在投票中，尽管用户仍会因为等待而少问后续问题。

本文提出了两项理论贡献和一项实践贡献。第一，本文通过实地研究表明，在措辞保持不变的情况下，人工智能系统呈现输出的方式会影响人们对该输出作出的重要判断。关于劳动幻觉和呈现方式的研究已表明，这些因素会影响人们对服务的评价以及实验室中的评分（Buell 和 Norton，2011年；Zhou 等人，2025年；Jin 和 Gong，2026年）。在本文中，呈现方式大幅影响了用户在不同模型之间的盲选，而且影响方向与将速度解读为流畅性或能力的观点相反（Efendi´c 等人，2020年；Termignon 等人，2026年）。第二，本文界定了在回答生成过程中，时间还能带来什么影响。拥堵导致的减速并未给投票带来可检测的净加分，因此，实验室证据发现的「首个词延后出现所带来的加分」（Tan 等人，2026年）并不适用于这类情况，反而会导致用户少问后续问题。第三，研究结果对人工智能系统的构建和评估具有启示。缓存会改变回答的评判方式，延迟会影响用户是否继续提问，而基于盲投票构建的偏好排行榜会反映回答的交付方式，却忽略等待的成本。

本文其余部分安排如下。 第2节 阐述相互竞争的解释及其预测。 第3节介绍compar:IA和WildChat。 第4至第6节 至 6 介绍三项研究，而第7节 讨论对理论、设计和评估的启示。

2 理论与假设

人工智能服务如何交付回答，是一种设计选择。运营方可以在模型生成新鲜回答时逐步输出，也可以立即提供存储文本，或者让速度较慢的模型慢慢生成。本节将这一选择置于近期有关人工智能代理设计的研究背景下，阐述关于同一回答的交付方式以及回答所需时间如何影响人们对该回答的判断的各种解释，并推导出三项研究所检验的相互竞争的预测。

2.1 将人工智能回答的交付方式作为一种设计选择

信息系统和管理学研究表明，人工智能代理的设计线索会改变用户对其的评价以及用户的行为。零售商客户服务聊天机器人的类人线索改变了顾客是否接受其提议（Schanke 等，2021年）。人工智能代理表达积极情绪，对服务评价的提升不如人类员工表达同样情绪所带来的效果；部分原因在于，这违背了顾客对人工智能代理的预期（Han 等，2023年）。响应时间也是一种此类线索。在一项实验室实验中，聊天机器人的延迟回复对用户的社会临场感和使用意向产生了相反影响，而且影响因新手用户和有经验的用户而异（Gnewuch 等，2022年）。在一项客户服务聊天现场实验中，顾客刚看到聊天机器人未能理解自己，随后便将人工智能辅助的人类代理异常迅速的回复解读为自己仍在与聊天机器人交谈的迹象，其情绪也因此变差（Zhang 和 Narayandas，2026年）。在人际交往中，快速回复意味着积极响应，并提高受聘的可能性（Hart 等，2026年）。

生成式人工智能还带来一个独特的线索：输出的呈现方式。能够在几秒内生成内容的平台，已开始试验展示多少内容以及何时展示。在一项现场实验中，注册前展示部分生成内容，比展示全部内容更能提高参与度（Jiang 等人，2026年）。这些研究考察了输出的来源、智能体的社交线索或展示的输出量，而与我们最接近的实验则考察同一段文本在屏幕上的呈现方式（第 2.3 节）。但没有一项研究观察过：在真实运营政策下，以两种方式呈现同一模型逐字不变的回答，并将其与另一个回答进行比较，以作出模型排名的选择。

2.2 速度 与时间作为回答的信息 wer

一种常见的观点认为，人们会把回答所需的时间视为有关回答的信息。这种观点有两种相反的版本。第一种认为，速度表明流畅性和能力。compar:IA 的设计者预计，“更快的回答可能会被认为更流畅或更可靠，而较慢的回答可能会降低用户

对回答的偏好，而不受内容质量影响”（Termignon 等，2026）。人们认为算法生成预测的速度较慢，准确性就较低，也较少依赖这类预测。对算法而言，慢速程度和对人而言一样，意味着付出了努力；但由于人们认为预测对算法来说很容易，这种努力并不会被视为质量的一部分（Efendi´c 等，2020）。实践中也持相同看法：大语言模型应用的响应能力“被普遍认为至关重要”（Tan 等人，2026）。人机交互研究从一开始就将响应时间视为核心设计参数。Miller（1968）提出了对话交互的时间阈值，Shneiderman（1984）则回顾了相关证据，指出更长且变化更大的延迟会拖慢用户并使其感到沮丧。延迟会加剧用户的焦虑（Guynes，1988），而寻找一种可接受的响应时间，几十年来一直是研究课题（Dabrowski和Munson，2011）。信息系统研究将这一问题延伸到了网络领域。下载延迟是预测网站成功与否的性能指标之一，其中包括用户是否会再次访问（Palmer，2002）。在受控实验中，较长的页面延迟会降低用户的表现、态度和行为意向（Galletta 等人，2004），而用户等待的耐心很有限（Nah，2004）。搜索引擎和大型网站也报告了相同的规律：结果页面变慢会改变人们与其互动的方式（Arapakis等，2014），在线实验也一再表明，速度对参与度至关重要（Kohavi等，2014）。

在第二种情况下，时间传递出审慎思考和用心的信号。人们得知一首诗或一幅画耗时较长后，会给予它更高的评价(Kruger 等人，2004)；观察者会根据他人做决定时投入了多少思考来评判他们，并倾向于认为与任务难度相称的思考量最合适(Kupor等，2014)。对于大语言模型，Tan 等人（2026年）改变了流式回答中第一个词出现前的等待时间。等待 2 秒的参与者认为回答不如等待 9 秒或 20 秒的参与者所评价的那样深思熟虑，也认为回答不如等待 9 秒的参与者所评价的那样有用；等待 9 秒和 20 秒后的评分没有差异，记录到的交互行为也没有随延迟而变化。作者得出结论：延迟是“可调节的设计变量”。

2.3 渐进式显示

现代聊天机器人会在模型生成回答时，将每条回答流式传输到屏幕上，因此用户可以看着文本随着机器书写的节奏逐步展开。我们将此称为渐进式显示。在聊天机器人中，这种显示方式与服务领域研究过的显示方式有两点不同。屏幕上逐步展开的内容本身就是回答，而不是一旁的动画，因此用户在屏幕上看到的正是他们将要评判的内容。而且，这也是默认方式：用户看到的几乎每条回答都会逐字呈现，一次性显示完整回答的情况则是例外；通常由运营方决定何时采用这种方式，以避免再次生成同一回答的成本。流式传输的目的是让回答更快开始显示，而非展示生成过程，因此我们不会将其视为设计层面的操作透明度。我们的数据所观察到的差异，是新生成的回答在生成时以流式方式呈现，与存储的答案在不到一秒内重放之间的差异。

两条研究线索预测，逐步展开的回答会被评判为优于一次性呈现的同一回答。第一条线索以努力为路径。关于劳动幻觉的研究发现，人们会认可他们能够看见的工作（Buell 和 Norton，2011年）。在模拟旅行和约会网站上，等待期间观看工作过程展示的用户，比一次性获得相同结果的用户更看重该服务。价值来自展示而非工作本身：改变服务实际搜索的网站数量并不会改变用户对其的重视程度，而展示搜索过程则会。展示需要等待，因为结果一次性到达时便没有可展示的内容，因此核心发现是，人们“实际上可能更偏好等待时间较长的网站，而非即时返回结果的网站”（Buell 和 Norton，2011年）。然而，一旦存在展示，时间本身并未增加任何价值。当同样的工作在10至60秒的等待中被展示时，在每一等待时长下，有展示时的价值都更高，并且随着等待延长而趋于下降，二者之间没有交互作用。消费者会奖励那些明显为他们付出努力的企业（Morales，2005年），而操作透明度能够同时为顾客和员工创造价值（Buell 等人，2017年）。这一逻辑也延伸至机器：当用户认为推荐代理在其推荐中投入了更多努力时，会给予更高评价（Tsekouras 等人，2022年），而对投入某一产品的努力所持的信念，有助于解释为何消费者对人工智能制造的产品评价较低（Rix 等人，2025年）。顾客还会把回复速度解读为其产生方式的证据（Zhang 和 Narayandas，2026年）。沿着这条路径，一次性呈现的回答看起来像是投入的工作更少，而人们会部分根据他们认为该项工作所耗费的努力来评判工作质量（Kruger 等人，2004年）。

第二条路径涉及阅读与注意力。在等待期间持续运行的界面能够吸引注意力并带来专注沉浸（Lee 等人，2012年），而传达进度的提示能减少不确定性，动画提示则会把注意力从等待上引开（Lee 等人，2017年）。逐字呈现的生成式人工智能回答，比逐行呈现的回答更容易被接受，因为用户会感到更加沉浸；但当用户需要理解内容时，这种效果会减弱（Jin 和 Gong，2026年）。在一项使用人工智能写作工具的实验中，研究者将相同的文本以不同速度呈现，Zhou 等人（2025年）原本预计“对文本质量的评价不会有差异”。但他们发现了较小的影响。质量评分最高的速度略快于平均阅读速度，同时也是读起来最舒服的速度。沿着这条路径，以适合阅读的速度逐渐呈现的文本更容易跟上；而一次性出现的文本则会被略读，或显得像是千篇一律的回答。

表1：相互竞争的预测。

可观察指标（研究）	各理论解释的预测
一次性显示相同的回答，而非逐步显示（研究1）	纯文本：没有差异。速度代表流畅性或能力：一次性显示回答至少表现同样出色。渐进式显示通过付出努力或阅读：表现更差（H1a 对比 H1b）
逐步展开的回答需要更长时间（研究2）	将时间视为审慎思考：更常胜出（H2a）。将时间视为成本：胜出次数更少（H2b）。渐进式显示：没有变化
用户等待回答的时间更长（研究2和研究3）	等待的成本：后续问题更少（H3）

两条路径对显示方式的要求不同：第一条要求用户将其视为这份回答背后所付出工作的迹象，第二条只要求文本更容易跟上。两者都预测，一次性显示的回答会得到较差评价。我们的数据记录了投票、选民附加在投票上的限定语、反应及后续行为，因此检验的是渐进式显示是否获得认可，而不是究竟哪条路径带来了这种认可；第7.5节描述了一项能够区分两者的检验。两条路径还都预测，一旦回答开始逐渐呈现，额外的时间便不会带来任何好处：无论呈现得快些还是慢些，显示的都是同一段文本；而慢于阅读速度的呈现，也不会让文本更容易跟上。

2.4 相互竞争的预测

表1 概述了各解释所作出的预测。第一个可观察指标是，同一个回答一次性展示与逐步呈现时的表现如何。如果投票只依据文本作出判断，正如 Zhou 等人 (2025年) 对评分的预期，那么两种呈现方式的表现应当相当。如果速度传递了流畅性或能力的信号，正如compar:IA的设计者所预期的那样，一次性展示的回答表现至少应当不逊色。如果渐进式显示会得到认可，那么它的表现就应更差。

> **H1a（渐进式显示）**。一次性展示的人工智能回答，其评价劣于内容相当、在生成过程中逐步呈现在屏幕上的回答。**H1b（纯文本，或速度体现流畅性）**。一次性展示的人工智能回答，其评价不劣于内容相当、在生成过程中逐步呈现在屏幕上的回答。

选民可以附加到投票上的限定语，例如“完整”“肤浅”或“不正确”，体现了他们如何解释自己的选择。我们用这些限定语来描述判断，而不是区分不同路径。

第二个可观察指标是屏幕上逐步呈现的回答所用的时间。在这里，对时间的两种解读方向相反，而渐进式显示预测不会产生变化。

> **H2a（时间代表审慎思考）**。在屏幕上逐步呈现的回答中，用时较长的回答会被认为更好。
> **H2b（时间代表成本）**。在屏幕上逐步呈现的回答中，用时较长的回答会被认为更差。

介于两者之间的结果　将符合渐进式显示。如果审慎思考带来的收益与等待的成本相互抵消，也会出现这种结果。因此限定了时间对投票的净效应，而非识别其各个组成部分。

第三个可观察指标是行为。聊天机器人对话由一系列服务接触构成：每次回答后，用户都会决定是否开始下一次对话。排队研究发现，延误会降低顾客对服务的评价，主要是通过引发愤怒和不确定性（Taylor，1994年）；较长等待时间也会导致人们放弃服务，无论是急诊科的患者（Batt and Terwiesch，2015年）还是呼叫中心的来电者（Ak.sin 等人，2013年）。如果时间是一种成本，那么等待一个回答所付出的成本会降低用户开始下一次对话的意愿。

