优化，而非另起炉灶：面向源文本忠实型印地语散文润色的 RASA-Bench

Abdur Rahman* **PrathamSingla** **Mohammad Saifullah** **VasuSharma**Pocket FM，印度班加罗尔{rahman.abdur，ext-pratham.singla，mohammad.saifullah，vasu.sharma}@pocketfm.com

arXiv:2611.29750v1 [cs.CL] 22 Nov 2026

摘要

对于印地语创意散文的创作者而言，起草成本低，编辑却不然：机器撰写的剧集、故事或散文读起来往往内容单薄、平淡，充斥着翻译腔，有时还会悄然改变源文本的含义。

RASA-Bench 面向三种印地语散文形式评测这项编辑任务，衡量模型优化欠充分撰写的人工智能初稿的能力；这与从零开始写作是不同的任务。基于 2,050 份草稿和八种精炼器（共 16,400 篇改写文本），四层协议只有在改写文本通过针对其源文本的忠实度检查后，才予以认可；随后，通过按形式进行的 Bradley–Terry模型拟合，根据整体编辑偏好对通过检查的文本进行排名，并将中心化系数标准化后等权平均。忠实度是一项决定性考验，且在不同形式中的表现各异：关卡失败率从短篇小说的 9.1% 到音频小说的 29.5% 不等；导致改写文本不合格的偏移，在叙事作品中涉及事件和身份，在散文中则涉及论断和证据。对 600 个案例进行的人工审查显示，门控一致性为 0.925（各形式为 0.91 至 0.94）；而主观偏好因形式而异，且在散文中最难达成一致，这一点在人类读者中同样如此，模型评审也不例外。基准测试、评测协议和印地语散文诊断可在 https://abdur75648.github.io/rasa-bench 获取。

1 引言

连载音频小说、文学短篇小说和文学散文都在大规模制作，以满足印地语受众的需求（Mann等人，2026年），而基于大语言模型的流程会先规划大纲，再据此生成散文草稿（Teleki等人，2025年；Wang等人，2024b；Kazi等人，2026b；Saifullah等人，2026年）。如此生成的草稿很少能直接发表：印地语文本往往显得内容单薄、平淡，还夹杂着翻译腔（Rahman等人，2026年），而且有时还会悄然改变源文本所确立的内容。因此，实际生产中需要进行润色稿：在保留内容完整的前提下，通过编辑工作将可用的草稿打磨成可发表的散文（Kazi等人，2026a）。高质量的润色能提升语言表达；而一项较差的润色稿会原样保留问题；而不忠实的润色稿则会改写故事或论点。要做好评估，就必须区分真正的散文改进与偏离源文本的流畅改写。

评估的发展未能跟上步伐。英语领域的创意写作和故事生成基准测试已相当成熟（Wu 等，2025年；Fein 等，2026年；Chhun 等，2022；Yang 和 Jin，2024；Chen 等，2022），印度语系基准测试则涵盖翻译和广泛能力（Sai B 等，2023年；Kumar 等，2022；Singh 等，2024；Verma 等，2025年；Gala 等，2024；Kamath 等，2025年），而偏好排行榜则根据开放式输出之间的比较对其进行排名（Zheng 等，2023）。即使在英语领域，基于源文本的改写也鲜少受到关注（Shu 等，2023；Raheja 等，2023；Madaan 等，2023；Chakrabarty 等，2025）；对于非英语散文而言，其失败模式更为突出，但相关研究仍寥寥无几。我们同时针对三种不同的印地语散文形式填补这一空白。

我们提出 RASA-Bench（**R**润色**A**源文本驱动的 **S**评估 **AI**），这是一项针对音频小说、短篇小说和文学散文中印地语散文润色的形式条件化、忠实度门槛化、经人工校准的基准测试。每个模型都会获得一份欠充分撰写的人工智能初稿，并必须对其进行润色；只有在通过针对源文本的忠实度门槛后，润色稿才能获得质量分数，而通过门槛的改写文本随后会在整体偏好上展开竞争（图 1）。该基准测试包含三种形式的 2,050 份源草稿，由八个模型生成 16,400 份润色稿，并通过四层协议进行评估，其中主要评审器经过印地语读者校准。

* 通讯作者。



图 1：以一个对象展示 RASA-Bench；机制以一个叙事例子说明，但适用于全部三种形式。对于一份描写不足的草稿，叙述者的双手紧握方向盘；违反忠实度的改写文本反转了动作，因此在评判前被门控排除；一份较弱但忠实的改写文本套用了“白 knuckles”的仿译，并保留了英语字母书写的术语；而一份出色的改写文本忠于原文，将仿译替换为适用于印地语的紧张感描写，并以天城文呈现该术语。通过门槛的改写文本随后按照整体偏好展开竞争。

两项发现构成了结果的主线。第一，忠实度本身就是一个决定性的、形式条件化的 信号：改写文本必须保留的内容因形式而异；不同形式的关卡失败率介于 9.1% 至 29.5% 之间，而不同模型之间的范围更广。因此，偏离源文本的流畅改写会在散文质量评分之前被筛除。第二，三种形式构成了一个难度梯度：客观的忠实度门槛在每种形式中都很可靠，但在音频小说中区分度最高；主观偏好在短篇小说中最为明确，而在散文中则确实呈现出较低的共识度——无论是对人类读者还是自动评审器而言，都是如此。排行榜顶部在不同形式间保持稳定；中部则重新排序，并通过等形式标准化宏观指标将三种形式结合起来。我们的贡献包括：

- **C1.** 首个基于专业编辑和出版的人类源文本构建的基准测试，用于优化三种不同形式的印地语散文：音频小说、文学短篇小说和文学散文（§3）。

- **C2.** 一种形式条件化、经忠实性门控的成对偏好协议，采用按形式均衡标准化的Bradley–Terry模型聚合，并经人工校准，用于区分基于源文本的改写质量与跨形式流畅但不忠实的改写（§4、§5）。该协议将忠实性门控过滤和 Bradley–Terry模型偏好方法适配到这一场景，并复用成熟的偏好评估组件。

- **C3.** 一种形式条件化的印地语散文诊断特征，主要体现为程式化短语负担以及仿译／翻译腔；通用指标无法捕捉这一特征。该特征以描述性方式报告，不纳入排名，在音频小说中最为明显，在书面形式中则较弱（§6）。

2 相关工作

创意写作和故事评估。 越来越多的研究开始对开放式写作质量进行基准测试，包括 WritingBench（Wu 等，2025年）、LitBench（Fein 等，2026年）、HANNA 人工标注故事语料库（Chhun 等，2022）、长篇连续音频剧（Mannet 等，2026年），以及关于故事质量测量的综述（Yang 和 Jin，2024；Chen 等，2022）。这些工作针对自由形式生成，并报告了主观散文中常见的中等偏低的人类一致性；例如，HANNA 仅报告了中等偏低的标注者间一致性，这与我们的观察一致，在散文中尤为明显。没有任何研究涉及

基于源文本的改写，尤其是文学散文，仍基本未得到研究。

改写、润色与低质量人工智能文本。 如何改进给定文本，已有研究将其作为指令微调编辑（Shu 等人，2023年；Raheja 等人，2023）、迭代式自我优化（Madaan 等人，2023），以及跨文化长篇故事的内容保持式改写（Kazi 等人，2026a，b）；基于编辑的质量奖励也可用于从草稿中筛选出更优的修订版本（Chakrabarty 等人，2025年）。另有一条研究路线定义并测量“低质量人工智能文本”：从可解释的维度衡量人工智能生成的低质量文本（Shaib 等人，2025），以及相较于人类写作中重复且过度常见的措辞（Paech 等人，2026年）。我们将优化而非生成的框架拓展至非英语语境，并加入经人工审核、针对特定语言的诊断指标。

印度语系与印地语评估。 印度语系评估主要聚焦于翻译、生成和广泛能力，包括 IndicMT Eval（Sai B 等人，2023年）、IndicNLG 基准（Kumar 等人，2022）、IndicGenBench（Singh 等人，2024），以及 MILU 等多任务套件（Verma 等人，2025年），同时也涵盖针对印地语的指令微调模型和基准测试（Gala 等人，2024；Kamath 等人，2025年）。这些工作针对翻译、事实性和理解任务；而印地语创意与文学创作，尤其是润色，仍基本缺乏基准测试。

大语言模型评审、成对偏好与校准。 成对比较是现代偏好排行榜的基础（Zheng 等人，2023年），并通过 Bradley–Terry模型对投票进行聚合（Bradley 和 Terry，1952；Chiang 等人，2024）；基于评分标准的评审器和模型评审器如今已成为标准做法（Liu 等人，2023年；Kim 等人，2024）。人们一直在积极研究其可靠性：只有提供正确参考答案时，评审器的判断才大体能与人类一致（Krumdick 等人，2025年）；而设计选择可能导致排名受噪声主导（Feuer 等人，2025年）。与我们的实验方案最接近的是 WritingPreferenceBench（Ying 等人，2025年），它会匹配在客观正确性上相当的回答配对，从而只评判主观质量。我们则采用来源忠实度门槛，并将其与整体成对偏好和人工校准相结合，专门用于三种散文体裁的印地语润色；据我们所知，尚无先前的基准测试在这一场景中将来源忠实度门槛与整体偏好结合起来。

3 基准构建

任务。 RASA-Bench 评估三种印地语书面体裁的散文润色：连载音频小说、文学短篇小说和文学散文。给定一个写得不够充分的人工智能草稿 S（一集音频小说、一个短篇小说或一篇散文），模型会生成一个润色版本 R，该版本应读来自然，使用符合语体的印地语，同时保留草稿确立的一切。这正是每种体裁中制片人或编辑实际面临的任务。模型的原始输出很少能直接用于出版或播出，但生成成本低；编辑工作的重点是在不改变草稿所述内容或情节的前提下提升语言表达。规划工作已在前期完成并作为给定条件；因此，我们衡量的是执行情况，即相对于给定草稿的改进，而且只有在改写忠实于源文本时才予以认可。这些体裁遵循相同的润色约定，但忠实度所保护的内容各不相同：叙事体裁保护事件、人物和揭示的信息，而散文则保护论断、证据和推理脉络。