> **H3（时间成本）**。等待人工智能回答的时间越长，用户继续对话的可能性就越低。

H3 并不能区分 H1 的两种解释。它指出了时间成本出现的位置，而这对评估很重要：投票是在两个回答之间进行比较，因此，同时拖慢两个回答的延迟不会改变胜出者；但再次提问的决定会体现出这种延迟。

2.5 偏好作为人工智能系统的衡量标准

这些问题也关系到如何评估生成式人工智能。公共竞答平台会向访问者展示针对同一提示词的两个匿名回答，并将他们的投票汇总到排行榜中（Chiang 等人，2024年；Termignon 等人，2026年）。这类人类偏好会受到准确性以外的特征影响，其中最显著的是回答长度（Singhal 等人，2024年）；此外，如上所述，compar:IA 的设计者也认为延迟会影响判断（Termignon 等人，2026年）。这些影响是否构成问题，取决于排行榜旨在衡量什么。如果衡量的是模型所写文本的质量，那么文本的交付方式就是一种混杂因素。如果衡量的是用户体验到的服务，那么交付方式就应计入评分；但这样一来，模型响应缓慢所带来的成本也应体现在评分中，而等待后投下的一票，可能无法反映这段等待让用户失去的后续问题机会。

3 设置与数据

3.1 compar:IA

compar:IA 是法国政府内部开发的一项公共服务，旨在让公众比较对话式人工智能模型，并收集法语偏好数据（Termignon 等人，2026年）。无需注册账户，该平台已广泛应用于学校和大学。访问者可以输入提示词，或选择一个建议的提示词，并选择抽取两个模型的方式：随机抽取、从访问者选择的列表中抽取，或在大型模型与小型模型、推理模型等主题范围内抽取。随后，两个匿名模型会并排作答。两个模型都开始回答后，两份回答会同时显示。之后，随着模型生成文本，每份回答都会显示在屏幕上；在模型完成回答前，页面会显示加载指示器。对于推理模型，推理文本会显示在回答上方的面板中。[1] 随后，访问者可以提出后续问题，该问题会发送给两个模型；也可以通过点赞或点踩来评价任意一份回答，并添加可选评价标签（有用、完整、有创意、格式清晰；不正确、肤浅、未遵循指令）。如果没有对单条消息作出反应，访问者还可以为其中一个模型进行最终投票，或宣布两者平局，同样可以添加可选评价标签。只有在收到这些反馈后，模型的身份才会被揭示。

该平台依据开放 Etalab 2.0 许可发布了三个数据集（Ministèrede la Culture，2026年）：对话、最终投票和消息反馈。我们使用截至2026年5月18日的版本。该版本包含自2024年10月以来的410,016段对话、149,209条投票和92,841条反馈，并提供匿名访客标识符，以及每个已完成回答的文本、推理文本和元数据。从2025年2月起，元数据会记录每条回答的耗时，该耗时由平台服务器测量：从平台向模型提供商发送请求，到收到回答的最后一部分所经过的时间。它既包括等待首批文字出现的时间，也包括生成文字所花的时间，但不包括将文本传送到访客浏览器所需的时间。元数据还会标记从平台缓存中提供的回答，我们将在第4节中对此进行说明。

我们保留平台五种标准模式下的对话（376,662段），并剔除自动化流量：119个访客标识符在单日内发起了超过20段对话，共涉及14,893段对话，而且几乎从不投票。另有一个标识符对应没有已存储 Cookie 的访问者发起的50,628段对话；我们保留这些对话，在聚类时将每个会话视为一位独立访问者，并将其排除在所有访客内比较之外。在其余数据中，访问者在22%的对话中提出了第二个问题，在36%的对话中投出了最终投票，并在13%的对话中对至少一条消息作出了反馈。附录中的表16追踪了发布数据中每个分析样本的筛选过程。
平台在我们研究期间的运行方式有两个特征与设计有关，这两点都记录在其源代码中（compar:IA 团队、2026年）。首先，从2026年3月起，平台保留了首个回答缓存，B 节对此加以利用。其次，直到2026年10月，服务器一直以一种会阻塞其他请求的方式等待模型提供商。开发者之后的修复描述了这一问题：答案流在请求处理程序中运行，“因此每次等待提供商响应时，都会占用整个事件循环。模型回答缓慢时，其他访问者的请求也都会随之等待，健康检查同样如此。”因此，生成时间会随平台和各个提供商的负载而变化，而不仅取决于问题和回答，这正是第 5 节所利用的因素。

3.2 WildChat

WildChat 通过公共网络应用程序向用户免费提供 ChatGPT，前提是用户同意匿名收集其聊天记录（Zhao 等人，2024年）。我们使用其最大的公开发布版本 WildChat-4.8M（艾伦人工智能研究所，2025年），该版本涵盖了2023年4月至2025年7月期间与多代 OpenAI 模型（GPT-3.5、GPT-4、GPT-4o、GPT-4o-mini、GPT-4.1-mini 和 o1）进行的对话。公开发布版本包含未被标记为有害内容的对话。对于每条回复，日志都会记录

[1] 我们通过平台公共源代码及其提交历史（compar:IA 团队、2026年）了解了界面和服务器的行为。B 节列出了与我们设计相关的代码更改，并附有相关链接及其合并日期。

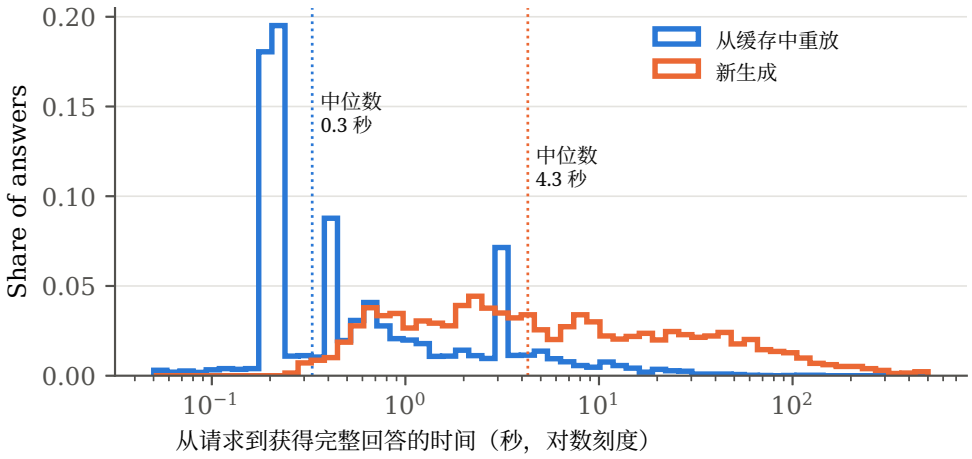


图1：2026年3月16日至2026年5月18日期间，符合条件的首个回答从请求到完整回答所需的时间，测量于平台的服务器。重放的回答来自平台存储的较早回答；新生成的答案则由模型生成并显示在屏幕上。虚线表示中位数。

提供商创建补全的时间（精确到秒）、WildChat服务器接收到完整回复的时间，以及一个哈希 IP 地址。该时间差从提供商为补全指定的创建时间起，计算至收到完整回复的时间，因此涵盖了该时间戳之后回复的生成和交付过程。WildChat 用户一次只与一个模型聊天，且未提供任何评分，因此这一数据源仅能反映持续对话。

4 研究1：一次性出现的回答

4.1 重放抽签

自2026年3月起，平台保留了一个缓存，即较早的首个回答的存储，以降低其成本（compar:IA 团队，2026年）。该存储以模型和初始提示的文本作为密钥，去除首尾空格并忽略大写字母；每个密钥最多保存最近的五个回答，并在该密钥最后一次使用48小时后过期。对于对话中的两个回答，平台分别生成一个随机数。在固定概率下，平台会在存储中查找该密钥；如果找到，就随机选取一个存储的答案进行重放；否则便请求模型生成一个新鲜答案，随后将其添加到存储中。数据中的首次重放发生在2026年3月16日，我们的时间窗口截至2026年5月18日。 B 节 记录了在此窗口运行的代码。

新生成的回答按照模型生成内容的速度显示在屏幕上。重放本应呈现相似效果：代码将存储文本拆分成约十五段，并“以短暂延迟逐段输出，以模拟真实的流式输出行为，确保一致的用户体验”（compar:IA 团队，2026年）。但实际延迟被缩短了。服务器同时读取一组比较中的两个回答，开发人员后来解释说，“只要另一个模型给出回答，较慢的读取就会立即取消”，因此每当另一个回答输出文本时，重放的停顿就会结束。服务器从重放开始到输出最后一段所测得的记录时长显示了这一结果（图 1）。重放时长的中位数为 0.3 秒，四分之三的重放不超过 1.1 秒。同一组提示词的新生成回答中位数耗时为 4.3 秒（四分位距为 1.4 至 20.3 秒）。在相同模型和提示词下，重放使记录时长缩短了 1.7 个对数单位，耗时减少五倍以上。因此，重放的回答只用几分之一秒就填满了屏幕上对应的一侧，远快于读者跟上文本显示的速度，而旁边的回答通常仍在生成。对于推理模型，存储的推理文本会一次性显示出来。我们将这种处理称为重放，并将其与新生成进行比较。

公开数据标明了哪些回答是重放的，却没有标明哪些回答本可以被重放。我们根据对话日志，按照平台自身的匹配规则重建资格：去除提示词首尾的空格，并忽略大小写后进行比较。如果同一模型在缓存期内已对相同的初始提示生成过新鲜回答，且该模型—提示词密钥在过去48小时内曾被使用过，那么该回答就符合资格。由此得到11,699个符合条件的回答，分布在2,460个模型—提示单元格中（1,077个不同的提示词、2,026名已识别访客），其中65%为重放。重放占比在4月5日前为42%，之后为74%，这与配置概率发生变化的情况一致，因此所有比较都固定周次。许多符合条件的提示词都是简短的问题，很多人都会

表2：研究1： 在模型和提示词固定的情况下，重放抽样与提问者和回答内容无关 d.

	Mean	重播减去新鲜	
		周固定效应	周 + 模型×提示词固定效应
预先确定的特征			
一天中的小时（0–23）	11.87	0.085 (0.112)	0.094 (0.132)
星期几（0 = 星期一）	2.72	0.038 (0.037)	−0.011 (0.036)
访客此前的对话（对数）	0.82	0.034（0.026）	0.028（0.017）
提示长度（字符数对数）	3.78	−0.052（0.027）	单元格中的常量
访客选择的模型（占比）	0.81	0.055 (0.008)	0.003 (0.003)
模型规模（参数对数）	5.21	−0.147 (0.031)	单元格常数项
科学与技术主题（占比）	0.14	−0.024 (0.007)	0.000 (0.003)
以右侧位置显示（占比）	0.50	−0.006 (0.010)	−0.007 (0.011)
答案内容			
回答长度（字符对数）	4.13		0.013（0.007）
输出词元（对数）	3.97		0.028（0.010）
推理文本（字符数对数）	2.04		0.047（0.017）

注：11,699个符合条件的首轮回答（65%被重放）。每个单元格显示重放指标的系数（标准误按访问者聚类）。提示词按照平台的匹配方式进行匹配，即去除首尾空格并忽略大小写。模型×提示单元格是平台在重放和新生成之间进行抽取的单位；单元格内的提示长度和模型规模不会变化。访问者历史行不包括用于没有Cookie的访问者的共享标识符。

访问者以完全相同的形式提问，例如询问法国首都。其他则是逐字重复的较长练习，例如总结给定文本。出现频率最高的提示词占符合条件的回答的19%，出现频率最高的十个提示词占62%。由于公开发布版本遗漏了一些对话，重建过程会漏掉部分符合条件的回答：在该时间窗口内，所有重播中有6%（8,604个中的509个）发生时，日志显示该密钥此前没有新生成的回答。若按精确文本而非平台规则匹配提示词，还会漏掉另外461个符合条件的回答。

4.2 抽取是否与随机抽取一样好?

有三项检验支持将模型—提示单元格内的抽取视为随机的。第一，重播确实如代码所示。在该时间窗口内的所有重播中，93%逐字重现了同一模型对同一提示词此前生成的新回答；其余部分最有可能重现了公开发布版本中未包含的对话里的回答。第二，重播答案与新生成的答案所测量的内容相似。在模型—提示单元格内，重播答案的字符数长1.3%（标准误0.7），词元数长2.8%（1.0），推理文本多4.7%（1.7）。第三，抽取与访问者和上下文无关（表2）。仅比较同一周内的重播答案和新生成的答案时，某些特征存在差异，因为测量的重放占比在不同提示词和模型之间有所变化；在模型—提示单元格内，每项差异都接近于零。因此，主要设定比较针对同一提示词的回答。

4.3 估计

主要结果是最终投票。对于对话 i，令 $W_i = 1$ 表示左侧回答是否在决定性投票中获胜， $R_{i\ell}$ 和 R_{ir} 表示左右两侧的回答是否为重播， $E_{i\ell}$ 和 E_{ir} 表示它们是否符合资格。我们估计

$$W_i = \beta\,(R_{i\ell} - R_{ir}) + \alpha_{(E_{i\ell}-E_{ir}),\,t(i)} + \gamma_{p(i)} + \delta_{q(i)} + \varepsilon_i, \tag{1}$$

其中， α 是第 t周符合条件模式的固定效应， γ 是无序模型对的固定效应， δ 是提示词的固定效应。左右位置随机分配，因此两侧对称； β 表示回答重放而非新生成时，其赢得决定性投票的概率变化。标准误按访问者聚类。样本包含对话中符合条件的回答至少有一个的1,005次决定性投票。重放可能改变访问者是否投票，以及投票是否为平局。因此，我们还报告了不以决定性投票为条件的结果：每段含有符合条件回答的对话如何结束、将平局计为半场胜利的得分，以及允许存在投票选择的界限。

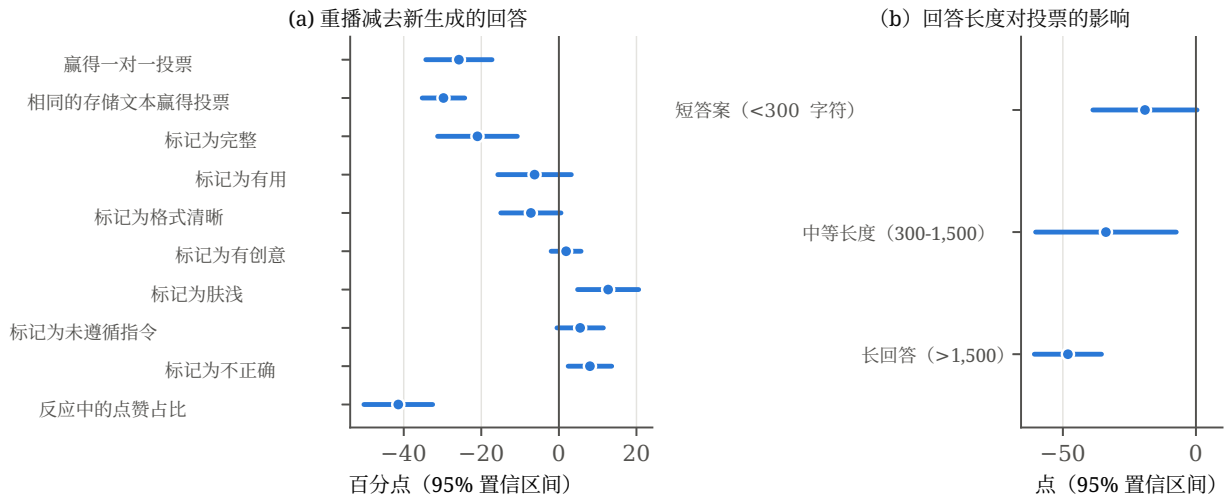


图2：研究1：与针对同一提示词新生成的回答相比，重播的回答会如何受到评价。面板（a）：投票、对相同存储文本的投票、投票附带的限定语（左侧减右侧差值），以及点赞占所有反应的比例。面板（b）：投票效应与两条回答的平均长度之间的关系。误差线表示按访问者聚类计算的标准误所对应的95%置信区间。

表3：研究1：重播的回答在盲投中失利。因变量：左侧回答在决定性投票中获胜（百分点）。

	(1)	(2)	(3)	(4)	(5)
重放（左）减去重放（右）	−26.2 (2.7)	−23.7 (3.2)	−25.8 (4.4)	−25.7 (4.3)	−30.3 (5.6)
按访问者和提示词聚类的标准误		(2.9)	(4.8)		
资格 × 周固定效应	Yes	Yes	Yes	Yes	
周固定效应					Yes
模型对 FE		Yes	Yes	Yes	Yes
提示 FE			Yes	Yes	Yes
模式 FE、长度和推理控制				Yes	
样本		至少一方符合资格			双方均符合资格
决定性投票	1,005	1,005	1,005	1,005	363
聚类（访问者）	862	862	862	862	300

注：第（3）列为主要设定。此处排除了将两个回答判定为相同的投票；表 4 报告了包含这些投票的结果。回归变量为 +1 （仅重放左侧回答时）， −1 （仅重放右侧回答时），其他情况下为 0。括号内为按访问者聚类的标准误。

4.4 结果

重播答案会落败。在474次涉及重播答案和新生成回答的决定性投票中，重播答案仅有22%的胜率。表3显示，这一惩罚在不同设定下都很稳定。在按周资格固定效应设定下，重放会使赢得决定性投票的概率降低26.2个百分点；在模型对固定效应设定下，降低23.7个百分点；在提示固定效应设定下，也就是我们的主要设定中，降低25.8个百分点。加入模式固定效应，并控制两个回答的长度和推理长度后，降幅为25.7个百分点。在两个回答都符合条件的对话中（此时两侧均为随机抽取），估计值为30.3个百分点（5.6）。在这些对话中，当两个回答都是新生成的时，左侧回答赢得50%的决定性投票；当两个回答都是重播答案时，左侧回答赢得49%；当一个回答被重放、另一个是新生成的回答时，重播答案赢得26%。除了按访问者聚类外，再按提示词聚类标准误以允许提示词重复出现，对结果影响甚微。结果支持H1a而非H1b：同时呈现的回答被评判得比在屏幕上逐步展开的回答更差。

相同文本。重放会再次呈现同一模型先前针对相同提示词给出的回答，而该回答此前曾以新出现的形式展示给另一位访问者。因此，数据中包含同一存储文本的重复呈现，其中一些是在屏幕上首次展示的，另一些则是重放的。1,948条存储文本以两种形式出现，其中265条在两种形式下都获得了决定性投票：310次新出现和334次重播呈现，由445位访问者针对118

个提示词作出判断。回答本身的文本完全相同，包括格式。在附带推理的文本中（占28%），88%的情况下，每次呈现的推理长度都相同。重放发生在该文本新出现后的中位时间为0.9小时。对每条文本赋予相同权重后，同一文本在新出现时有57%的决定性投票判其获胜，而重放时这一比例为29%，相差27.3点（3.7）。加入文本固定效应后，重放使其获胜概率降低29.8点（2.8；表4c）。按文本聚类计算标准误后，估计结果相近（3.5）；加入对手模型的固定效应并控制其回答长度后，结果也相近（30.2点）。将重放限制在新出现后的前48小时内（28.0点），以及仅考察44条其新出现和重播呈现均符合抽签资格的文本时（28.2点；按文本聚类的标准误为9.1），结果也相近。相同文本在重放时获得的点赞更少、点踩更多。这种比较固定了文本内容，而不是随机抽取交付方式，因此是对抽签结果的补充：它表明，文本之间的差异无法解释这种惩罚。

对话如何结束。表4a 跟踪发布记录中每个符合条件的回答，直至其对话结束，不以是否投票为条件。重放并未改变对话以最终投票（−0.3 点，0.6，未投票时为 62.5%）或平局（−0.5，0.4）结束的概率。回答获胜的对话占比下降了 1.5 个百分点，败选的占比上升了 2.3 个百分点，基准值分别为 11.2% 和 11.0%。在投过票的回答中，将平局计作半场胜利的得分下降了 15.9 个百分点（2.3）。Lee界限允许重放改变投票者身份，其假设是重放对所有访问者投票行为的影响方向相同（Lee，2009年）。据此，重放对决定性投票获胜份额的影响介于 −39.6 与 −28.9 个百分点之间（95%区间 −52.4 至 −20.3）；对包含平局的得分的影响介于 −28.6 与 −17.0 个百分点之间（−36.0 至 −8.9）。分别针对重放概率的两个时期计算，决定性投票获胜份额的界限分别为 −41.8 至 −22.9，以及 −37.8 至 −33.2 个百分点。一方出现重放，也使投票者中的平局概率降低了 9.8 个百分点（2.6），基准值为 36.1%：投票者看出了重放回答与新生成回答之间的差异，并因此判重放回答落败。

选民怎么说。 投票附带的限定语将重放的回答描述为实质内容更差（表5 和图2a）。与新生成的回答相比，重放的回答被认为不完整的可能性低了21.0点（5.2），被认为肤浅的可能性高了12.7点（4.0），被认为不正确的可能性高了8.0点（2.9）。在有用性、格式清晰和遵循指令方面的差异方向相同，但估计精度较低。即使文本相同，评价也一样：控制文本固定效应后，重放的回答被认为不完整的可能性低了23.2点（3.5），被认为肤浅的可能性高了12.2点（3.0），被认为不正确的可能性高了7.0点（2.1）（表5最后一列）。因此，当相同的措辞同时出现时，选民用实质内容方面的语言解释自己为何作出不利选择：不完整、肤浅、不正确。限定语是可选的，且由投票者随投票一并选择，因此它们展现的是选民如何为自己的选择辩护，而不是他们对某个具体主张的看法。即便如此，这些评价仍与将速度解读为流畅性或能力的观点相悖；按照这种观点，一个立刻给出回答的模型，至少应显得同样可靠。对单条回答的点赞和点踩也呈现出相同的情况。对于同一模型针对同一提示词给出的新生成回答，重放的回答获得的点赞少了62%，点踩多了86%；而在任何回应都获得的回答中，点赞所占比例低了41.4点（模型—提示单元格内低了32.2点）。

绝对惩罚与相对惩罚。 投票会比较两个回答，因此，重放可能部分是因为旁边的回答当时明显正在生成而落败。点赞和点踩反应有助于进行检验，因为它们对应单个回答，而且有些对话包含两次重放。参照组是两个回答都为新生成回答的对话中的回答；这些回答获得点赞的比例为 5.6%，获得点踩的比例为 1.5%。相对于该组，并在模型—提示单元格内，当另一个回答为新生成的回答时，重播的回答获得的点赞少 3.0 个百分点（0.5）；当另一个回答同样为重放时，少 1.4 个百分点（0.5）。与重放回答并列的新生成回答的点赞率变化了 −1.0 个百分点（0.6）。在这两种情况下，重放回答的点踩率分别上升 2.1 个百分点（0.5）和 0.9 个百分点（0.4）；与重放回答并列的新生成回答的点踩率变化了 +0.3 个百分点（0.3）。因此，即使屏幕上没有任何内容正在生成，重放仍会被评价得更差；而当旁边正在生成新生成的回答时，惩罚更大：在一个包含回答自身是否重放、相邻回答是否重放及二者交互项的模型中，交互项对点赞的影响为 +2.56 个百分点（0.62）。

不同回答长度下的惩罚。 短回答、中等长度回答和长回答均会出现惩罚（图 2范围表 6）：平均而言，少于 300 个字符的回答会被惩罚 19.1 点，中等长度回答会被惩罚 33.8 点，超过 1,500 个字符的回答会被惩罚 48.1 点。惩罚与符合条件的回答的预期长度（即同一模型此前针对同一提示词生成的新鲜回答的平均对数长度）之间存在线性交互，−3每对数单位为 1 点（2.8）。无论配对中是否包含推理模型，都会出现惩罚（分别为 19.4 点和 28.8 点）。

稳健性。 如果问题取决于日期（例如询问今天的日期），存储的回答可能已经过时。排除涉及时间或时事的提示词后，估计值为24.7点。去掉获得最多投票的提示词（占投票样本的6%）后，估计值为24.3点；仅考虑至少被三位访问者提出的提示词时，估计值也为24.3点。不使用共享的无 Cookie 标识符时，估计值为26.0点；当每次重放都

表4：研究1：重　　放惩罚，不以决定性投票为条件，且针对相同文本（百分点）　　).