构建流程。 图2 总结了构建过程，三种形式的结构完全相同。每个项目都从一份黄金标准的人类源文本开始 G：对于音频小说，是来自私人当代语料库、经过专业编辑的一集节目；对于书面形式，则是取自公开及授权文学作品集的已出版印地语短篇小说或文学随笔（附录 J）。一份形式中立的英文大纲，即上游流程会生成的结构规划，驱动一个留出生成器池；池中的每个生成器都将其执行为一份欠写作的草稿 S；一个形式条件化的源文本对黄金标准筛选器会接受 S 核心内容和忠实度（对于音频小说，还包括安全性）；随后 G 被分离。通过构建过程可得出两个性质。生成器池与润色器名单相互隔离，因此没有任何排行榜模型会润色自己的草稿。并且 G 构建以其为锚点，但它从不参与竞争：由于 R 经由有损且由大纲介导的草稿，源自 G，针对 R 进行评分 G 会奖励对原文的重构；因此，门控会将每个 R 与其自身的草稿进行比较 S，而通过门槛的润色结果则相互排名，因此该基准测试衡量的是对给定草稿的润色。逆向构建的基准测试同样从一份完成的人类文档中派生出每项任务，然后使用不断优化的检查清单对输出进行评分，直至该文档满足每项标准（Pham 等人，2026年）；RASA-Bench 在评估前将 G 分离。只有实例化输出

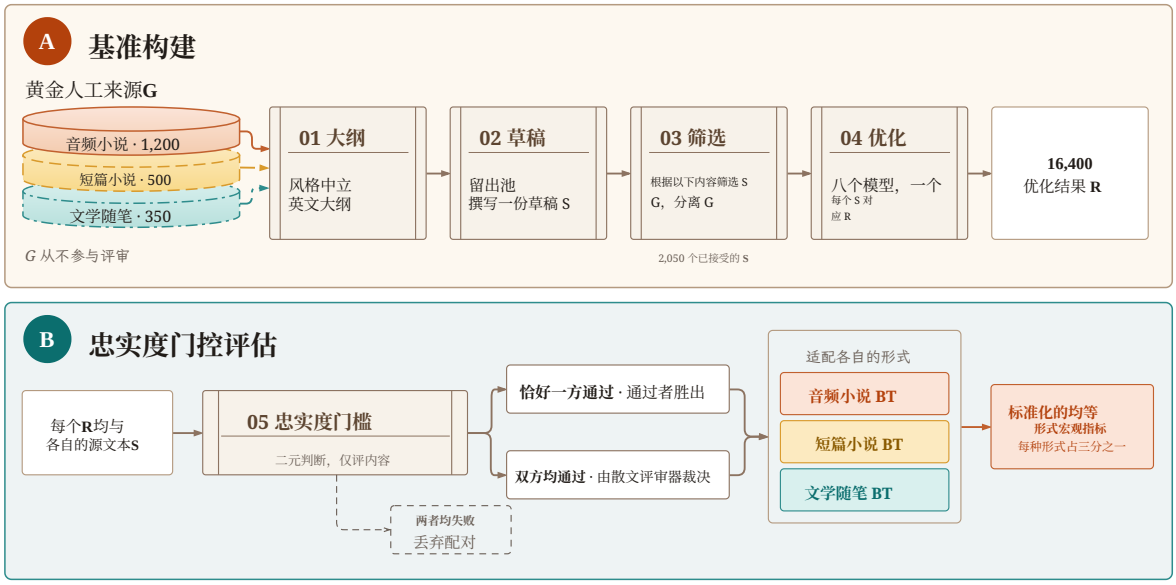

图 2：RASA-Bench 的构建与评估。(A) 将三种散文形式的人工黄金源文本 G 转换为风格中立的英文大纲；留出生成器据此生成欠充分撰写的草稿 S，按形式筛选的筛选器会逐一核验草稿是否符合 G，然后再将其分离 G。八个精炼器基于 2,050 份通过筛选的草稿生成 16,400 份输出 R。(B) 每份 R 仅根据其自身的源草稿 S进行检查：只有一份通过时，该份胜出；两份都失败时则弃用；两份都通过时，则比较散文质量。按形式计算 Bradley–Terry 系数，中心化并缩放至标准差为 1，再按各形式等权取平均。 G 绝不参与排行榜评审。

行、草稿和可接受性标准会因形式而异。

转换保留了什么。筛选器会拒绝 19.0% 的候选音频小说草稿、19.9% 的短篇小说草稿和 11.2% 的文章草稿；在音频小说中，被拒绝的草稿大多是因为事件发生了变化或缺失、角色或关系发生了变化，或结局有所改变。被拒绝的草稿会被替换，因此每份保留的草稿都通过了核验，确认其保留了 G的核心故事或论点（附录J）。保留的草稿比 G 更短（在两种书面形式中，其平均长度均为 G的 0.66 倍）：按照设计，大纲不提供写作技巧，而润色所恢复的正是这些内容。八个精炼器接收完全相同的草稿，因此比较是在相同的草稿分布上进行的。我们将手头的草稿视为忠实度目标，因为编辑润色的是流程交付的内容；若要求精炼器忠于其从未见过的内容，评估的就会是对原文的重建。

语料库。最终纳入的源草稿共 2,050 篇，由八个模型进行润色，得到 16,400 份优化结果：1,200 篇音频小说草稿，涵盖 82 部剧集（9,600 份优化结果）；500 篇短篇小说（4,000 份）；以及 350 篇文学散文（2,800 份）。源文本按类型和长度进行分层抽样。音频小说源文本涵盖约十一个以当代题材为主的类型家族，其中奇幻与修仙（27.9%）以及家庭与社会剧（24.4%）占比最高；短篇小说涵盖一般文学叙事（71.6%）、寓言与民间故事（23.4%）、儿童故事（4.6%）以及实验性故事；散文则包括文化（59.7%）与文学（31.7%）反思、个人反思、文学批评，以及议论性或说明性写作（完整分类见附录 J）。草稿刻意写得不够充分：忠实的润色稿会提升散文质量，同时保留源文本的内容。对于音频小说，忠实的润色稿通常也会加长草稿，因此该流程会对此控制篇幅（§5）；所有形式的草稿和润色稿篇幅均见附录 J。

4 评估协议

该协议是一种形式条件化、由忠实度门控的成对偏好评估。它借鉴了两种成熟设计：秉承 Ying 等人（2025年）理念的忠实度门控筛选，以及秉承 Zheng 等人（2023年）理念的整体成对比较，并采用 Chiang 等人（2024）所用的基于成对投票的 Bradley–Terry 排名；该协议专门针对基于源文本的印地语优化而设计，并依据印地语读者进行了校准。评估分为四个层次进行。每个层次在三种体裁之间共享同一架构，仅通过体裁卡进行专门化；附录 H 列出了提示堆栈和

Rank	模型	标准BT（95% 置信区间）	Win	Fail	J1	J2	J3
1	GPT-5.5	1.783 [1.729, 1.831]	0.873	7.7%	**3.95**	**3.56**	**3.59**
2	Claude Opus 4-6	1.268 [1.206, 1.322]	0.790	13.9%	3.82	3.55	3.54
3	DeepSeek V4-Pro	0.146 [0.102、0.194]	0.544	8.9%	3.55	3.43	3.40
4	Gemini 3.1 Pro	-0.033 [-0.105, 0.031]	0.502	33.8%	3.74	3.46	3.45
5	Grok 4.3	-0.409 [-0.460, -0.355]	0.410	27.3%	3.61	3.46	3.43
6	Gemma 4 26B	-0.613 [-0.659, -0.566]	0.358	27.8%	3.55	3.40	3.38
7	Sarvam 105B	-1.005 [-1.034, -0.970]	0.268	10.9%	3.27	3.37	3.31
8	Qwen3.5 27B	-1.136 [-1.171, -1.100]	0.239	18.3%	3.17	3.37	3.34

表1：跨体裁头条排行榜，按等权形式标准化 Bradley–Terry模型（标准BT）排名。先将按形式居中的系数缩放至标准差为 1，再进行等权平均。胜率和失败率是未标准化的等权形式平均值，分别对应经过门控的实证胜率（平局计半）和门控失败率。J1–J3 是按形式划分的 1–5 分诊断均值（各列最佳值以粗体标出），用于描述性地解读各形式内的表现，不参与排名。各形式的表格见附录 A。

形式条件化内容，并简要说明任务、门控和标题评判提示词的规范。

第一层：忠实度门槛。 在进行任何质量比较之前，二元门控会评审每份润色稿R 是否忠于其源草稿 S ，以检验内容保真度；这种做法承袭了生成文本的源文本一致性检查思路（Laban 等人，2022年）：无论改写读起来多么流畅，只要背离源文本，就会判定为失败。违规类型体系依形式而定。对于叙事体裁，它涵盖事件、事实、时间线、揭示信息、对话含义、角色关系的改变，以及虚构内容；这些都是长篇虚构作品系统所追踪的叙事状态（Saifullah 等人，2026年）；对于文章，它涵盖论断被更改或删去、证据被篡改或虚构，以及限定或推理发生转变。门控会检查S 并首先运行，因此，偏离源文本的流畅改写无法凭散文质量胜出。这正是「优化，而非另起炉灶」*refine, don't reinvent*的保障。

第2层：标题成对评审器。对于每个源文本，通过门控的润色稿由整体评审器进行正面对比，并给出单一判定结果（A、B 或平局），而非按维度评分的评分表；评审时遵循体裁特定的编辑标准：剧集采用音频表演文稿的标准，短篇小说采用叙事技巧的标准，散文采用论说文写作技巧的标准。在这种设置下，相对成对标题指标更为合适：在控制客观信号后，评审器对主观写作偏好的判断与人类的一致性仍然较弱（Ying 等人，2025年），因此我们仅将绝对分数作为次要诊断指标，并像公开偏好排行榜那样按成对比较进行排名（Zheng 等人，2023年）。每一对润色稿都按两种呈现顺序进行评审，并合并为一个源级结果，以解决呈现顺序偏差（Wang 等人，2024a）；若两种顺序下的评审仍存在分歧，则合并为平局（附录 B）。门控与质量评审器共同决定每次比较的结果：若一对润色稿中恰有一份通过门控，则判该润色稿胜出；若两份均未通过，则舍弃该对；若两份均通过，则由质量评审器裁定。合并后的结果用于拟合 Bradley–Terry 模型，并在每种体裁内，基于该体裁相同源文本的比较单独进行中心化；平局（包括两种顺序下仍存在分歧的情况）均按各模型各得半胜计入。标题分数为各体裁等权标准化 *Bradley–Terry 宏观分数*：将每个中心化的八模型系数向量除以其总体标准差，再对每个模型在三种体裁中的标准化系数取平均。因此，各体裁权重相等，系数离散程度也相同。