	新鲜组均值（%）	重播减去新鲜
（*a*）所有符合条件的回答：对话如何结束（*11,699* 个回答）		
自己的回答在决定性投票中获胜	11.2	−1.5 (0.4)
其他回答在决定性投票中获胜	11.0	2.3 (0.4)
投票宣布双方平票	15.3	−0.5 (0.4)
没有最终投票	62.5	−0.3 (0.6)
净偏好（胜负差）	0.2	−3.8 (0.6)
（*b*）获得最终投票的回答，以及投票选择的入选投票界限		
平局计为半胜的评分（全部 2,180 个投票答案）	50.3	−15.9 (2.3)
Lee界限、决定性投票获胜份额		[−39.6, −28.9]
95%区间		[−52.4, −20.3]
Lee界限、所有投票中的得分		[−28.6, −17.0]
95%区间		[−36.0, −8.9]
Lee界限，决定性投票获胜份额，3月16日—4月5日		[−41.8, −22.9]
Lee界限，决定性投票获胜份额，4月6日—5月18日		[−37.8, −33.2]
(c) 同一存储文本以新鲜形式和重放形式呈现（两种形式下均有决定性投票的*265*条文本）		
赢得决定性投票，每条文本等权	56.8	−27.3 (3.7)
赢得决定性投票，文本固定效应	55.0	−29.8 (2.8)
标准误按文本聚类		(3.5)
加入周固定效应		−29.8 (2.8)
加入其他模型及其回答长度		−30.2 (2.9)
新出现后48小时内的重播		−28.0（3.0）
仅符合条件的出现（44 个文本）		−28.2（6.4）
按文本聚类		（9.1）
对回答点赞，文本固定效应		−1.33 (0.22)
对回答点踩，文本固定效应		1.10 (0.19)

注：面板(a)和(b)：答案层级回归，以该回答自身的重放指标和另一回答的重放指标为自变量，纳入模型×提示固定效应，并纳入与另一回答的资格交互的周固定效应；新鲜答案均值针对两条回答都未被重放的对话。Lee界限使用恰有一条符合条件回答的对话（2,097）；对重放使投票变得更常见（决定性投票）或更不常见（任何投票）的组进行修剪，并假设重放使每位访问者的投票都朝同一方向变化；区间为对访问者进行500次自助法抽样后，下界的第2.5百分位数和上界的第97.5百分位数。重放概率各时期的界限使用该时期的对话。在恰有一条符合条件回答的对话中，32.2%的重放和29.0%的新生成答案之后出现了决定性投票；在这些投票中，符合条件的回答分别赢得20.6%和51.7%。面板(c)：在出现决定性投票的对话中，同一模型针对同一提示词给出的存储文本（新生成和重放）的出现情况；回归纳入另一回答的重放指标以及文本固定效应。等权重行对有决定性投票的265条文本（两种形式均有）取平均，计算文本重放时与新生成时获胜份额之差；其标准误假设文本相互独立。点赞行使用所有出现情况，并纳入另一模型的固定效应。均值为文本以新生成形式展示时赢得决定性投票的比例。除非另有说明，标准误按访问者聚类。

该样本与较早的新生成回答逐字匹配。重放概率在两个时期都存在惩罚（24.1和27.1点）。采用更严格的资格定义后，结果大致不变：如果密钥必须在过去24小时内使用过，则为24.0点（4.6）。在较小的样本中，存储曾在过去48小时内为该密钥提供过重放，因此可以确定该密钥当时仍在使用，估计值为25.4点（10.1）。从存储中提供一个回答，也使其旁边的新生成回答的记录时间缩短了约14%。举例来说，将研究2中首选的投票估计值及其95%区间应用于这一幅度的变化，意味着重放获胜概率会改变 −1.1至+0.3点。

持续对话。 重放并未明显改变访问者是否会提出第二个问题：每次重放的首个回答都会使这一概率变化 +0.8 个点（0.6；95%区间 −0.3 至 +1.9），基准值为 8.8%（附录中的表11 ）。符合条件的提示词都很短，而且往往是分配给访问者的，因此很少有访问者会继续对话；接下来的拥堵研究旨在衡量持续对话。

5 研究2：耗时更长的回答

研究1比较同时出现的回答与在屏幕上逐步呈现的回答。研究2探讨当屏幕上逐步呈现的回答耗时比平常更长时会发生什么。它使用了2025年2月至2026年5月期间的首次交流，

表5：研究1：选民对重播的回答的评价（百分点）。

	均值（%）	重播减去新鲜	相同的存储文本
附加到投票的限定标签（*1,005* 次决定性投票）			
有用	18.0	−6.3 (4.8)	−9.1 (3.1)
完整	19.5	−21.0 (5.2)	−23.2 (3.5)
格式清晰	16.3	−7.2 (4.0)	
有创意	5.0	1.9 (2.0)	
肤浅	9.8	12.7 (4.0)	12.2 (3.0)
未遵循指令	6.5	5.5 (3.1)	
不正确	5.8	8.0 (2.9)	7.0 (2.1)
答案本身的拇指评价（*11,699* 个符合条件的回答）			
任何拇指反应	3.50	−0.21 (0.30)	
点赞	2.17	−1.36 (0.28)	
点踩	1.32	1.14 (0.23)	
已获得反应的回答中的点赞占比	62.1	−41.4 (4.5)	
包含模型×提示词 FE		−32.2 (7.0)	

注：标签行使用投票样本，以及表3中第（3）列的设定；因变量是标签的左侧减右侧差值。最后一列使用针对存储文本的1,336次决定性投票，这些文本既以新鲜形式出现，也以重放形式出现，并纳入文本固定效应；因变量是该文本标签与其他回答标签之间的差值。前三个点赞行是以重放指标为自变量的答案层级回归，纳入周固定效应和模型×提示固定效应，并包含每个符合条件的回答。对引发反应的409个回答中点赞所占比例进行估计时，先仅纳入周固定效应，再加入模型×提示固定效应。标准误按访问者聚类。

表 6：研究1：投票惩罚的稳健性与异质性（百分点）。

样本	重播减去新鲜	决定性投票
主要设定（第3列）	−25.8（4.4）	1,005
排除获得最多投票的提示词	−24.3 (5.0)	943
排除时效性提示词	−24.7 (5.8)	789
由三位或以上访问者提出的提示词	−24.3 (4.5)	797
排除共享的无 Cookie 标识符	−26.0 (4.8)	903
与先前回答逐字相同的重播	−27.9 (4.5)	983
其密钥在过去24小时内使用过的符合条件的回答	−24.0 (4.6)	905
其存储在过去48小时内提供过重播的符合条件的回答	−25.4 (10.1)	202
3月16日—4月5日（重放占比约为 0.4）	−24.1 (5.8)	556
4月6日—5月18日（重放占比约为 0.7）	−27.1 (8.1)	449
短答案（均值低于 300 个字符）	−19.1 (10.0)	188
中等长度答案（300–1,500）	−33.8 (13.5)	272
长答案（超过 1,500）	−48.1 (6.4)	545
这对模型中的一个推理模型	−19.4（11.5）	479
该组合中没有推理模型	−28.8（5.3）	526

注：表 3 第（3）列的设定。括号内为按访问者聚类的标准误。回答长度是两个首个回答的平均长度。将重放与符合条件的回答的预期长度（同一模型对同一提示词先前新生成的回答的平均对数长度）进行合并交互，得到 −3.1 个对数单位对应的点（2.8）；在不加入提示固定效应的情况下，该数值为 −2.5（2.2）。在资格条件行中较为严格的两行里，仅当对话中的每个符合条件的回答都满足更严格的规则时，才保留该对话。

两个回答均为新生成且计时，并经过 第3节中的筛选：共有324,029段对话，来自134,071名已识别访客，涉及110个模型。答案中位数耗时20秒（四分位距为9.4至39秒），而访客中位数等待较慢回答的时间为26秒。

5.1 设计

生成时间取决于模型、回答长度以及模型在回答前进行多少推理。我们将慢速程度定义为回答生成时间对数中无法由这些因素解释的部分：即对每个模型在每个月分别进行回归时，将持续时间对数回归于回答长度对数和推理长度对数所得的残差。慢速程度为 0 的回答，其耗时与该模型生成相同长度文本时的典型耗时相当；慢速程度为 0.7 则意味着耗时约为两倍。慢速程度的标准差为 0.65 对数单位。

慢速程度仍可能反映内容，例如某些类型的文本生成速度较慢。因此，我们使用拥塞度量作为工具变量：同一天该模型向其他访问者提供的回答的平均慢速程度，但排除目标访问者自己的回答。由于所有访问者共享平台服务器和各模型提供商，提供商繁忙的一天会使该模型生成的每个回答都变慢（第3节）。回归吸收了每个日历日的固定效应，因此用于识别的变异，是某模型的拥塞程度与同一天其他模型拥塞程度之间的差异；在控制固定效应后，其标准差为 0.10 对数单位。拥塞度量已根据回答长度进行调整，因为原始生成时间还包含当天请求构成的影响：基于原始时间构建的拥塞度量能够预测目标模型生成更长的回答（+0每个对数单位对应 0.034 个对数单位，标准误为 0.017），而经过长度调整的度量则不能预测（表12）。我们报告普通最小二乘法（OLS）估计值，该方法利用慢速程度的全部变异；同时报告工具变量（IV）估计值，该方法仅利用由拥塞程度驱动的变异。IV 估计值依赖于两个假设。第一，在给定固定效应的情况下，某模型在某一天的拥塞程度与访问者是谁以及他们提出了什么问题无关。第二，拥塞程度仅通过回答所需的时间影响访问者的行为。

拥塞度量能够有力地预测回答自身的慢速程度（系数为0.94，第一阶段 F 统计量为1,066），并且与其不应影响的对话特征无关（附录中的表12）。在拥堵日，提示词并没有变长，访问者的经验也没有增加，主题和模式也没有发生变化。回答并没有变短，但其中包含的推理文本更少（表12）。由于拥堵可能以这种方式改变回答本身，我们还报告了回归中不包含长度和推理控制变量的估计结果，以及将生成时间对数本身作为回归变量的估计结果（表13）。

对于持续对话，分析单位是对话，回归变量是两个回答慢速程度的均值。所有回归均吸收每月模型对的固定效应、日历日固定效应、星期时段固定效应和对话主题固定效应，并控制回答长度、推理长度、提示长度以及访问者在平台上的经验。访客内估计为每位访问者加入固定效应，因此比较同一个人的较快对话和较慢对话。对于投票，回归变量是左侧与右侧回答之间慢速程度的差异，工具变量是两个模型拥塞程度之间的差异。主要投票设定吸收每月有序模型对的固定效应，从而控制哪个模型位于左侧，以及日历日固定效应。由于许多访问者共享同一天的拥塞程度，标准误采用双向聚类，分别按访问者和日历日聚类（Cameron等人，2011年）。

5.2 等待没有带来可检测的优势

表7报告了投票回归结果。在屏幕上逐步展开的两个回答中，就其长度而言耗时更长的回答并未更频繁地获胜。在主要设定中，迟缓程度每增加一个对数单位，获胜概率就变化 +0.2个百分点（0.4）；当迟缓程度由服务提供商拥塞导致时，变化为 −2.9个百分点（2.5）。这两个估计值均无法与零区分；对于等待时间翻倍的情况，工具变量估计值的95%区间为 −5.4至 +1.4个百分点，排除了大幅提升的可能。仅使用无序模型对的设定给出了负的估计值（普通最小二乘法 −1.0、0.4；工具变量 −6.6、3.2）。差异来自模型对的排序，该排序解释了哪个模型位于左侧，而不是来自日历日效应：仅使用有序模型对时，估计值为 +0.2和 −2.7个百分点。作为比较，文本长度多一个对数单位的回答，其获胜频率高出13.9个百分点，这是偏好数据中熟悉的长度效应（Singhal 等人，2024年），而研究1中的一次重放则低了25.8个百分点。

对单条回答的反应也呈现出相同趋势。在控制其旁边那条回答的慢速程度后，较慢的回答获得的点赞更少（每增加一个对数单位减少−026点，标准误为0.07），点踩更多（增加+026点，标准误为0.05）；两者的基准比例分别为6.0%和2.8%。而慢回答旁边的那条回答获得了更多点赞（增加+037点，标准误为0.07）：访问者等待较慢的回答时，已经可以对先到的回答做出反应。

因此，投票结果并未为H2a或H2b提供可检测到的支持：对于一条已经开始生成的回答，等待更久并没有让它更常胜出。这正是渐进式显示所预测的模式，而上述区间限定了对时间的两种解读所产生的净效应（表1）。就单条回答而言，较慢的回答获得的点踩略多，这与等待会带来成本的推断一致。实验室证据考察的是等待过程中的另一个环节。 Tan 等人(2026年)延迟了第一个词的出现，随后以固定速度逐步输出回答；与等待2秒相比，等待9秒提升了感知到的思考深度，而等待20秒并未比等待9秒带来更多提升。我们的测量涵盖了回答生成所需的全部时间，直到最后一个词出现；这类回答的生成耗时长达数十秒。因此，这两组估计描述的是不同的时间变化，并不处于同一条估计曲线上。

5.3 等待会产生持续对话成本

等待确实会影响访问者接下来的行为。表8 显示，较慢的首次交流使访问者提出第二个问题的概率每增加一个迟缓的对数单位便降低1.8点（0.2）；基准概率为23.1%，因此降幅约为

表7：研　究2：等待在投票中未带来可检测的认可（每对数单位慢速程度的百分点　　s).

	OLS	IV	第一阶段 F 值	观测值
盲投：左侧回答获胜（慢速程度，左减右）				
模型对 × 月份固定效应、顺序效应和日历日固定效应	0.2 (0.4)	−2.9 (2.5)	469	80,205
按访问者和周聚类的标准误		−2.9 (2.8)	232	80,205
不含日历日固定效应	0.2 (0.4)	−2.7 (2.5)	466	80,205
模型对 × 月份固定效应，未排序	−1.0 (0.4)	−6.6（3.2）	159	80,205
添加日历日固定效应	−1.0（0.4）	−6.5（3.2）	159	80,205
没有长度和推理控制		−7.4 (3.5)	152	80,205
单轮对话	−0.7 (0.5)	−4.9 (3.3)	155	66,321
每个流式传输的首个回答获得的点赞数				
回答获得的点赞数	−0.04 (0.06)	0.47 (0.36)	1,609	659,969
对回答点踩	0.20 (0.04)	0.61 (0.21)	1,609	659,969
获得反馈的回答中，点赞占比	−1.6（0.4）			57,299

注：除非另有说明，投票回归均控制左侧与右侧在对数长度和推理长度对数上的差值。有序设定为左右两侧每一种双模型排列分别设置固定效应；无序设定中，额外增加一个对数单位的长度可赢得 +13.9点（0.4）。工具变量是其他访问者所遇到的两个模型当日缓慢程度之左侧减右侧差值。答案层级回归吸收模型 × 月份和日历日固定效应，并控制长度和推理长度；回归纳入所有有记录时间的流式传输首个回答，包括其搭档未记录时间的回答。标准误按访问者和日历日两个维度聚类。

基准率。基于拥堵的估计值相近（2.0 点，1.0）。比较同一访问者的对话，估计值分别为 1.5（0.2）和 2.4 点（1.1）。因此，等待时间加倍会使提出后续问题的概率降低约 1.3 点。附录中的表13显示，在回归中不加入长度和推理控制变量，并以生成时间对数而非慢速程度作为回归变量时，点估计值相同（−2.0，1.0）。使用仅剔除焦点对话的工具变量、以较慢回答的等待时间（对数秒）作为回归变量，以及在不含推理模型的对话中，估计值也相同。将聚类方式从访问者和日期改为访问者和周，以允许拥堵冲击持续数日，会使工具变量估计值的标准误升至 1.1（普通最小二乘法为 0.2）。简化式，即拥塞度量本身的影响，为 −1每对数单位 0.8 点（0.9）。在工具变量实际提供的变异范围内，拥堵程度每增加一个标准差，回答所需时间就会延长约 10%，提出后续问题的概率则会变化 −00.19 点。

图3a 展示了持续对话如何随慢速程度的变化而变化。在调整了相同的固定效应和控制变量后（普通最小二乘法），在对话中位等待时间为10至40秒的最快八成对话中，曾提出第二个问题的访问者占比与最快十分位相比，差异不到半个点。第九个十分位的这一比例低0.9个点（0.4；中位等待时间为55秒），而最慢十分位低3.6个点（0.5；中位等待时间为111秒）。在慢速程度的第80百分位数以下和以上分别采用分段斜率的回归显示，在以下区间为 +0每个对数单位0.2个点（0.3），在 −5以上为0.1个点（0.4）。在这些调整后比较中，中等程度的慢速回答伴随的损失很小，而损失集中在最慢的五分之一对话的长等待时间上。

这种成本会在对话后续再次出现。在已经继续对话的访问者中，第2至第6轮次较慢的回答使再次提出问题的概率降低了每个对数单位3.7个点（0.3），在访问者内部则降低4.2个点（0.4），基准值为51.2%。拥堵估计值为 −1.6个点（1.5）。

其他结果则呈现出不同的变化。回答速度变慢并未改变最终投票的可能性（工具变量 0.0，0.8）；而采用拥堵作为工具后，估计结果显示，一周内另一天返回平台的可能性很小，且无法与零区分（−0.6，0.5）。回答速度变慢提高了访问者对回答作出反应的可能性（工具变量 +1.7，0.7），这与上文所述用户对等待中的聊天对象的反应一致。这些结果支持 H3：等待的成本体现在用户是否决定再次提问上。

6 研究3：日常聊天机器人使用中的等待

compar:IA 是一个比较网站，其访问者知道自己正在测试模型。研究3探讨等待的成本是否也会出现在用户日常使用单个聊天机器人的过程中。WildChat 用户向 ChatGPT 寻求各种帮助，每次只使用一个模型，且没有评分，也不会揭示结果（Zhao 等人，2024年）。

表8：研究2：慢速程度 较慢的首个回答会减少后续问题（每迟缓的对数单位对应的百分点）

	OLS	IV	OLS	IV
	所有对话		访问者内部	
提出第二个问题	−1.8 (0.2)	−2.0 (1.0)	−1.5 (0.2)	−2.4 (1.1)
7天内其他日期的回访	−0.1 (0.1)	−0.6 (0.5)		
投出最终票	1.3 (0.3)	0.0 (0.8)		
回应消息	0.3 (0.1)	1.7 (0.7)		
“提出第二个问题”的平均值（%）		23.1		
第一阶段系数（F统计量）		0.94（1,066）		(973)
对话	324,029	324,029	280,993	280,993
后续轮次：在第 *2–6* 轮后继续（均值 *51.2%*）				
继续对话	−3.7 (0.3)	−1.6 (1.5)	−4.2 (0.4)	
轮次	140,671	140,368	121,794	

注：慢速程度是两条回答的生成时间对数均值，在模型和月份范围内扣除其长度与推理长度的影响后计算得出。工具变量是同一天内相同模型回答其他访问者时的平均慢速程度（服务提供商拥塞）。所有列均吸收模型对 × 月份、日历日、星期小时和主题固定效应，并控制回答长度、推理长度、提示长度和访问者经验；同一访问者列还加入访问者固定效应。七日回访使用发布数据中有七天后续跟踪记录的319,912段对话。标准误按两种方式聚类：按访问者聚类（无 Cookie 的访问者按会话聚类），并按日历日聚类。

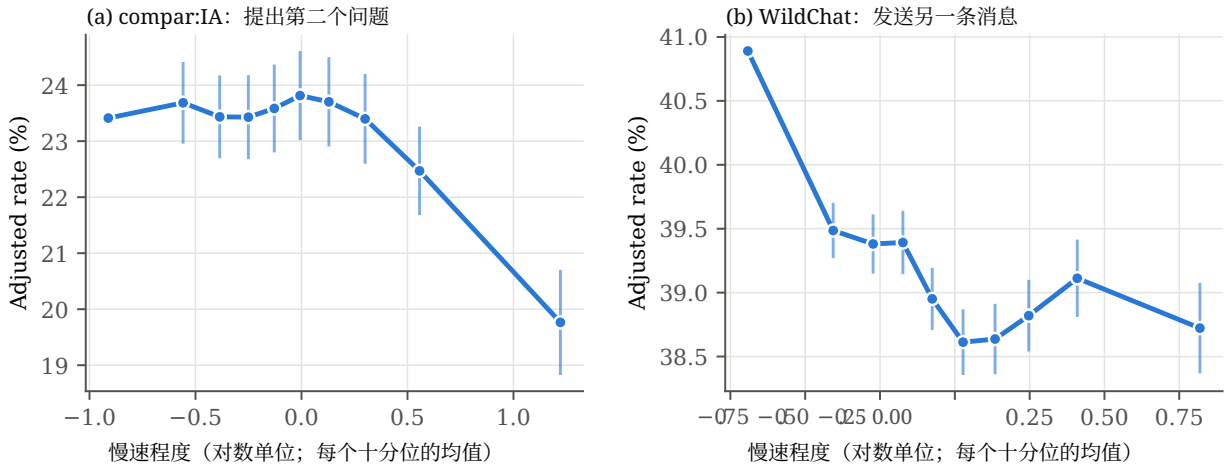


图3：按慢速程度十分位划分的持续对话。面板（a）：compar:IA 的首次交流；访问者提出第二个问题的比率，已根据表8中的固定效应和控制变量进行调整。面板（b）：WildChat 回复；用户发送另一条消息的比率，已根据表 9中的固定效应进行调整。柱形表示各十分位与最快十分位之间差异的95%区间，标准误的聚类方式与主表相同。面板（a）中的慢速程度是相对于同一模型相似回答的对数时间，因此0.7表示耗时约为通常的两倍；面板（b）中的慢速程度则按秒为单位的延迟加一后取对数计算。

6.1 设计

对于每条回复，我们将延迟计算为 WildChat服务器收到完整回复的时间减去提供商生成补全的时间，两者均精确记录到秒（第3节）。发布数据中的 5,167,182 条回复没有一条延迟为负，其中 461,750 条的延迟不足一秒，记录为零，只有 26 条超过十分钟。延迟与回复长度密切相关：在同一模型和月份内，回复长度每增加一个对数单位，延迟就增加 0.61 个对数单位；如果延迟涵盖完整回复的撰写时间，这一结果正如预期。我们保留延迟不超过十分钟的回复，并剔除单日创建超过 50 段对话的哈希 IP 地址。样本包含 2023年4月至2025年7月期间来自 1,829,326 个哈希 IP 地址的 2,694,924 段对话中的 4,378,514 条回复。回复耗时的中位数为 6 秒（四分位距为 2 至 13 秒；第 95 百分位数为 43 秒）。结果变量是用户是否在收到回复后于同一段对话中发送了另一条消息。用户在收到首条回复后有 17.6% 的情况下发送了后续消息，在收到后续回复后则有 73.5% 的情况下发送了后续消息。

研究设计沿用研究2，但有一处调整。简短回复存在固定开销，且许多延迟记录为零，因此我们以延迟秒数加一后的对数衡量慢速程度，并以非参数方式调整内容。慢速程度是该对数值减去以下回复的平均值：来自同一模型、同一月份、回复长度相近、消息长度相近且在对话中处于相同位置的回复。拥堵程度则是同一模型在同一时钟小时内向其他用户发送的回复的平均慢速程度。OpenAI 的服务由所有 WildChat 用户以及该提供商的其他客户共享，因此每小时的拥堵程度会因任何单个用户无法控制的因素而变化。

拥堵也反映了需求：控制日历日和星期小时固定效应后，拥堵时段的消息和回复较短，美国用户也较少（表12）。因此，首选设定吸收了 19,077 个时钟小时各自的固定效应。这些效应吸收了同一小时内所有模型共同面临的需求，因此比较的是同一小时内使用的模型；特定于某个模型的需求则仍然存在。在多个模型同时被使用的时段发送的回复占 59%，为识别提供了变异。在控制这些固定效应后，拥塞度量的标准差为 0.06 对数单位，仍能有力预测回复本身的慢速程度（系数为 0.72，第一阶段 F统计量为 10,068）。消息和回复长度方面的差异消失了，美国用户占比的系数也从 −0.051 降至 −0.023（0.010）。我们还报告了对来自同一哈希 IP 地址的回复进行比较的估计结果，这消除了用户之间稳定的差异；此外，也报告了以日历日和星期小时固定效应替代时钟小时固定效应的估计结果。所有回归均控制模型与月份、对话中的位置以及长度分组单元的交互固定效应；标准误采用双向聚类，分别按哈希 IP 地址和时钟小时聚类。

6.2 结果

表 9 显示，回复越慢，用户继续对话的可能性越低。控制时钟小时效应后，慢速程度每增加一个单位，用户发送另一条消息的概率就降低 1.3 个点（0.1）；当慢速程度由拥堵导致时，概率降低 1.8 个点（0.7）。基准概率为 39.2%。在同一设计中比较来自同一哈希 IP 地址的回复，降幅分别为 2.4 个点（0.1）和 6.7 个点（1.2）。改为控制日历日和星期小时效应后，估计值分别为 1.5 个点和 3.5 个点（在同一哈希 IP 地址内分别为 2.6 个点和 6.1 个点）。在时钟小时设计中，拥堵估计值主要来自后续回复（−30.6、0.9）和 GPT-3.5 模型（−10.9、0.2）。对于首次回复、GPT-4 级模型和美国用户，时钟小时工具变量估计值分别为 −10.2（1.1）、+10.3（1.4）和 +10.4 个点（1.8）；首次回复和 GPT-4 级模型的普通最小二乘法估计值为负。排除记录为零秒的回复后，估计值相近（−20.1、0.7）；这类回复占样本的 8%，且全部处于慢速程度最快的四个十分位内。在确实继续对话的用户中，回复越慢，从收到回复到创建下一条请求的时间就越长（+17慢速程度每增加一个单位，时间增加 0.0 个对数点，0.7）。图3b 展示了按慢速程度十分位调整后的持续对话率。与 compar:IA 不同，下降主要发生在较早阶段：收到第二快十分位的回复后，用户继续对话的可能性比收到最快十分位回复后低 1.4 个点（0.1）；收到最慢十分位的回复后，这一可能性则低 2.2 个点（0.2）。