第3层：维度评判器。 三位采用 1–5 分制的诊断评审器会分解模型之间的差异所在，并在不同形式间共享一套核心维度：语言与自然度（J1）、构成与戏剧化（J2），以及声音与对话（J3）；每个维度均依据特定形式评分标准进行评判。这是一种次级且可解释的分解；由于这些评审器与 Headline 评审器属于同一模型族，因此它们与 Headline 评审器的一致性仅具描述性，且未收集人工维度标签。记录的验证依据来自第2层中的人工比较。

第4层：诊断取证。 一个确定性的词汇层用于衡量印地语特有的散文特征，主要包括经人工审阅的程式化短语负担，以及仿译／翻译腔诊断。这些内容将在此处定义，并在§6中测量；它们仅作描述，绝不纳入排名。

评判器栈与校准。单一主-

要评审器 GPT-5.5 负责门控、标题比较和各项维度评估；该评审器是在印地语与编辑一致性试点中选定的，且独立于排行榜结果，并通过标题任务中的人工评审进行了验证。为探究是否由同一评审器家族决定排名，我们让两个独立家族的审计评审器（Claude Opus 4-6 和 Kimi-K3）重新评估了涉及顶级模型且通过门槛的音频小说标题比较，以检验稳健性（§5）；书面形式则由单一主要评审器评估，这一范围限制已在“局限性”中说明。人工校准被纳入设计：印地语读者在每种形式中都执行了完全相同的整体成对任务，并且忠实度门槛在每种形式中都通过人工标注样本进行了验证（§5）。我们校准的目标是“处于人类一致性范围内”，而不是追求高一致性，因为底层偏好具有主观性，且大量结果属于难分高下的情况；MT-Bench 用于验证大语言模型评审的标准正是如此，即与人类的一致程度达到人类彼此之间所能达到的水平（Zheng 等，2023年）。最后，由于同一形式内的比较并不相互独立（每一对模型都基于相同的来源接受评审），我们使用 Bradley–Terry 分数报告排名不确定性，并给出 95% 源文本项目 自助法置信区间（1,000 次重复；在每种形式内独立地对完整来源进行重新抽样）。每次重复都会重新执行完整的拟合、中心化、缩放和平均流程；我们报告百分位区间。

5 结果

5.1 跨形式排行榜

在按形式均等标准化的 Bradley–Terry 宏观分数上，相邻模型的边际 95% 源自助法区间互不重叠（图 3，表 1）；部分按形式划分的中游模型配对区间有所重叠（附录 A）。GPT-5.5 排名第一，其次是 Claude Opus 4-6；这两者在每种单独形式中均位居第一和第二。DeepSeek V4-Pro 在宏观指标上稳居第三，领先于名次紧密的中游模型（Gemini 3.1 Pro、Grok 4.3、Gemma 4 26B）和排名靠后的两款模型（Sarvam 105B、Qwen3.5 27B）。各形式的排名分歧主要出现在中游：DeepSeek、Gemini 和 Grok 在音频小说、短篇小说和散文中的名次交替变化（按形式划分的排行榜见附录 A，表 4）；标准化宏观分数将各形式特有的名次变化汇总为一个跨形式排名。维度评判器提供了易于解读的补充视角：其平均分显示，领先的是 GPT-5.5 和 Claude Opus 4-6，排名靠后的是两款得分最低的模型；这与整体排名在两端的结果一致。自然度（J1）的分布范围比构成（J2）或声音（J3）更广。

5.2 门控至关重要

忠实度是基准测试的一项关键指标，而且在不同形式中的表现各异。音频小说的输出级门控失败率为 29.5%，短篇小说为 9.1%，散文为 17.1%；按形式等权计算的宏观指标为 18.6%，语料加权率为 22.4%（16,400 份润色稿中有 3,677 份）。各模型的失败率差异很大，而且同一模型在不同形式中的失败率也有所不同（附录 D，图 7）：在近五分之三音频小说草稿中出现偏离的模型（Gemini 3.1 Pro，57.5%），在其散文中的失败率要低得多（29.4%），在其短篇小说中的失败率则更低（14.6%）；因此，忠实度既取决于模型，也取决于形式。在润色任务中，忠实但平淡的改写文本优于流畅却改变源文本的改写，因此，任何对门控通过与门控失败的输出进行的比较，都要先依据忠实度作出判断，然后才考虑散文质量。图 4 展示了不同模型在忠实度与散文质量之间的权衡。排行榜标题结合了各模型的通过率，以及其忠实输出的散文质量：GPT-5.5 和 Claude Opus 4-6 在这两个维度上都表现出色，而 Gemini 3.1 Pro 的通过率较低，但纯文案质量较高。这确实影响了排名：如果只比较双方改写文本都通过门控的情况，Gemini 在忠实输出中升至第三位，尽管忠实度使其排行榜标题排名降至第四位（附录 B）。

门控还会为每个失败标注内容违规标签；由于分类体系受形式制约，主要失败模式也同样如此（表 2）。音频小说的失败主要表现为事实、对象或时间线偏移（失败输出的40.9%）；短篇故事的失败主要表现为关系、身份或能动性转变（39.2%）；而文章的失败则主要表现为论断变更、询问或限定条件变更（65.1%）。叙事体裁最常通过改变发生了什么或由谁实施来背离源文本，而文章则最常通过改变论证内容来背离源文本。我们按形式分别保留这些类别，因为它们的失败类型在不同形式之间并不对应。附录 I （表 18）给出了这种偏移的人工确认最小对示例，其中源文本和改写文本片段并列展示。

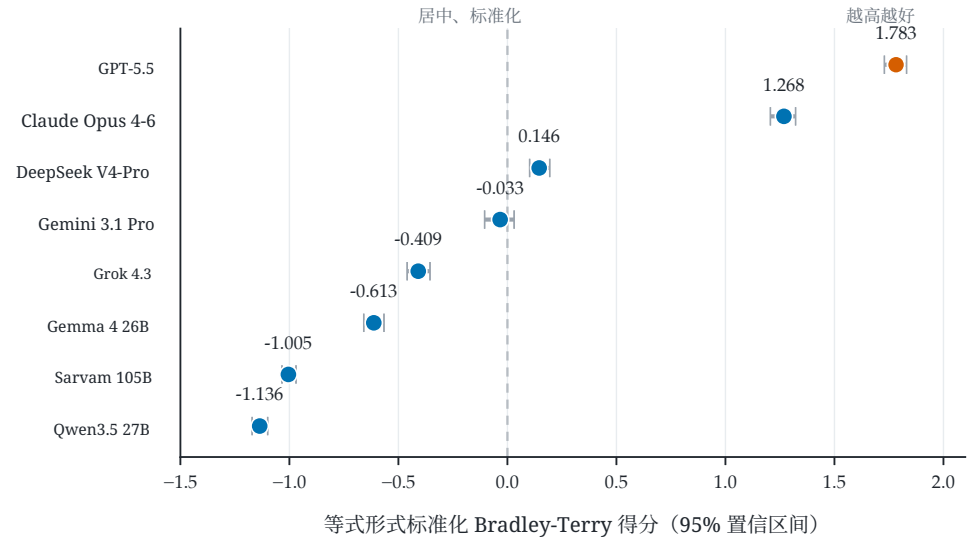

图3：涵盖2,050个源文本和八个精炼器的等形式标准化 Bradley–Terry 排行榜。各点表示三个经过离散程度标准化的形式特定系数的平均值；柱状图表示完整流程计算得出的源文本—项目级自助法95%区间。越高越好。

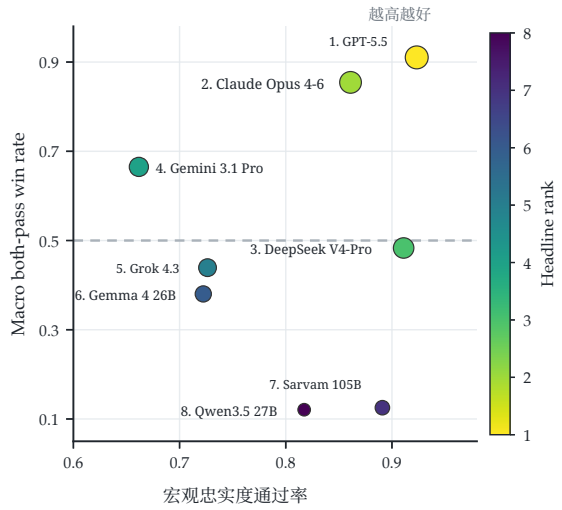

图4：忠实度—文案前沿：各模型的忠实度门槛通过率（横轴）与在双方改写文本均通过门槛的比较中所对应的散文质量（纵轴）。GPT-5.5 和 Claude Opus 4-6 在两个轴上均处于领先位置；Gemini 3.1 Pro 以较低的通过率实现了较高的纯文案质量，而 DeepSeek V4-Pro 和 Sarvam 105B 虽然保持忠实度，但散文质量较弱。

Form	主要失败类别	占比
音频小说	事实、对象或时间线偏移	40.9%
短篇小说	关系、身份或能动性	39.2%
文学随笔	主张、探究或限定	65.1%

表2：按散文形式划分的主要忠实性脆弱点；占比为该形式失败输出中的占比。税-分类法是形式条件化的；完整分布见附录表10。

5.3 稳健性与校准

长度匹配保留了三种形式中的前两个模型，以及短篇小说和文章中的完整排名；只有音频小说中排名较低的模型发生变化，因为匹配还改变了这些模型所遇到的对手（附录 B，表 5，图 6，以及表6）。排名对于生成器家族亲缘关系同样稳健：四个精炼器在留出生成器池中各自拥有一个同级项；然而，在移除所有同级项匹配的比较后重新计算每种形式的排行榜，三种形式中的顺序均完全保持不变（斯皮尔曼 $\rho = 1.0$；附录 E）。

消融实验。 两项消融实验检验了这一设计（附录 C）。对于 门控，我们对门控已作出判断或已剔除的 22,214 个配对进行了文本评判，因此评审器覆盖全部 57,400 个同源配对；并根据门控唯一的严重程度信号——违规标签数量——对其进行评估。应用已发布阈值后，结果与已发布排行榜完全一致；阈值设为两个标签时，排序不变；阈值设为三个标签以及完全不设门控时，只有 Gemini 3.1 Pro 和 DeepSeek V4-Pro 互换第三、第四名（斯皮尔曼相关系数为 0.976）。没有门控，就无法将改写稿与其草稿进行核对；而从中获益的模型，恰恰是门控最常判定其失败的模型。对于 绝对基准，我们将每份草稿分别与其八份润色稿进行比较（共 16,400 个配对）：六个精炼器在所有形式中都优于各自的草稿；Sarvam 105B 与草稿难分高下；Qwen3.5 27B 在音频小说形式下则不如草稿。

人工校准表明，评审器的表现处于读者一致性的范围内，且这两个层面的表现各不相同。客观门控在所有形式中都可靠：在一项分层抽取的 600 例审计中（每种形式 200 例），其 555 例与人工标注一致（0.925, [0.901，0.943]）；按形式统计，其一致性分别为：音频 0.930、故事 0.935、散文 0.910。各形式的误判失败率和误通过率都很低；附录 G报告了这些指标及其条件定义，还列出了按精炼器统计的一致性（0.873 至 0.970，各区间均有重叠），以及按书面失败类别统计的一致性（0.879 至 0.973）。主观偏好对形式更为敏感。在整体任务中，Headline 评审器与个别读者在 540 项标注中的 284 项上达成一致（0.526）；相比之下，人与人之间的两两一致性为 0.472（189/400）。评审器与留出读者多数意见的一致性为 0.524（33/63），与个别读者和其余读者的留一法多数意见一致时的 0.529（83/157）相当。正如主观的、难以抉择的创意散文偏好研究所预期的那样，读者一致性本身较低（Chhun 等人，2022年；Feuer 等人，2025年）；无论是人类还是自动评审器，在散文形式下的一致性最低，且评审器的一致性与人类水平相当（附录 G，表 12）。因此，排行榜的可靠性建立在整体排名的稳定性上，而非逐项一致性。

GPT-5.5 既是评审器，也是参赛者。GPT-5.5 既是主要评审器，也是排名第一的模型，因此我们检查每个阶段是否存在自我偏好。在门控阶段，人类标注员推翻了门控放行的 37 条经审计 GPT-5.5 输出中的 2 条（0.054，审计总体为 0.076）；在重新门控的音频样本中，对于 GPT-5.5 的输出，我们的门控通过率为 0.856，而独立家庭评审员的通过率为 0.981。因此，对于顶级模型，主要门控更严格（附录 F）。在偏好评估中，两个独立模型家族——Claude Opus 4-6 和 Kimi-K3——对涉及 GPT-5.5、通过门控的音频小说比较重新评审，其判断与主要评审器分别有 0.902 和 0.866 的一致性；无论基于哪一方重新拟合，完整排名顺序都不变，GPT-5.5 仍居首（斯皮尔曼 $\rho = 1.0$）。540 条人类偏好标签不涉及评审器，也同样将 GPT-5.5 排在首位；该样本旨在衡量一致性，规模太小，无法对其他模型排名。即使完全不设门控，GPT-5.5 仍排名第一。书面形式的评估依赖单一主要评审器，我们在局限性部分对此作了说明。

仿译（原文写法）	≈ 英语习语	自然印地语
声音 ठंडी	“冷淡的声音”	生硬的声音
मुस्कसेन到眼睛没有传到	“笑容没有传达到眼睛”	做作的微ऽस्कसेन笑
一阵ा寒意顺着 ा背发凉 गई	“一阵寒意顺着脊背窜过”	起鸡 皮疙瘩了
空气中弥漫着 紧张感	“空气中弥漫着紧张气氛”	氛围 m̃ 紧张

表 3：代表性仿译词（真实模型输出）及母语写作者会采用的自然印地语。三种形式均使用相同的 120 项仿译列表进行评分。

6 印地语散文的诊断特征

除了排名之外，RASA-Bench 还揭示了一种通用、与语言无关的指标所无法捕捉的印地语散文特征。我们报告了两项基于同一审核词库、经人工审核的诊断指标：一项是程式化短语负担（FPB），即一份经审核的印地语惯用表达词汇表，涵盖对话标签、身体反应描写、氛围与沉默填充语，以及文章修辞模板，其灵感来自低质文本取证研究（Shaib等人，2025年；Chakrabarty等人，2025年；Paech等人，2026年）；另一项是仿译／翻译腔诊断指标，用于检测将英语习语逐字译成印地语的情况。这些程式化表达是模型相较于人类写作过度使用的惯用结构（Paech等人，2026年），因此，负担较高表明文本类似于人工智能“低质量文本”，即便语法仍然正确。仿译是多语言生成中有据可查的现象，源语言的痕迹会在目标语言中显现（Riley等人，2020；Koppel和Ordan，2011；Rabinovich和 Wintner，2015；Rahman等人，2026年）；我们据此改编出一项经审核的印地语诊断指标。120 种仿译表面形式由全部三种形式共用，而程式化表达词汇表则包含 272 种音频小说表面形式和 274 种书面形式表面形式；这两个层面的指标均为描述性指标，完整词汇表见附录 I。

这种特征在音频小说中最为显著，在各种书面形式中则逐渐减弱（图 5；各形式的完整数值见附录表 16）：经审阅的措辞在故事中出现得较少，在散文中出现得最少，但这并不能证明其他形式特有的程式化语言不存在。在同一种形式中，程式化表达的负担与质量并不相符；若要说有什么关系，表现力更强、戏剧性更足的模型反而使用得更多。在音频小说中，每千词的程式化表达负担，顶级模型 GPT-5.5 为 3.39，而排名第二的

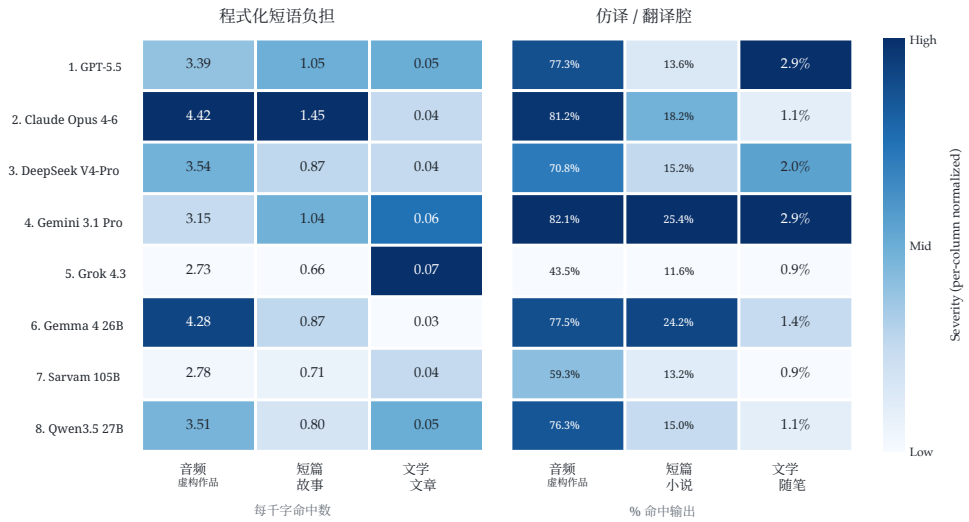

图 5：按排行榜排名排列的各模型及形式的印地语诊断特征：程式化短语负担和仿译出现率。这一特征在音频小说中最为显著，在短篇小说和散文中逐渐减弱；同一形式内的负担与质量并不相关，质量最高的模型反而负担最重。

Claude Opus 4-6 为 4.42，Grok 4.3 则仅为 2.73；Gemini 的音频输出中有 82.1% 出现仿译，Grok 的这一比例为 43.5%，因此负担较低并不意味着润色稿更好。以音频小说为参照，经审阅的同一词汇表在人类语料库中每千词仅有 0.17 次不同词项的命中，远低于所有模型的音频负担。仿译最容易被母语者听感察觉（表 3）：英文短语“笑容没有传达到眼睛”被逐字译成印地语后，读起来像机器生成；印地语母语写作者则会使用一个地道的习语。表 3 列出了四组这样的表层表达，每组均取自真实的模型输出，并配上印地语母语写作者会采用的地道表达。文字泄漏很少见：八个模型的平均拉丁字符比例均低于 0.3%。附录 I （表 17）列出了这些程式化短语族的代表性例子，并突出显示了反复出现的短语。

7 结论

优化草稿与创作草稿的要求不同，而这正是生产环境对模型在各种形式下提出的要求。RASA-Bench 针对三种印地语散文进行测评，方法是判断哪种改写文本更出色，并且只认可忠于源文本的改写文本。结果显示，顶尖表现稳定，中游表现因形式而异，而忠实度信号至关重要且受形式条件化：改写文本必须保留什么，以及保留起来有多难，取决于它是在讲述故事还是在提出论点。同样的门控偏好设计也可以扩展到其他印地语形式和其他印度语言。给精炼器的启示很明确：优化，而非另起炉灶。

局限性

这些结果受到若干边界条件的限制。底层偏好具有主观性，且难以判定的情况很多：读者之间的一致性仅属一般，在散文上尤其低（Fleiss $\kappa = 0.012$ 针对三位读者评定的20组共享散文比较；散文排行榜本身则基于9,800组配对数据拟合），因此任何单个标签都带有噪声，排名的可靠性来自大量汇总比较（Chhun 等人，2022；Feuer 等人，2025）。三个层级均由同一个模型族（GPT-5.5）担任主要评审器（Krumdick 等人，2025）；§5 会在每个阶段对此进行检查。我们的独立模型族检查是有意针对特定范围进行的，因此覆盖面相应有限：两个外部模型族重新评审了顶级模型在音频小说比较中通过门槛的结果，排名并未改变；但该检查未扩展至完整排行榜，也未涵盖短篇故事和散文体裁，这些部分仍依赖单一评审器；完整的跨模型族、全体裁审查仍有待未来开展。忠实度门槛通过一个包含600个案例的人工样本进行了验证，足以表明它能在不同体裁和模型间跟踪人类判断，但尚未达到对每个模型进行穷尽式校准的程度；而用于筛查每份草稿是否符合 G 的构建筛选器本身未经人工审查。维度评判器未经校准，其绝对分数也无法跨体裁进行比较。