研究3表明，H3 在竞技场之外同样成立。在日常使用单个聊天机器人时，无论是在比较网站上，等待回答的时间越长，人们就越不愿继续对话。

7 讨论

这三项研究将通常相伴出现的两件事区分开来：人工智能回答的交付方式，以及用户等待回答的时间。交付方式影响了投票。对每段文本赋予同等权重时，同一模型生成的同一段存储文本，在屏幕上逐步展开时赢得了57%的决定性投票，而一次性填满整个屏幕时只赢得29%。选民在对照重放文本解释自己的选择时，称其不够完整、更加肤浅，而且更常出现错误。等待则影响了行为。在屏幕上逐步展开的回答中，额外的等待时间并未在投票中带来可检测到的净支持；等待时间较长也降低了人们继续提问的可能性，这种情况在公共竞技场和日常使用 ChatGPT 时都存在。

7.1 哪种解释符合证据

表10 将各种解释与相关结果逐一对照。只评判文本的投票，以及偏好更快回答的投票，都未能通过第一项检验：一次性显示的回答以明显劣势落败，选民也没有认为它更可靠。渐进式显示则通过了这项检验。对于逐步展开的回答所需的时间，无论采用哪种时间解读，都没有发现投票中存在可检测到的支持，这正是渐进式显示理论所预测的。时间成本体现于用户是否会再次提问。

该设置的若干特征可能会在没有这些解释的情况下，使同时展示的回答受到惩罚，而数据排除了这些可能性。当对话中的两个回答都是重放时，单个回答仍会受到惩罚，因此这种惩罚并不要求屏幕另一侧存在一个明显正在生成的回答，尽管与此类回答形成对比会使惩罚更大。在逐步展开的回答中，较晚完成的回答并没有更经常获胜，因此这种惩罚不是近因效应。相同文本也会引发这种惩罚；排除依赖日期的提示词后，这一现象仍然存在，因此存储

表 9：研究3：在日　常使用 ChatGPT 时，较慢的回复会结束对话（每单位慢速程度对应的百分点　　）.

	（1）普通最小二乘法	（2）工具变量	（3）普通最小二乘法	（4）工具变量
			在哈希 IP 地址内	
精确到小时的固定效应（首选）				
用户发送另一条消息	−1.3 (0.1)	−1.8 (0.7)	−2.4 (0.1)	−6.7 (1.2)
第一阶段 F 值		10,068		3,950
日历日和星期小时固定效应				
用户发送另一条消息	−1.5 (0.1)	−3.5 (0.4)	−2.6 (0.1)	−6.1 (0.7)
第一阶段 F 值		58,073		12,754
均值（%）		39.2		
回复		4,378,514		

子样本，精确到小时的固定效应	OLS	IV	回复
对话中的首条回复	−0.4 (0.1)	−1.2 (1.1)	2,687,209
后续回复	−2.2 (0.1)	−3.6 (0.9)	1,691,305
GPT-3.5 模型	−3.0 (0.2)	−1.9 (0.2)	1,232,870
GPT-4 级别和 o1 模型	−0.3 (0.1)	1.3 (1.4)	3,145,644
美国用户		1.4 (1.8)	910,395
用时不超过五分钟的回复		−1.8 (0.7)	4,378,055
排除记录为零秒的延迟	−1.8 (0.1)	−2.1 (0.7)	4,036,299
标准误仅按哈希 IP 聚类		−1.8 (0.6)	4,378,514
到下一次请求的时间（对数秒，连续回复）	0.170（0.007）		1,650,194

注：慢速程度是 log（1 + 以秒计的延迟）减去同一模型在同一个月内、具有相似回复长度（二十分位组）、用户消息长度（十分位组）和对话中位置的回复的均值。工具变量是同一模型在同一时钟小时内回复其他用户时的平均慢速程度。所有回归均吸收模型、 × 月份、轮次位置和长度单元固定效应。首选行加入每个时钟小时一个固定效应（19,077 个小时），因此只保留同一小时内所使用模型之间的差异；59% 的回复是在包含多个模型的小时内发送的。第二组行则使用日历日和星期小时固定效应。标准误按两种方式聚类：哈希 IP 地址和时钟小时。下一次请求所需时间这一行的单位是秒的对数单位，而不是百分点，衡量的是从收到回复到为用户下一条消息创建补全之间的时间。样本中有 7,750 个对话没有第一条回复，因为其第一条回复落在回复数少于 20 条的长度单元中、缺少拥塞度量，或超过了延迟上限。

内容并不能解释这一点。重播并未改变访问者是否投票，也没有改变他们在所有对话中宣布平局的频率。在投票者中，单次重播降低了出现平局的可能性，这使比较更加清晰，而不是决定谁会投票。

7.2 对理论的启示

劳动幻觉认为，当人们能看到一项服务背后的工作时，他们会更加看重这项服务；相关实验表明，起作用的是展示本身，而非实际工作量（Buell 和 Norton，2011年）。后续研究发现，人工智能文本在屏幕上的呈现方式会小幅改变实验室对其质量的评分（Zhou 等人，2025年），也会改变用户接受它的意愿（Jin 和 Gong，2026年）。我们的结果将这些发现拓展到了人工智能输出判断会产生实际后果的情境：人们在盲测中对模型进行排序；这些人自行选择问题，而比较的是同一模型在运营方实际策略下交付的、逐字不变的回答。在这种情境下，交付方式使结果出现了明显差异，而且方向与平台设计者的预期相反（Termignon 等人，2026年）。关于消费者对人工智能制造的产品打折看待的研究，部分将这种现象归因于人们相信算法创造的产品缺少投入其中的爱与心力（Rix 等人，2025年）。在这里，两个回答都由人工智能生成，措辞也完全固定，因此仅交付方式就改变了结果。在注册前展示部分生成内容会提高参与度（Jiang 等人，2026年）；我们的结果进一步表明，完整回答的揭示方式会改变人们对它的评价。机器运行越快就越被认为可靠的预测（Efendi´c 等人，2020年）并不适用于用户能看到文本逐步呈现的情形：一次性显示的回答，反而更常被投票者描述为不正确。对于默认以逐步呈现方式生成回答的生成式人工智能，劳动幻觉引出的实际问题恰好相反：不是要不要增加工作过程的展示，而是缓存等后台决策会在何时将这种展示取消。究竟是哪条路径产生了这种效应，是实地数据尚未解答的理论问题。答案将决定这种效应反映的是对努力程度的推断——告知用户节奏由软件控制应会削弱这种推断——还是更容易阅读，而告知用户这一点不应产生影响。

表10：与证据相悖的解释。

解释	预测	研究发现
投票者只评判文本	同一个回答在展示于……时表现同样出色	在决定性投票中，支持该方案的比例为57%，而支持另一方案的比例为29%
速度体现的是流畅性还是能力	一次性呈现或逐步展开 一次性呈现的回答至少表现得同样好而且似乎至少同样可靠	相同文本；抽签中 −25.8点 它表现更差，选民认为它不正确的次数也更多（+8.0 点）
渐进式显示（付出努力或阅读）	一次性显示的回答表现更差；额外的花在展开式回答上的时间毫无益处	与两者均一致
作为审慎思考的时间	耗时更长的展开式回答胜出	没有可检测的净增益：等待时间翻倍时，等待， −5.4 至 +1.4 点（95%区间）
时间作为一种成本	more 耗时更长的展开式回答胜出更少；用户较少再次提问	投票中未发现净损失；后续问题更少 compar:IA 上的后续问题，以及在 WildChat
仅限屏幕对比	两种回答同时出现时不受惩罚	点赞 −1.4 点，可与另一个重放并列
最后完成的回答 wins	展开得越慢的回答，获胜越多	另一个重放 它们不会（每对数单位对应 +0 个普通最小二乘法估计值、 −2 个工具变量估计值）
存储的内容不同	相同文本不受惩罚，或仅对取决于日期的提示词进行惩罚	对相同文本进行惩罚； −24.7 点 不含时效性提示词
重播改变了谁投票	投票或平局在所有对话中发生变化	两者均未改变；Lee界限排除零

这些结果也限定了人们将时间视为机器输出信息的解读范围。在屏幕上逐步呈现的回答中，拥堵造成的延迟并未带来可检测到的净投票优势。关于首个词出现得更晚会提高感知到的思考深度的实验室证据（Tan 等人，2026年）与这一结果并不矛盾，而是与之并存。各项研究设计改变的是等待过程的不同部分：实验室研究延迟了首个词的出现，同时保持书写速度不变；我们的数据中的重播则完全取消了逐步呈现；而拥堵延长了从开始到最后一个词出现的整个时间。首个词出现前的延迟如何与文本逐步呈现相结合，仍有待研究；若能分别改变这两个因素，便可揭示其作用。劳动幻觉实验本身也指向与我们的估计相同的结论：在展示相同工作的情况下，较长等待时间并未增加价值（Buell 和 Norton，2011年）。

对于信息系统中响应时间的研究，这些发现将等待成本从态度转移到行为边际。早期研究测量了满意度、态度和意图（Shneiderman，1984年；Palmer，2002年；Galletta 等人，2004年；Nah，2004年），而关于等待界面的研究则探讨显示如何改变感知到的等待（Lee 等人，2012年，2017年）。对话系统提供了一种直接的行为测量，即用户在回答到达后是否再次提问；在我们的数据中，这一测量随等待而变化，而偏好投票则没有。该决策类似于在高成本等待之后选择开始另一次服务接触，而非放弃正在进行的排队（Batt与Terwiesch，2015年；Ak.sin 等人，2013年）。在compar:IA上的调整后比较中，与排队情形一样，大部分损失来自最长的等待；在WildChat中，回复快得多，损失主要来自脱离最快的十分位。

7.3 对设计的启示

缓存对所有生成式人工智能服务运营者都很有吸引力，因为它能节省计算资源和时间（Bang，2023）。我们的结果表明，存储的答案如何呈现，会影响人们对它的评价。一次性铺满屏幕的答案，其评价不如逐步呈现在屏幕上的同一答案，因此，能够节省成本的缓存也可能降低感知质量。compar:IA 自身的设计原本打算分段重放存储的答案，“以模拟真实的流式输出行为”（compar:IA 团队，2026年）；但我们在观察窗口中看到的重播速度远快于预期。研究结果支持以可读的速度测试重放，也引出了关于推测解码及其他会改变文本呈现速度的方法的同一问题（Leviathan 等人，2023）。设计实验由此可以确定最能满足用户需求的呈现速度，并检验较慢地显示存储的答案究竟能否帮助用户，还是仅仅提高他们对答案的评价。

结果并不支持通过增加延迟来让回答显得更经过深思熟虑。在屏幕上逐步展开的回答中，较长等待时间在投票中并未获得可检测到的净加分，反而损失了后续问题。这一点很重要，因为推理模型在作答前会花费更多时间（Guo 等人，2025年；Snell 等人，2025年），正成为默认选择。额外时间只有在换来更好的回答时才值得花费；在我们的数据中，它本身并不能换来更好的印象。

7.4 对评估的影响

由盲投票构建的偏好排行榜已成为生成式人工智能的标准衡量尺度（Chiang 等人，2024年；Termignon 等人，2026年）。交付方式对它们意味着什么，取决于它们旨在衡量什么。如果排行榜衡量的是模型所写文本的质量，那么交付方式就是一个混杂因素：某个模型的回答若碰巧被更频繁地重放，就会因与其输出无关的原因而处于劣势。竞技场可以通过在相同的展示条件下呈现每个回答来消除这一因素，或者记录缓存、流式传输速度和推理展示并对其进行调整，正如它们已经对回答长度所做的那样（Singhal 等人，2024年）。如果排行榜衡量的是用户所体验的服务，那么交付方式就应计入分数，但这样一来，我们所衡量的延迟成本便未包含在其中。投票并未显示出较慢回答有任何可检测到的回报，但较慢的回答降低了用户继续对话的可能性。当两个回答都较慢时，偏好分数无法记录这一成本，因为投票是对两者进行比较：使两者都变慢的延迟不会改变胜者，而用户仍需承担这一成本。在比较系统时，后续提问率等持续对话指标应与偏好分数并列。

7.5 局限性与未来工作

该研究设置的若干特征指向了未来工作的方向。首先，重放会同时改变回答出现的速度、展开的方式，以及读者能够跟随回答的时间，因此实地数据能够识别这种交付方式的影响，但无法揭示其背后的心理路径。劳动幻觉提出了推断出的努力这一解释，而逐步展开的文本也可能帮助读者跟随回答、保持注意力，或让回答显得像是专门为该请求撰写的。设计实验可以区分这些路径：在相同时间安排下展示相同的回答，让文本以可读的速度逐步展开，或主要在最后突然出现，并告知一半读者，节奏由软件独立于模型所付出的工作量来设定。如果披露这一信息后某种影响减弱，而阅读舒适度保持不变，则说明作用来自推断出的努力；如果影响仍然存在，则说明作用来自阅读过程。在读者做出总体判断之前，询问他们某一具体且可核查的断言有多大可能正确，并提供准确率已知的回答，就能显示交付方式改变的是对准确性的信念，还是仅仅改变了选民给出的理由；我们的数据中的可选评价标签无法解决这一问题。单独改变第一个词出现前的延迟时间，还可以将这些结果与有关起始延迟的实验室证据联系起来。其次，重放资格是根据一个公开发布版本重建的，而该版本遗漏了一些对话，只包含已完成的回答，并且持续时间是在平台的服务器上测量的，而不是在访问者的浏览器中测量的。记录每一次请求以及每位访问者看到的内容，将会使这三方面的信息都更加精确。第三，拥堵估计值建立在这样一个假设之上：提供商的负载只会通过回答所需的时间影响访问者。受阻服务器也可能使界面本身变慢，而且拥堵还会略微缩短推理文本。在 WildChat 中，为每个时钟小时设置的固定效应可以吸收该小时内所有模型共同面临的需求，但无法吸收特定于某个模型的需求。提供商自身关于服务中断和负载的记录，可以提供更为清晰的变异来源。第四，持续对话是一种行为，而不是满意度或收益的衡量指标：用户可能只是因为问题已经得到回答而停止。将持续对话与任务成功联系起来，就能显示那些未被提出的问题究竟具有多大价值。第五，WildChat 只能通过哈希处理的 IP 地址识别用户，而且其公开发布版本会遗漏任何位置包含有害内容的对话，而这可能与对话持续了多长时间有关。最后，重放抽签涵盖的是那些被反复提出的提示词；与平均提示词相比，这些提示词更短，也更常由他人建议或布置。compar:IA 的受众主要讲法语，其中有许多学生，而 WildChat 用户选择的是免费服务。在商业助手中、使用其他语言，并采用独特且要求更高的问题进行复现研究，将有助于显示这些结果的适用范围究竟有多广。

8 结论

人工智能回答值得等待吗？在这些情境中，对于用户观看着逐步生成的回答，额外等待并未在其投票中带来可检测的净认可，而长时间等待则降低了他们继续提问的可能性。不过，相比于同样的回答在屏幕上逐步展开，用户对那些一出现便立即显示的回答评价更为苛刻。真正影响投票的是回答的呈现方式，而非所需时间。对于构建和评估人工智能系统的人而言，以下有两项测试：以可读的速度显示存储的答案是否能够消除惩罚，以及更短的等待时间是否能让用户继续提问。偏好投票能够反映第一点，却遗漏了第二点。

参考资料

Zeynep Akşin、Barış Ata、Seyed Morteza Emadi 和 Che-Lin Su。呼叫者的延迟敏感度的结构估计。*Management Science*，59(12):2727–2746，2013年。doi：10.1287/mnsc.2013.1730。

艾伦人工智能研究所。WildChat-4.8M。Hugging Face，https://huggingface.co/datasets/allenai/WildChat-4.8M，2025年。

Ioannis Arapakis、Xiao Bai 和 B. Barla Cambazoglu。网页搜索中响应延迟对用户行为的影响。载于《第37届国际ACM信息检索研究与开发大会》论文集，第103–112页，2014年。doi：10.1145/2600428.2609627。Fu Bang。GPTCache：一种用于大语言模型应用的开源语义缓存，可实现更快的回答并节省成本。载于《自然语言处理开源软件（NLP-OSS 2023）第3届研讨会》论文集，第212–218页，2023年。doi：10.18653/v1/2023.nlposs-1.24。Robert J. Batt 和 Christian Terwiesch。耐心等待：急诊科排队放弃的实证研究。《Management Science》，61(1)：39–59，2015年。doi：10.1287/mnsc.2014.2058。Ryan W. Buell 和 Michael I. Norton。劳动幻觉：操作透明度如何提升感知价值。《Management Science》，57(9)：1564–1579，2011年。doi：10.1287/mnsc.1110.1376。Ryan W. Buell、Tami Kim 和 Chia-Jung Tsay。通过操作透明度创造互惠价值。《Management Science》，63(6)：1673–1695，2017年。doi：10.1287/mnsc.2015.2411。A. Colin Cameron、Jonah B. Gelbach 和 Douglas L. Miller。多维聚类下的稳健推断。《商业与经济统计期刊》，29(2)：238–249，2011年。doi：10.1198/jbes.2010.07136。Wei-Lin Chiang、Lianmin Zheng、Ying Sheng、Anastasios Nikolas Angelopoulos、Tianle Li、Dacheng Li、Banghua Zhu、Hao Zhang、Michael Jordan、Joseph E. Gonzalez 和 Ion Stoica。Chatbot Arena：一个基于人类偏好评估大语言模型的开放平台。载于《第41届国际机器学习大会》论文集，《机器学习研究论文集》第235卷，第8359–8388页，2024年。compar:IA 团队。compar:IA 源代码。GitHub 代码仓库，https://github.com/betagouv/ComparIA，2026年。Jim Dabrowski 和 Ethan V. Munson。寻找最佳计算机系统响应时间的40年历程。《与计算机交互》，23(5)：555–564，2011年。doi：10.1016/j.intcom.2011.05.008。Emir Efendi´c、Philippe P. F. M. Van de Calseyde 和 Anthony M. Evans。缓慢的响应时间会削弱人们对算法预测的信任，但不会削弱对人类预测的信任。《组织行为与人类决策过程》，157：103–114，2020年。doi：10.1016/j.obhdp.2020.01.008。Dennis F. Galletta、Raymond Henry、Scott McCoy 和 Peter Polak。网站延迟：用户的容忍度有多高？《Journal of the Association for Information Systems》，5(1)：1–28，2004年。doi：10.17705/1jais.00044。Ulrich Gnewuch、Stefan Morana、Marc T. P. Adam 和 Alexander Maedche。人机聊天机器人交互中响应时间的相反影响。《商业与信息系统工程》，64(6)：773–791，2022年。doi：10.1007/s125 99-022-00755-x。Daya Guo、Dejian Yang、Haowei Zhang、Junxiao Song 等。DeepSeek-R1 通过强化学习激励大语言模型进行推理。《自然》，645(8081)：633–638，2025年。doi：10.1038/s41586-025-09422-z。Jan L. Guynes。系统响应时间对状态焦虑的影响。《美国计算机协会通讯》，31(3)：342–347，1988年。doi：10.1145/42392.42402。Elizabeth Han、Dezhi Yin 和 Han Zhang。带有情感的机器人：人工智能代理是否应在客户服务中表达积极情绪？《Information Systems Research》，34(3)：1296–1311，2023年。doi：10.1287/isre.2022.1179。Einav Hart、Eric M. VanEpps、Ovul Sezer 和 On Amir。速度是一种信号：更快的回复何时会提高录用可能性。《Management Science》，2026年。doi：10.1287/mnsc.2024.06185。Shenyang Jiang、Akshat Lakhiwal、Che-Wei Liu 和 Jiang Duan。看得更少，参与更多：重新思考生成式人工智能共创平台的早期用户体验——来自现场实验的发现。《Information Systems Research》，2026年。doi：10.1287/isre.2024.1200。Fei Jin 和 Shuying Gong。“逐字”还是“逐行”呈现？生成式人工智能中的响应生成动态如何影响用户接受度。《消费者行为期刊》，2026年。doi：10.1002/cb.70256。Ron Kohavi、Alex Deng、Roger Longbotham 和 Ya Xu。网站实验人员的七条经验法则。载于《第20届ACM SIGKDD国际知识发现与数据挖掘大会》论文集，第1857–1866页，2014年。doi：10.1145/2623330.2623341。Justin Kruger、Derrick Wirtz、Leaf Van Boven 和 T. William Altermatt。努力启发式。《实验社会心理学期刊》，40(1)：91–98，2004年。doi：10.1016/S0022-1031(03)00065-9。Daniella M. Kupor、Zakary L. Tormala、Michael I. Norton 和 Derek D. Rucker。思考校准：恰到好处的思考如何提升个人的影响力和吸引力。《社会心理与人格科学》，5(3)：263–270，2014年。doi：10.1177/1948550613499940。David S. Lee。培训、工资与样本选择：处理效应的精确界限估计。《经济研究评论》，76(3)：1071–1102，2009年。doi：10.1111/j.1467-937X.2009.00536.x。

Younghwa Lee、Andrew N. K. Chen 和 Virginia Ilie。在线等待可以得到管理吗？填充界面和呈现模式对在线感知等待时间的影响。MIS Quarterly，36(2):365–394，2012年。doi: 10.2307/41703460。Younghwa Lee、Andrew Chen 和 Traci Hess。在线等待体验：利用时间信息和干扰项缩短在线等待的感受时长。Journal of the Association for Information Systems，18(3):231–263，2017年。doi: 10.17705/1jais.00452。Yaniv Leviathan、Matan Kalman 和 Yossi Matias。通过推测解码实现 Transformer 的快速推理。载于《第40届 International Conference on Machine Learning 论文集》，《机器学习研究论文集》第202卷，第19274–19286页，2023年。Robert B. Miller。人机对话事务中的响应时间。载于《1968年12月9日至11日秋季联合计算机会议论文集》第一部分，第267–277页，1968年。doi: 10.1145/1476589.1476628。Ministère de la Culture。compar:IA 数据集：comparia-conversations、comparia-votes 和 comparia-reactions。Hugging Face，https://huggingface.co/datasets/ministere-culture/comparia-conversations，2026年。Andrea C. Morales。给予企业“努力”评价：消费者对高投入企业的反应。Journal of Consumer Research，31(4):806–812，2005年。doi: 10.1086/426615。Fiona Fui-Hoon Nah。关于可容忍等待时间的研究：Web 用户愿意等多久？Behaviour & Information Technology，23(3):153–163，2004年。doi: 10.1080/01449290410001669914。Jonathan W. Palmer。网站的可用性、设计和性能指标。Information Systems Research，13(2):151–167，2002年。doi: 10.1287/isre.13.2.151.88。Jennifer Rix、Benedikt Berger、Thomas Hess 和 Christine Rzepka。算法折价：解释消费者对人类创作与算法创作的数字产品的估值。Journal of Management Information Systems，42(2):633–668，2025年。doi: 10.1080/07421222.2025.2487308。Scott Schanke、Gordon Burtch 和 Gautam Ray。估算“拟人化”客户服务聊天机器人的影响。Information Systems Research，32(3):736–751，2021年。doi: 10.1287/isre.2021.1015。Ben Shneiderman。人类计算机操作绩效中的响应时间和显示速率。ACM Computing Surveys，16(3):265–285，1984年。doi: 10.1145/2514.2517。Prasann Singhal、Tanya Goyal、Jiacheng Xu 和 Greg Durrett。任重道远：探究基于人类反馈的强化学习中的长度相关性。载于《第一届 Conference on Language Modeling (COLM) 论文集》，2024年。Charlie Snell、Jaehoon Lee、Kelvin Xu 和 Aviral Kumar。对于推理任务，最优扩展大语言模型测试时计算可能比扩展参数更有效。载于 International Conference on Learning Representations (ICLR)，第10131–10165页，2025年。Felicia Fang-Yi Tan、Moritz Alexander Messerschmidt、Wen Yin 和 Oded Nov。响应延迟和任务类型对人类与大语言模型交互及感知的影响。载于 Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems，第1–17页，2026年。doi: 10.1145/3772318.3790716。Shirley Taylor。等待服务：延迟与服务评价之间的关系。Journal of Marketing，58(2):56–69，1994年。doi: 10.1177/002224299405800205。Lucie Termignon、Simonas Zilinskas、Hadrien Pélissier、Aurélien Barrot、Nicolas Chesnais 和 Elie Gavoty。compar:IA：法国政府用于收集法语人类提示词和偏好数据的大语言模型竞技场。arXiv 预印本 arXiv:2602.06669，2026年。Dimitrios Tsekouras、Ting Li 和 Izak Benbasat。你帮我挠背，我也帮你挠背：用户努力和推荐代理努力对感知推荐代理质量的影响。Information & Management，59(1):103571，2022年。doi: 10.1016/j.im.2021.103571。Shunyuan Zhang 和 Das Narayandas。通过在线聊天中的人工智能吸引客户：来自随机现场实验的证据。Management Science，72(1):73–95，2026年。doi: 10.1287/mnsc.2022.03920。Wenting Zhao、Xiang Ren、Jack Hessel、Claire Cardie、Yejin Choi 和 Yuntian Deng。WildChat：真实场景中的100万条 ChatGPT 交互日志。载于 International Conference on Learning Representations (ICLR)，第34590–34605页，2024年。David Zhou、John R. Gallagher 和 Sarah Sterman。深思熟虑、困惑还是不可信：文本呈现如何影响人们对人工智能写作工具的看法。载于《2025年 Conference on Creativity and Cognition 论文集》，第573–589页，2025年。doi: 10.1145/3698061.3726907。