在不同体裁间进行比较。RASA-Bench 对润色稿进行相互排名，同时将每份润色稿与其各自的草稿进行门控评估 S （§3）。由于短篇故事和散文的源文本公开可用，它们可能出现在模型预训练数据中，存在污染风险，而私有音频语料库则不受此风险影响；尽管如此，这项任务评估的仍是对一份单独生成的草稿所做的润色 S，并且门槛会筛除润色器凭借对源文本的记忆而添加的内容。最后，该基准涵盖三种印地语散文形式，无法覆盖其他体裁、其他印度语言，也不支持经过校准的逐维度评分。

伦理声明

RASA-Bench 使用两类源文本。音频小说来源是一个由专业编辑过的剧集构成的私有语料库，仅用于启动数据构建和为忠实性过滤器提供依据；这些内容绝不会发布或展示，也不会用于排行榜评审（构建模型只会在大纲提取和门控阶段读取这些内容）。短篇故事和散文来源取自印地语维基文库、BHAAV语料库（Kumar 等人，2019）以及 Hindi Discourse（Dhanwal 等人，2020）；版权与发布按条目处理，部分源材料受到限制，因此发布决定与我们为本研究处理文本所获的许可彼此独立。所有发布和展示的文本，包括本文中的每个例子，均由模型生成，我们不会披露可识别私有语料库的详细信息。人工验证由印地语母语使用者完成；他们是本组织支付报酬的成员，均已知情同意，并在不知道模型身份的情况下进行评审：音频小说偏好评审有六位读者，书面偏好评审组各有三位成员，第二轮门控审计有三位标注员（附录 G）；个别标注、读者身份及人口统计信息均不予公开。三种形式的已发布草稿与优化结果、提示词、配置、评判器栈、评估代码和汇总结果，均已通过内容、隐私及许可审查。只有人工黄金源文本是私有的：它们仅用于构建，不会分发，也不是复现任何结果所必需的。音频小说在按形式均等计算的总指标中占三分之一；其余三分之二来自公开来源。评估除了使用大语言模型评审来处理主观任务，也采用人工验证集，并透明地报告一致性水平。我们认为，这项工作总体影响积极：改进工具和评估方法，以优化印地语散文，能够服务于使用这种代表性不足语言的广大受众，同时让人工编辑参与其中。

参考文献

Anthropic. 2025年。 Claude Sonnet 4.5 系统卡。

Anthropic. 2026年。 Claude Opus 4.6 系统卡。

Ralph Allan Bradley 和 Milton E. Terry. 1952年。 不完全区组设计的秩分析：成对比较法。 *Biometrika*，39(3–4)：324–345。

Tuhin Chakrabarty、Philippe Laban 和 Chien-Sheng Wu。2025年。从 AI-slop 到 AI-polish? 通过基于编辑的写作奖励和测试时计算来对齐语言模型。*arXiv* 预印本 *arXiv:2504.07532*。Hong Chen、Duc Minh Vo、Hiroya Takamura、Yusuke Miyao 和 Hideki Nakayama。2022年。StoryER：通过排名、评分和推理自动评估故事。载于*2022*年自然语言处理实证方法会议论文集，第1739–1753页。Cyril Chhun、Pierre Colombo、Fabian M. Suchanek 和 Chloé Clavel。2022年。人工标准与自动指标：故事生成评估基准。载于第*29*届国际计算语言学会议论文集，第5794–5836页。国际计算语言学委员会。Wei-Lin Chiang、Lianmin Zheng、Ying Sheng、Anastasios Nikolas Angelopoulos、Tianle Li、Dacheng Li、Banghua Zhu、Hao Zhang、Michael I. Jordan、Joseph E. Gonzalez 和 Ion Stoica。2024年。Chatbot Arena：一个依据人类偏好评估大语言模型的开放平台。载于第*41*届国际机器学习会议（*ICML*）。DeepSeek-AI。2026年。DeepSeek-V4：迈向高效的百万词元上下文智能。*arXiv* 预印本 *arXiv:2606.19348*。Swapnil Dhanwal、Hritwik Dutta、Hitesh Nankani、Nilay Shrivastava、Yaman Kumar、Junyi Jessy Li、Debanjan Mahata、Rakesh Gosangi、Haimin Zhang、Rajiv Ratn Shah 和 Amanda Stent。2020年。一个印地语故事话语模式标注数据集。载于第十二届语言资源与评测会议论文集，第1191–1196页。Daniel Fein、Sebastian Russo、Violet Xiang、Kabir Jolly、Rafael Rafailov 和 Nick Haber。2026年。LitBench：用于可靠评估创意写作的基准测试与数据集。载于第*19*届欧洲计算语言学协会分会会议论文集（第一卷：长篇论文），第7740–7755页。

Benjamin Feuer、Chiung-Yi Tseng、Astitwa Sarthak Lathe、Oussama Elachqar 和 John P. Dickerson。2025年。当评判沦为噪声：大语言模型评审基准的设计失误如何悄然损害效度。*arXiv* 预印本 *arXiv:2509.20293*。

Jay Gala、Thanmay Jayakumar、Jaavid Aktar Husain、Aswanth Kumar M、Mohammed Safi UrRahman Khan、Diptesh Kanojia、Ratish Puduppully、Mitesh M. Khapra、Raj Dabre、Rudra Murthy 和 Anoop Kunchukuttan。2024。Airavata：用于印地语的指令微调大语言模型。*arXiv* 预印本 *arXiv:2401.15006*。

Gemma Team。2026年。Gemma 4 技术报告。*arXiv* 预印本 *arXiv:2607.02770*。

GLM-5-Team。2026年。GLM-5：从氛围编程到智能体工程。*arXiv* 预印本 *arXiv:2602.15763*。

Google DeepMind。2025年。Gemini 3 Flash 模型卡。

Google DeepMind。2026a。Gemini 3.1 Flash-Lite 模型卡。

Google DeepMind。2026年。Gemini 3.1 Pro 模型卡。

Anusha Kamath、Kanishk Singla、Rakesh Paul、Ravi- raj Bhuminand Joshi、Utkarsh Vaidya、Sanjay Singh Chauhan 和 Niranjan Wartikar。2025年。对印地语大语言模型进行基准测试：一套全新的数据集及比较分析。载于《首届印度语言以人为本人工智能基准测试、协调、标注与标准化研讨会论文集（*BHASHA 2025*）》，第52–68页。Taaha Kazi、Aditya Sanjiv Kanade、Ivory Yang、Mohammad Saifullah、Vasu Sharma、Bhavesh Pant、Kunal Bhatia、Abhinav Prakash、Abdur Rahman、Varun Agarwal、Sarjita Jain、Viveca Pavate 和 Nishita Maithul。2026a。改编，而非翻译：长篇跨文化故事本地化的架构与评估框架。载于《*2026*年自然语言处理实证方法会议论文集：产业论坛》，第2093–2114页，匈牙利布达佩斯。计算语言学协会。Taaha Kazi、Vasu Sharma、Mohammad Saifullah 和 Abdur Rahman。2026b。PAUSE：用于长篇文化故事改编的可编辑策略工件。预印本，arXiv:2608.28633。Seungone Kim、Juyoung Suk、Shayne Longpre、Bill Yuchen Lin、Jamin Shin、Sean Welleck、Graham Neubig、Moontae Lee、Kyungjae Lee 和 Minjoon Seo。2024年。Prometheus 2：一种专门用于评估其他语言模型的开源语言模型。载于《*2024*年自然语言处理实证方法会议论文集》，第4334–4353页。Kimi Team。2025年。Kimi K2：开放智能体智能。*arXiv*预印本 *arXiv:2507.20534*。

Kimi Team。2026年。Kimi K3：开放前沿智能。*arXiv*预印本 *arXiv:2607.24653*。

Moshe Koppel 和 Noam Ordan。2011年。翻译腔及其方言。载于第*49*届计算语言学协会年会：人类语言技术论文集，第1318–1326页。计算语言学协会。

Michael Krumdick、Charles Lovering、Varshini Reddy、Seth Ebner 和 Chris Tanner。2025年。没有免费的标签：缺乏人类基准的大语言模型裁判的局限性。*arXiv* 预印本 *arXiv:2503.05061*。

Aman Kumar、Himani Shrotriya、Prachi Sahu、Amogh Mishra、Raj Dabre、Ratish Puduppully、Anoop Kunchukuttan、Mitesh M. Khapra 和 Pratyush Kumar。2022年。IndicNLG 基准：面向印度语言中多样化自然语言生成任务的多语言数据集。载于*2022*年自然语言处理经验方法会议论文集，第5363–5394页。

Yaman Kumar、Debanjan Mahata、Sagar Aggarwal、Anmol Chugh、RajatMaheshwari 和 RajivRatn Shah。2019 年。BHAAV：用于印地语故事情感分析的文本语料库。*arXiv* 预印本 *arXiv:1910.04073*。

Philippe Laban、Tobias Schnabel、Paul N. Bennett 和 Marti A. Hearst。2022年。SummaC：重新审视用于摘要不一致性检测的基于 NLI 的模型。计算语言学协会汇刊，10：163–177。Yang Liu、Dan Iter、Yichong Xu、Shuohang Wang、Ruochen Xu 和 Chenguang Zhu。2023年。G-Eval：使用 GPT-4 并提升人类对齐效果的自然语言生成评估。载于*2023*年自然语言处理实证方法会议论文集，第 2511–2522 页。Aman Madaan、Niket Tandon、Prakhar Gupta、Skyler Hallinan、Luyu Gao、Sarah Wiegreffe、Uri Alon、Nouha Dziri、Shrimai Prabhumoye、Yiming Yang、Shashank Gupta、Bodhisattwa Prasad Majumder、Katherine Hermann、Sean Welleck、Amir Yazdanbakhsh 和 Peter Clark。2023年。Self-Refine：通过自反馈进行迭代式优化。神经信息处理系统进展，36：46534–46594。Logan Mann、Abdur Rahman、Mohammad Saifullah、Taaha Kazi 和 Vasu Sharma。2026年。Narrative-WorldBench：用于长时程协作创作音频剧情的前沿饱和基准测试和潜在世界模型。载于*Culture x AI*：将 *AI* 作为文化技术进行评估（*ICML 2026*）。MiniMax。2026年。MiniMax M2.7：自我进化的早期迹象。

Mistral AI。2026年。MistralMedium 3.5。

OpenAI。2026a。GPT-5.5 系统卡。

OpenAI。2026b。推出 GPT-5.4 mini 和 n ano.

Sam Paech、Allen Roush、Judah Goldfeder 和 Ravid Shwartz-Ziv。2026 年。Antislop：用于识别和消除语言模型中重复模式的综合框架。载于国际学习表征会议。

Chau Minh Pham、Zichao Wang、Puneet Mathur、Alexa Siu、Akriti Jain、Aparna Garimella、Ananya B. Sai、Nedim Lipka、Mohit Iyyer 和 Varun Manjunatha。2026年。AnalystBench：利用网络挖掘的多模态任务对专业长篇报告生成进行基准测试。载于计算语言学协会发现：*ACL 2026*，第 23894–23926 页，美国加利福尼亚州圣迭戈。计算语言学协会。

Qwen Team。2026年。Qwen3.5：迈向原生多模态智能体。

Ella Rabinovich 和 Shuly Wintner。2015年。无监督识别翻译腔。《计算语言学协会会刊》，3：419–432。

VipulRaheja、Dhruv Kumar、Ryan Koo 和 Dongyeop Kang。2023年。CoEdIT：通过任务特定的指令微调进行文本编辑。载于计算语言学协会论文集：*EMNLP 2023*，第5274–5291页。

Abdur Rahman、Vasu Sharma 和 Mohammad Saiful- lah。2026年。为什么仿译词在印地语模型输出中看起来很少见：以人为锚定的测量研究。载于跨文化语言开放推理研讨会论文集（*ORACLE*），第316–328页，匈牙利布达佩斯。计算语言学协会。

Parker Riley、Isaac Caswell、Markus Freitag 和 David Grangier。2020年。翻译腔作为“多语言”神经机器翻译中的一种语言。载于计算语言学协会第*58*届年会论文集（*ACL*），第7737–7746页。

Ananya Sai B、Tanay Dixit、Vignesh Nagarajan、Anoop Kunchukuttan、Pratyush Kumar、Mitesh M. Khapra 和 Raj Dabre。2023年。IndicMT Eval：用于元评估印度语言机器翻译指标的数据集。载于计算语言学协会第*61*届年会论文集（第*1*卷：长篇论文），第14210–14228页。计算语言学协会。

Mohammad Saifullah、Thomas Kornmaier、Taaha Kazi、Vasu Sharma、Aditya Sanjiv Kanade 和 Aanand KumarYadav。2026年。叙事世界模型：以叙事学为基础的长篇虚构作品作者记忆。预印本，arXiv:2607.05577。

Sarvam Foundation Models Team。2026年。介绍 Sarvam 的自主模型。

Chantal Shaib、Tuhin Chakrabarty、Diego Garcia-Olano 和 Byron C. Wallace。2025年。衡量文本中的人工智能“低质量人工智能文本”。*arXiv* 预印本 *arXiv:2509.19163*。

Lei Shu、Liangchen Luo、Jayakumar Hoskere、Yun Zhu、Canoee Liu、Simon Tong、Jindong Chen 和 Lei Meng。2023年。RewriteLM：一种用于文本改写的指令微调大语言模型。*arXiv* 预印本 *arXiv:2305.15685*。Harman Singh、Nitish Gupta、Shikhar Bharadwaj、Di- nesh Tewari 和 Partha Talukdar。2024。IndicGen-Bench：用于评估大语言模型在印度语言上的生成能力的多语言基准。在计算语言学协会第*62*届年会论文集。

Maria Teleki、Vedangi Bengali、Xiangjue Dong、Sai Te-jas Janjur、Haoran Liu、Tian Liu、Cong Wang、Ting Liu、Yin Zhang、Frank Shipman 和 James Caverlee。2025年。大语言模型用于故事生成的综述。载于计算语言学协会论文集：*EMNLP 2025*，第13954–13966页。计算语言学协会。

Sshubam Verma、Mohammed Safi Ur Rahman Khan、Vishwajeet Kumar、Rudra Murthy 和 Jaydeep Sen。2025年。MILU：印度语系多任务理解基准测试。载于*2025*年美洲分会计算语言学协会会议论文集：人类语言技术（第*1*卷：长篇论文），第10076–10132页。计算语言学协会。

Peiyi Wang、Lei Li、Liang Chen、Zefan Cai、Dawei Zhu、Binghuai Lin、Yunbo Cao、Lingpeng Kong、Qi Liu、Tianyu Liu 和 Zhifang Sui。2024a。大型语言模型并非公平评估者。载于计算语言学协会第*62*届年会论文集（第*1*卷：长篇论文），第9440–9450页，泰国曼谷。计算语言学协会。

Qianyue Wang、Jinwu Hu、Zhengping Li、Yufeng Wang、Daiyuan Li、Yu Hu 和 Mingkui Tan。2024b。使用动态分层大纲与记忆增强生成长篇故事。*arXiv* 预印本 *arXiv:2412.13575*。

Yuning Wu、Jiahao Mei、Ming Yan、Chenliang Li、Shaopeng Lai、Yuran Ren、Zijian Wang、Ji Zhang、Mengyue Wu、Qin Jin 和 Fei Huang。2025年。WritingBench：一个全面的生成式写作基准测试。*arXiv* 预印本 *arXiv:2503.05244*。

xAI。2026年。亚马逊 Bedrock 上的 Grok。

Dingyi Yang 和 Qin Jin。2024。什么造就了一个好故事，我们又该如何衡量？关于故事评估的综合综述。*arXiv* 预印本 *arXiv:2408.14622*。

Shuangshuang Ying、Yunwen Li、Xingwei Qu、Xin Li、Sheng Jin、Minghao Liu、Zhoufutu Wen、Xeron Du、Tianyu Zheng、Yichi Zhang、Letian Ni、Yuyang Cheng、Qiguang Chen、Jingzhe Ding、ShengdaLong、Wangchunshu Zhou、Jiazhan Feng、Wanjun Zhong、Libo Qin及其他4位作者。2025年。超越正确性：评估跨文化的主观写作偏好。*arXiv*预印本 *arXiv:2510.14616*。

Lianmin Zheng、Wei-Lin Chiang、Ying Sheng、Siyuan Zhuang、Zhanghao Wu、Yonghao Zhuang、Zi Lin、Zhuohan Li、Dacheng Li、Eric P. Xing、Hao Zhang、Joseph E. Gonzalez及Ion Stoica。2023年。使用 MT-Bench 和 Chatbot Arena 评测大语言模型裁判。*arXiv*预印本 *arXiv:2306.05685*。

A 按形式划分的主榜排行榜

表 4 给出了形式特定的主榜表。Bradley–Terry模型分别在每种形式内进行拟合；表中展示原始中心化点估计，发布版本提供其95%的源级自助法区间。标准化宏观区间见表 1。胜率是忠实度门控经验胜率，失败率是输出级门控失败率，J1–J3则是形式条件化的1–5分诊断均值。表 1 在求三个标准化值的平均数之前，先将每种形式的中心化系数向量除以其总体标准差；此处保留原始的形式特定系数。表格中部按不同形式重新排序（例如，Gemini 3.1 Pro在短篇小说中排名第三，但在散文中排名第四；DeepSeek V4-Pro在短篇小说中排名第四，但在音频小说和散文中排名第三）。

B 排行榜稳健性

位置偏差。 每个配对都会按两种呈现顺序分别进行评审，并折叠为单一的源级结果；这种交换折叠从设计上消除了呈现顺序偏差，任何两种顺序之间的不一致都会折叠为平局。

长度控制。 忠实的润色稿可能会扩展其草稿。近长度匹配使 GPT-5.5 和 Claude Opus 4-6 在每种形式中保持前两名，并保留完整的短篇故事和散文排名（表 5）。音频小说在前三名之外更为敏感，因此图 6分别报告其不受限、近长度和大间隔结果。

长度匹配保留了哪些配对。近长度匹配会改变模型遇到的对手。在音频小说中，这完全解释了 Gemini 的大幅变化，也部分解释了 Grok 的变化（表5 和表6）。Gemini 3.1 Pro 和 GPT-5.5 对音频草稿的扩写幅度最大（分别达到草稿长度的 1.661 倍和 1.598 倍；附录J），因此 Gemini 的 1,011 次近长度比较中，有 691 次是与 GPT-5.5 进行的；无论长度如何，Gemini 在这种配对中都会落败（近长度情况下为 0.054，整体为 0.055）。对于其他每个交手次数超过一次的对手，Gemini 的近长度胜率都不低于其整体胜率。保持其对各个对手的整体胜率不变，并按

A. 音频小说

排名	模型	BT	Win	失败率	J1	J2	J3
1	GPT-5.5	2.02	0.891	13.4%	3.95	4.10	3.87
2	Claude Opus 4-6	1.13	0.752	21.2%	3.76	4.12	3.83
3	DeepSeekV4-Pro	0.28	0.578	12.9%	3.60	3.99	3.66
4	Gemma 4 26B	-0.46	0.399	42.8%	3.64	3.97	3.66
5	Gemini 3.1 Pro	-0.57	0.373	57.5%	3.68	4.02	3.73
6	Sarvam 105B	-0.69	0.350	14.8%	3.27	3.95	3.52
7	Grok 4.3	-0.71	0.341	45.8%	3.70	4.06	3.70
8	Qwen3.5 27B	-1.01	0.277	27.6%	3.01	3.89	3.54

B. 短篇小说

排名	模型	BT	Win	Fail	J1	J2	J3
1	GPT-5.5	2.03	0.877	3.0%	3.93	3.46	3.41
2	Claude Opus 4-6	1.59	0.814	8.2%	3.82	3.50	3.41
3	Gemini3.1Pro	0.36	0.587	14.6%	3.72	3.35	3.29
4	DeepSeekV4-Pro	-0.16	0.476	3.8%	3.40	3.33	3.24
5	Grok 4.3	-0.26	0.456	12.0%	3.41	3.34	3.26
6	Gemma 4 26B	-0.72	0.356	15.4%	3.39	3.26	3.19
7	Qwen3.5 27B	-1.38	0.223	7.2%	3.11	3.27	3.19
8	Sarvam 105B	-1.46	0.209	8.8%	3.13	3.24	3.16