表11：研究1：重播与参与度（每次重播的回答对应的百分点）。

	均值 (%)	资格 × 周固定效应	+ 提示词固定效应
提出第二个问题	8.8	0.8 (0.6)	0.2 (0.4)
投出最终票	22.8	−1.7 (0.7)	−0.3 (0.6)
回应消息	5.1	−0.8 (0.4)	0.2 (0.4)
7天内其他日期的回访	1.6	−0.5 (0.2)	−0.2 (0.2)
投票认为两个回答相同（在投票中）	36.1	−9.8 (2.6)	

注：共有 6,898 次对话，至少包含一个符合条件的首个回答。回归变量统计重放的首个回答数量（0、1 或 2）；固定效应将符合条件的回答数量与周进行交互。平局一行使用 1,574 张投票，并采用一个指标，标示是否恰好有一个回答被重放。标准误按访问者聚类。

表12：拥塞工具变量与对话特征。

	Mean	工具变量系数	
		日历日效应	确切时钟小时
研究2（*compar:IA*）：相同模型，相同日期			
提示长度（字符数对数）	4.790	−0.045 (0.026)	
访问者此前的对话（对数）	0.911	−0.007 (0.026)	
科学与技术主题（占比）	0.209	−0.013 (0.008)	
访客选择的模型（占比）	0.261	0.006 (0.010)	
回答长度（字符对数）	7.467	−0.018 (0.019)	
推理长度（字符对数）	0.919	−0.091 (0.028)	
研究3（*WildChat*）：相同模型，相同时段			
用户消息长度（字符数对数）	5.682	−0.276 (0.047)	−0.017 (0.049)
回复长度（字符数对数）	6.346	−0.243 (0.054)	−0.047 (0.038)
美国用户（占比）	0.208	−0.051 (0.008)	−0.023 (0.010)

注：每一行均将该特征对拥堵工具变量进行回归，并纳入日历日固定效应（研究2，主要设定：模型对 × 月、日、星期小时；研究3：模型 × 月、日、星期小时、轮次）；最后一列为研究3的首选设计，用确切时钟小时替代日和星期小时。研究2工具变量的标准差为0.25对数单位（加入固定效应后为0.10）；研究3工具变量的标准差为0.18。标准误采用双向聚类（研究2：访问者和日；研究3：哈希 IP 和时钟小时）。

A 补充结果

A.1 研究1

表11 报告了研究1中重播与持续对话及反馈之间的关系。回归变量统计一段对话中被重放的首个回答数量；固定效应则将符合条件的回答数量与周进行交互，因此比较的是符合条件的回答恰好被重放或未被重放的对话。最后一行显示，恰好有一个回答被重放的对话，以投票判定两个回答相同而结束的可能性较低。

A.2 研究2

表12 将拥堵工具变量与拥堵不应影响的对话特征联系起来。对于 compar:IA，模型的同日拥堵与提示词长度、访问者经验、自定义模型选择的占比以及回答长度无关。在拥堵日，回答包含的推理文本较少。对于 WildChat，当回归仅控制星期几和小时的固定效应时，拥堵时段的消息和回复更短，美国用户也更少。当为每个时钟小时设置固定效应后，长度差异消失，美国用户占比的差异也缩小（第6节）。表13 报告了后续结果的替代设定，而表14则按主题以及限定标签分别报告了投票回归结果。主题估计属于探索性分析，在10个主题中，每对数单位的估计值介于 −8.7至 +4.3点之间；只有文化与艺术主题的一个正估计值（+4.1，2.0）约为距离零值两个标准误，这在未针对多重比较进行调整的十个估计值中并不罕见。

表13：研究　2：后续结果的稳健性（除非另有说明，单位为每迟缓程度对数单位的百分点）　.

因变量：提出第二个问题	OLS	IV
主要设定（表8）	−1.8 (0.2)	−2.0 (1.0)
工具变量仅排除焦点对话		−1.9 (1.0)
仅按访问者聚类的标准误		−2.0 (0.8)
按访问者和周聚类的标准误	−1.8 (0.2)	−2.0 (1.1)
访问者内部，按访问者和周聚类		−2.4 (1.2)
回归中不包含长度和推理控制变量	−2.2 (0.2)	−2.0 (1.0)
回归变量：生成时间对数，未净除长度影响		−2.0 (1.0)
简化式：拥堵工具变量本身		−1.8 (0.9)
回归变量：较长等待时间的对数秒数	−1.9 (0.2)	−2.2（1.1）
该组中没有推理模型	−1.6（0.3）	−1.5（1.0）
首次对话的访问者	−2.1（0.3）	
最慢的五分之一对话（指标）	−2.3（0.3）	
缓慢程度的第80百分位数以下的斜率	0.2（0.3）	
缓慢程度的第80百分位数以上的斜率	−5.1（0.4）	

注：固定效应、控制变量和双向聚类均与表8 相同，除非另有说明。简约式行报告了拥堵工具变量每增加一个对数单位的影响。在不含长度和推理控制的行中，慢速程度和工具变量仍按净除长度影响进行测量；在以生成时间对数为回归变量的行中，工具变量不变，回归控制了长度和推理长度。指标和样条函数行采用具有相同固定效应的普通最小二乘法；慢速程度的第80百分位数为0.41个对数单位。

表14：研究　2：按主题划分的慢速程度与盲投，以及限定标签（普通最小二乘法，每对数单位的百分点).

	较慢的回答	决定性投票
按对话主题划分，较慢的回答获胜		
社会	−8.7 (4.1)	1,680
科学与技术	−3.5 (1.0)	15,513
教育	−2.1 (0.8)	19,551
个人与日常生活	−1.7 (1.7)	7,531
商业与金融	−0.9 (1.7)	7,514
健康与环境	0.2 (1.9)	5,821
政治与法律	0.6 (1.6)	7,603
娱乐与爱好	1.1 (1.5)	8,303
文化与艺术	4.1 (2.0)	5,184
其他	4.3 (5.7)	1,505
限定标签（左减右），所有投票		
有用	0.3 (0.3)	
完整	1.0 (0.4)	
格式清晰	−0.7 (0.4)	
有创意	0.1 (0.2)	
肤浅	−0.1 (0.3)	
未遵循指令	0.1 (0.2)	
不正确	0.4 (0.2)	

注：采用无序模型对的普通最小二乘法 × 月份固定效应、长度和推理控制变量（表7），按主题分别估计（至少有1,500 次决定性投票的主题）；探索性分析，未对多重比较进行调整。标准误按访问者和日期聚类。

B 缓存和服务器的实现

表15 列出了平台公共源代码的变更（compar:IA 团队，2026年），这些变更定义了两种设计，并附有拉取请求和提交的链接。代码仓库记录的是代码编写和合并的时间，而不是部署时间。因此，合并日期被推断为部署日期；对于缓存，数据确认了这一日期，因为首次重放出现在2026年3月16日，也就是缓存合并当天。

数据中的重放占比——截至4月5日为符合条件的回答的42%，此后为74%——与部署概率先约为0.4、随后约为0.75的情况一致。10月12日的第二次变化解释了为何我们观察窗口中的重播速度很快。在阅读循环的每一轮中，服务器会同时读取比较中的两个回答，并在另一个回答到达后取消较慢的读取；重放的暂停就是这样一次读取，因此每当另一个回答生成文本时，重放就会结束。

表15：公共代码仓库中与设计相关的 compar:IA 代码。

已合并	更改
2026年3月16日	首轮回答缓存。密钥：模型和初始提示，移除首尾空格并忽略大小写字母。每个密钥最多缓存五个回答，在该密钥最后一次写入或读取后 48小时过期。每次首次轮次请求中，均匀随机抽取决定是否查找密钥；如果找到，则会随机选取一个存储的答案并重放。该概率是部署设置（当前代码中默认值为 0.5）。拉取请求 #379，提交 83eb461。
2026年3月16日	重放显示：存储文本会分段发送，每段至少 20 个字符（较长的回答大约分为十五段）每段发送后暂停 0.2 秒；存储的推理文本会一次性发送。如果暂停被取消，重放会立即继续处理下一段。记录的时长从重放开始一直计算到其最后一段。相同的拉取请求。
2026年3月16日	回答取自所显示回答的最终文本。提交 7d5db50。
2026年10月12日	修复服务器在等待提供商时“占用唯一的事件循环”的问题，从而避免缓慢的模型导致延迟每位访问者的请求。第二项变更让每个答案流按自身的速度进行；其说明指出在此之前，“较慢的读取会在另一个模型回答后立即被取消”，而缓存重放忽略了取消操作，“只是为了撑过那些被取消的读取”。拉取请求 #795，提交 2364b6c 和 34ae6bc。

较慢的读取；重放的暂停就是这样一次读取，因此每当另一个回答生成文本时，重放就会结束。

C 重建重放资格

我们针对从第一次已记录的重放开始的每个首次回答，重建存储中是否保存了其密钥。若满足以下两个条件，回答便符合条件。第一，自缓存开始后且在该回答之前，同一模型至少针对同一初始提示词给出过一次新生成的首次回答；比较提示词时，采用平台的比较方式，即去除首尾空格并忽略字母大小写。第二，该模型和提示词此前的首次回答（无论是新生成的还是重放的）距今不超过48小时。仍有两个误差来源。例如，未包含在公开发布版本中的对话，可能因为包含个人数据而被移除（Termignon 等人，2026年），可能创建或刷新了一个我们无法观察到的密钥。因此，我们归类为不符合条件的一些回答实际上符合条件：窗口期内的8,604次重放中有509次（6%）属于这一组，因而未被使用。反过来，某个密钥可能提前过期或未能写入，因此我们归类为符合条件的一些回答实际上不可能被重放；这些回答始终是新生成的，而将回答与同一提示词进行比较，也限制了它们可能造成的影响。将投票样本限制为密钥在过去24小时内曾被使用的回答，或限制为存储在过去48小时内曾提供过重放的回答，所得估计基本相同（表 6）。按照精确文本而非平台规则匹配提示词后，在全部流量中识别出的符合条件的回答为12,057个，而不是12,518个。在样本中的2,460个模型—提示单元格中，有474个同时包含重放回答和新生成的回答；它们包含11,699个符合条件的回答中的9,339个，并提供了提示词内部的比较。

D 样本构建

表16 展示了各分析样本如何从公开发布版本中逐步筛选得出。对于 compar:IA，每项研究都从五种标准模式下且不含自动化流量的对话开始。研究1随后保留重放窗口内双方首个回答均有记录的对话，其投票和答案级样本则依据C节的资格规则筛选得出。研究2保留双方回答均为新鲜生成、经过计时且长度足以进行长度调整，并且能够计算拥塞度量的首次交流。对于 WildChat，筛选条件会剔除延迟不合理或文本为空的回复，以及活跃度过高的自动化用户；随后还要求有足够多的相似回复，以便调整内容因素，并要求同一小时内有足够多其他用户的回复，以衡量拥堵。

表16：分析样本。

compar:IA（截至*2026*年*5*月*18*日的发布版本）	
发布版本中的对话	410,016
五种标准模式中的对话	376,662
剔除119个自动化标识符对应的13,886段对话后	362,776
研究*1*：*2026*年*3*月*16*日至*2026*年*5*月*18*日	
记录了首个回答的对话	57,859
至少包含一个符合条件的首个答案	6,898
两个首个回答均符合条件	4,801
含有符合条件回答的对话中的最终投票	1,574
决定性投票（主要投票样本）	1,005
符合条件的回答（回答层面分析）	11,699
研究*2*：*2025*年*2*月至*2026*年*5*月	
首个回答均计时的对话	345,809
均为新鲜、计时且长度可调，并包含工具变量（主样本）	324,029
其中的决定性投票	80,205
后续轮次（第2–6轮）	140,671
具有记录时间的流式传输的首个回答（回答层面分析）	659,969
WildChat-4.8M（*2023*年*4*月至*2025*年7月）	
发布版本中的回复	5,167,182
延迟不超过十分钟、非空，且来自每天最多进行 50 次对话的 IP地址	4,543,659
在至少包含 20 条回复的内容单元格中，使用拥塞度量（主样本）	4,378,514
对话	2,694,924
哈希处理的 IP 地址	1,829,326

注：自动化标识符涵盖所有模式下的 14,893 个对话，其中 13,886 个对话采用标准模式。研究2的计时计数包括所有模式和流量；主样本应用了其上方的筛选条件。研究2中的回答层面分析包括其配对对象未计时的回答，因此回答数量多于对话样本数量的两倍。