C. 文学散文

排名	模型	BT	Win	Fail	J1	J2	J3
1	GPT-5.5	1.80	0.852	6.6%	3.97	3.11	3.49
2	Claude Opus 4-6	1.48	0.803	12.3%	3.89	3.04	3.39
3	DeepSeekV4-Pro	0.32	0.578	10.0%	3.64	2.96	3.31
4	Gemini3.1Pro	0.18	0.545	29.4%	3.84	3.00	3.32
5	Grok 4.3	-0.33	0.432	24.3%	3.73	2.98	3.33
6	Gemma 4 26B	-0.86	0.318	25.1%	3.63	2.95	3.28
7	Sarvam 105B	-1.22	0.245	9.1%	3.42	2.93	3.24
8	Qwen3.5 27B	-1.37	0.216	20.0%	3.40	2.94	3.28

表4：特定形式的主要排行榜。请参见表1 ，其中给出了根据这些原始特定形式拟合得出的等形式标准化宏观指标。

近长度混合进行重新加权后，胜率为 0.167，低于观测到的 0.217；因此，对手变化解释了胜率从 0.373 全部下降至该水平的原因。Grok 4.3 是最短的音频改写器（长度为草稿的 0.775 倍），没有与 GPT-5.5 进行近长度比较，其 633 次比较中有 339 次是与 Sarvam 105B 进行的。同样的重新加权得出 0.433，因此，配对构成解释了其胜率从 0.341 升至 0.563 的约五分之二；其余部分则来自配对内胜率的提高，主要体现在与 Sarvam 105B 和 Qwen3.5 27B 的比较中。在书面形式中，这两个模型的长度都保持在草稿的 0.84 至 1.12 倍之间；在使用同一评审器时，它们的近长度胜率与整体胜率相差不超过 0.03，这表明差异与模型在音频上的长度行为有关。这种分解仅为描述性分析。

双通过纯散文诊断。将范围限制在两篇改写文本均通过门控的比较中，可以得出忠实输出的散文质量排序（表7）。这是一项排除门控失败输出的诊断；附录C中的无门控拟合也会对散文进行评审。GPT-5.5 和 Claude

模型	音频 WR (n)	故事 WR (n)	短文 WR (n)
GPT-5.5	**92.6% (1,056)**	**87.8% (1,824)**	**84.4% (1,664)**
Claude Opus 4-6	**83.5% (3,751)**	**85.1% (2,407)**	**83.1% (2,060)**
DeepSeek V4-Pro	64.4% (3,926)	51.0% (2,700)	59.9% (2,072)
Gemini 3.1 Pro	21.7% (1,011)	60.6% (2,034)	56.1% (1,597)
Grok 4.3	56.3% (633)	45.7% (1,431)	45.8% (1,494)
Gemma 4 26B	39.8% (3,614)	37.8% (2,624)	31.8% (2,059)
Sarvam 105B	30.2% (3,345)	23.1% (2,592)	23.8% (2,040)
Qwen3.5 27B	24.9% (3,956)	23.6% (2,646)	21.1%（2,026）

表5：在输出词数比为 $[0.85, 1.15]$ 的已判定比较中，近邻长度的实证胜率。平局计为0.5；排除双方均失败和无效结果。 n 是模型比较次数。这是描述性敏感性分析，而非长度调整后的排行榜。

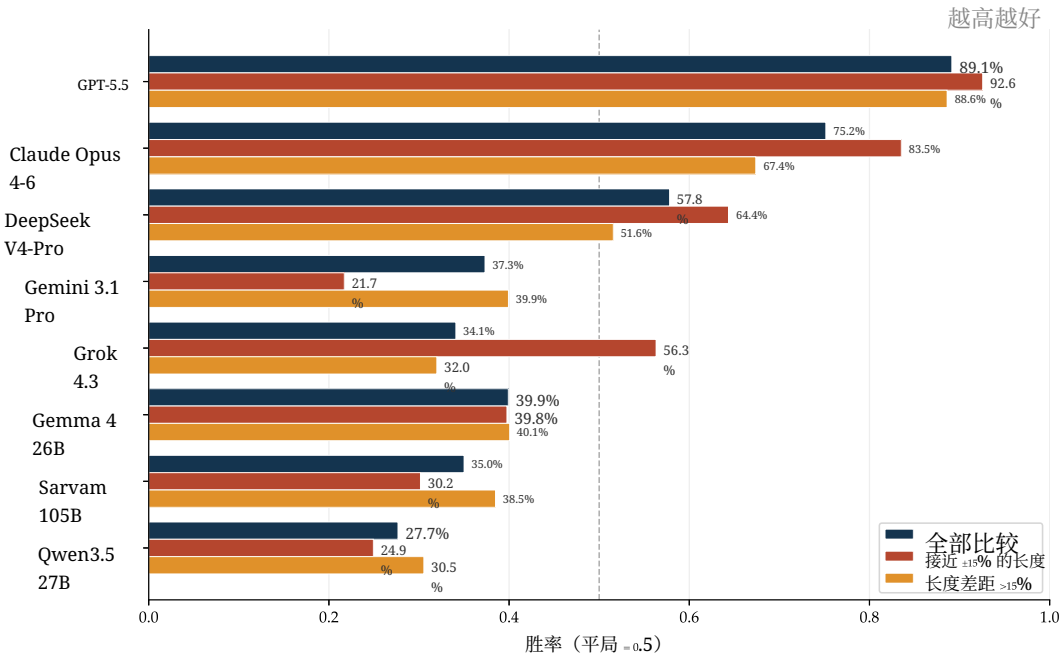

图6：音频小说长度控制敏感性：涵盖全部已判定比较、近邻长度比较，以及输出长度差距较大的比较的实证胜率。在近邻长度匹配下，排名前两位的模型保持不变，但部分较低排名发生变化。结合表 5可见，头部领先模型在全部三种形式下均保持稳定，并且完整的书面形式排名得以保留。

	Gemini 3.1 Pro		Grok 4.3	
对手	近邻 (n)	All	近邻 (n)	All
GPT-5.5	0.054 (691)	0.055	(0)	0.071
Claude Opus 4-6	0.192 (73)	0.138	0.103 (68)	0.123
DeepSeek V4-Pro	0.500（71）	0.407	0.338（145）	0.252
Gemini 3.1 Pro			0.000 (1)	0.440
Grok 4.3	1.000 (1)	0.560		
Gemma 4 26B	0.734 (77)	0.552	0.380 (25)	0.387
Sarvam 105B	0.893 (14)	0.467	0.724 (339)	0.548
Qwen3.5 27B	0.750(84)	0.506	0.827(55)	0.606
All	0.217(1,011)	0.373	0.563 (633)	0.341

表6：按对手划分，Gemini 3.1 Pro 和 Grok 4.3 在音频小说近长度比较（词数比为 $[0.85, 1.15]$）及所有已判定比较中的胜率。平局按半胜计算。

Opus 4-6 在每种形式中都稳居第一和第二，Gemini 3.1 Pro 则升至总体第三，这与其在保持忠实度时散文表现出色的情况一致。

Rk	模型	宏观指标	音频	故事	文章
1	GPT-5.5	0.910	0.972	0.888	0.871
2	Claude Opus 4-6	0.854	0.837	0.866	0.860
3	Gemini 3.1 Pro	0.665	0.663	0.648	0.684
4	DeepSeek V4-Pro	0.483	0.466	0.438	0.546
5	Grok 4.3	0.439	0.388	0.466	0.463
6	Gemma 4 26B	0.380	0.471	0.366	0.304
7	Sarvam 105B	0.125	0.113	0.152	0.111
8	Qwen3.5 27B	0.121	0.073	0.161	0.128

表7：两项均通过时的仅散文实证胜率（仅统计通过门控的比较），按形式等权计算的宏观指标及各形式的胜率。此处选取了一个子集，用于说明忠实输出中散文质量的贡献；这并非未经过门控的排行榜。

C 消融实验

两项消融实验都完全沿用已发表研究中的评审方式：使用固定的标题提示词，由 GPT-5.5 进行评审，将两种呈现顺序合并，平局计作半场胜利，并采用按形式拟合、居中和缩放后的标题综合得分等权平均，同时进行 1,000 次源文本条目自助法重复抽样。这两项实验在已发表的实验基础上新增了 38,614 个配对（两种呈现顺序共计 77,228 次评审）。

门控消融。已发布的运行只对两份改写文本都通过门控的配对进行散文评审。我们对其余 22,214 对（16,867 对音频小说、2,395 对短篇小说、2,952 对文章）进行了散文评审，因此评审器覆盖了全部 57,400 对同源配对；随后，我们根据违规标签数量调整门控阈值，这是唯一的严重程度信号：当输出包含至少 t 个违规标签时，该输出会在阈值 t 下被判定为失败。作为对照，$t = 1$ 精确复现了已发布的排行榜，且每个被排除和由门控判定的配对数量都完全一致。表格 8 显示，排名顺序在 $t = 2$时保持不变；而在 $t = 3$ 以及无门控时，只有 Gemini 3.1 Pro 和 DeepSeek V4-Pro 的第三、第四名互换（斯皮尔曼相关系数为 0.976；在无门控条件下，相邻模型在每次自助法重复抽样中仍能彼此区分）。无门控条件是我们自行设计的消融实验，仅用于诊断：散文评审器看不到草稿，因此没有门控就无法检查保真度，而排名上升的模型恰好是门控失败率最高的模型。

绝对基准：将草稿作为候选项。我们还将每份草稿与其八个优化结果逐一比较（共 16,400 个配对）。草稿本身没有门控判定，因为门控比较的是优化结果与其草稿，因此它只能纳入无门控拟合，作为第九个候选项（表 8，最后一列）。表 9 列出了每个精炼器对其对应草稿的直接胜率。六个精炼器在三种形式下的表现都优于草稿，胜率均达到 0.748 或更高。Sarvam 105B 在每种形式下都以微弱优势胜出，但在九个候选项的拟合中无法与草稿区分（$P = 0.834$，区间重叠）；而 Qwen3.5 27B 在音频小说这一形式下以及该拟合中的表现都不及草稿。因此，草稿确实是一个有竞争力的候选项。

模型	$t \geq 1$	$t \geq 2$	$t \geq 3$	无	+草稿
GPT-5.5	1.783	1.675	1.602	1.545	1.673
Claude Opus 4-6	1.268	1.278	1.283	1.271	1.397
DeepSeek V4-Pro	0.146	0.123	−0.007	−0.041	0.073
Gemini 3.1 Pro	−0.033	0.108	0.285	0.399	0.496
Grok 4.3	−0.409	−0.348	−0.212	−0.136	−0.023
Gemma 4 26B	−0.613	−0.537	−0.461	−0.427	−0.294
Sarvam 105B	−1.005	−1.112	−1.221	−1.274	−1.085
Qwen3.5 27B	−1.136	−1.187	−1.268	−1.337	−1.142
源草稿					−1.096
斯皮尔曼	1.000	1.000	0.976	0.976	

表 8：门控消融。当输出包含 t 个或更多违规标签时（$t \geq 1$ 为已发布门控）、不设门控（无），以及将源草稿作为第九个候选项加入无门控拟合（+Draft）时，等形式标准化 Bradley–Terry 宏观分数。斯皮尔曼相关系数以已发布的八个精炼器排序为基准。

D 忠实度失败类型

图 7 展示了各模型的失败率，表 10列出了各形式完整的失败家族分布，图 8 则按形式分别展示了各模型的这些家族。失败类型可以多选，因此占比之和不为 1。

模型	音频	故事	文章
GPT-5.5	0.999	0.978	0.986
Claude Opus 4-6	0.997	0.980	0.997
DeepSeek V4-Pro	0.951	0.748	0.911
Gemini 3.1 Pro	0.829	0.910	0.929
Grok 4.3	0.911	0.783	0.860
Gemma 4 26B	0.883	0.779	0.761
Sarvam 105B	0.625	0.520	0.554
Qwen3.5 27B	0.458	0.555	0.576
配对	9,600	4 ,000	2 ,800

表9：各精炼器相对于自身源草稿的纯文本胜率（平局按半胜计算），统计所有优化结果，无论是否通过门槛。

Form	失败类别	计数	占比
音频	事实、对象或时间线偏移	1,160	40.9%
音频	关系、身份、或能动性转变	1,103	38.9%
音频	揭示或信息状态转变	1,058	37.3%
音频	虚构内容或背景故事	939	33.1%
音频	删除或改变的事件节拍	911	32.2%
音频	对话含义变化	741	26.2%
故事	关系、身份或能动性	143	39.2%
故事	事件、结果或时间顺序	141	38.6%
故事	对话或母题含义	116	31.8%
故事	揭示、信息状态或结局	107	29.3%
故事	形式、压缩或其他	8	2.2%
文章	主张、探究或限定	312	65.1%
文章	证据、例子或归属	206	43.0%
文章	推理、反讽或结局	59	12.3%
文章	形式、压缩或其他	4	0.8%

表10：各形式失败家族的完整分布（占各形式失败输出的比例）。音频语料库包含2,833个门控失败输出，短篇故事语料库包含365个，散文语料库包含479个。

E 生成器家族亲缘关系

在留出的生成器池中，四个精炼器都有一个同级项属于同一家庭。移除所有匹配同级项的比较后，重新计算各形式的 Bradley–Terry 排行榜，三种形式的排名顺序均完全不变：音频小说中移除了 33,600 个结果中的 4,144 个，短篇小说中移除了 14,000 个结果中的 1,813 个，散文中移除了 9,800 个结果中的 1,330 个，且各自的 Spearman $\rho = 1.0$ 与主结果相符。各精炼器在同级项与非同级项比较中、双方均通过时的胜率划分见表11。

F 评审器家族稳健性

对于音频小说，两个独立的评审器家族在相同的 4,997 项通过门槛的 GPT- 涉及标题比较（双方均通过门槛的散文比较）上重新进行了评审，采用冻结的标题评分标准，并合并两种呈现顺序。

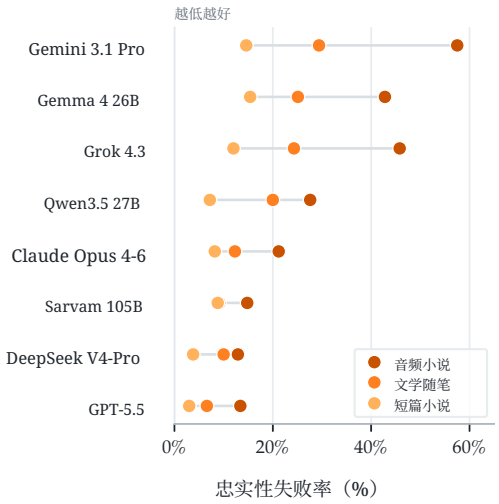

图7：按模型和形式划分的忠实度失败率。每个模型对应三个数据点，分别代表一种散文形式；失败率既因模型而异，也会在同一模型的不同形式之间变化（例如，Gemini 3.1 Pro 在音频小说上的失败率远高于在散文或故事上的失败率）。越低越好。

精炼器	同级项 WR	非同级项 WR	Δ
音频小说			
GPT-5.5	0.964	0.976	-0.012
Claude Opus 4-6	0.872	0.834	+0.037
Gemini 3.1 Pro	0.639	0.667	-0.027
DeepSeek V4-Pro	0.470	0.466	+0.004
短篇小说			
GPT-5.5	0.8612	0.8927	-0.0316
Claude Opus 4-6	0.8796	0.8639	+0.0158
Gemini 3.1 Pro	0.5628	0.6640	-0.1012
DeepSeek V4-Pro	0.4255	0.4391	-0.0136
文学散文			
GPT-5.5	0.7838	0.8879	-0.1041
Claude Opus 4-6	0.8566	0.8610	-0.0044
Gemini 3.1 Pro	0.6431	0.6920	-0.0489
DeepSeek V4-Pro	0.5469	0.5462	+0.0007

表11：生成器族亲缘关系审计。对于与留出草稿生成器属于同一模型族的四个精炼器，表中列出了它们在同族源草稿和非同族源草稿上的双方均通过实证胜率。$\Delta = \mathrm{WR}_{\mathrm{sib}} - \mathrm{WR}_{\mathrm{non}}$；平局计为 0.5，正值表示同族生成草稿更占优势。

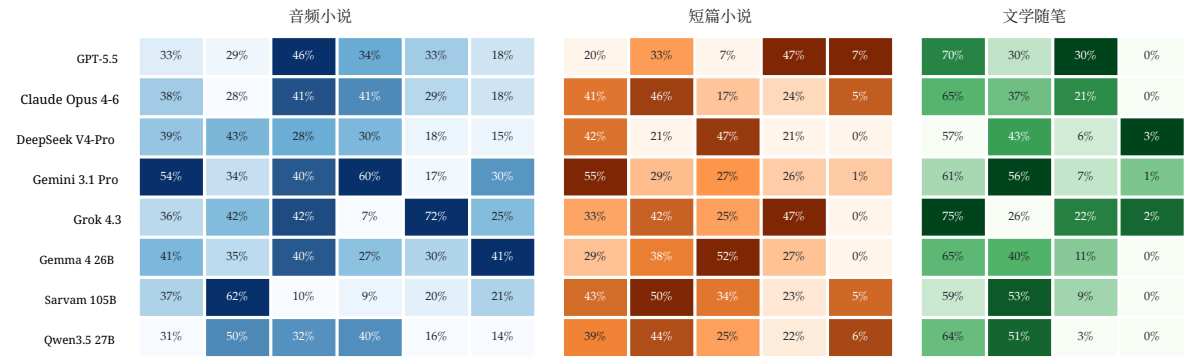

图 8：按模型和形式划分的失败类别构成。图中包含三个评审组（音频小说、短篇小说、文学散文）；每个评审组内，行按排行榜排名列出八个精炼器，列出该形式对应的失败类别；每个单元格表示模型门控失败输出中带有该类别标签的比例。失败为多标签，因此各行的数值之和不为 1。

Claude Opus 4-6 保留了完整的排名顺序（斯皮尔曼 $\rho = 1.0$），并在 0.902 的合并结果上与主要评审器意见一致；Kimi-K3 是单独评估的另一个模型族，在温度为 0 时也保留了完整顺序（斯皮尔曼$\rho = 1.0$），GPT-5.5 位居第一，且在 0.866 的比较中意见一致（4,326/4,997；95% Wilson 置信区间 [0.856, 0.875]）。用任一模型族的评审结果替换主要评审结果并重新拟合后，GPT-5.5 仍排名第 1，门控判定的比较结果也没有变化。在 1–5 分维度上，主要评审器与审计评审器之间存在约 0.2 分的整体宽松偏差，因此只有排序具有可比性；而在一个确定性重新门控的音频样本中，主要门控让顶级模型的输出通过的比例为 0.856，Claude Opus 4-6 审计评审器则为 0.981（$n = 160$；对于其他模型的输出，分别为 0.714 和 0.971，$n = 140$），因此主要门控更严格，没有迹象表明其对自家模型族有所宽待。人工标注也与此一致：在涵盖 600 个案例的门控审计中，标注员推翻了门控放行的 37 个 GPT-5.5 输出中的 2 个（0.054），低于整个审计的误通过率 0.076。短篇故事和文章排行榜依赖单一的主要评审器；未来工作将把独立模型族审计扩展到这些内容。

G 人工验证详情

读者。 所有读者和标注员均为印地语母语使用者，是本组织的有偿成员；他们已提供知情同意，并在评审时不知道模型身份。六位读者标注了音频小说的偏好——度，其中五位针对一组共同的 28 项比较进行评审；每份表单由三位读者组成书面偏好评审组，每组共享 20 项比较；三位标注员完成了第二轮书面门控审计；四位标注员则生成了 40 个沿用的音频门控标签。个体标注、读者身份和人口统计信息均不予公开。

门控审计。 表 12 汇总了合并结果。门控审计对每种形式各抽取 200 个案例，共 600 个：其中包括提交的 200 个音频小说案例，保持不变；对于每种书面形式，则包括首轮的 50 个案例和第二轮的 150 个案例，这些案例均依据同一盲评标准进行标注，并按模型和门控判定结果分层。一致性在音频小说中为 0.930，在短篇小说中为 0.935（187/200），在散文中为 0.910（182/200）。各项审计中的门控误判失败率和误通过率采用不同条件，因此不予合并：音频小说中，误判失败率为 0.073（96 个人工判定通过的案例中有 7 个），误通过率为 0.067（104 个人工判定失败的案例中有 7 个）；书面审计围绕自动门控标签进行平衡，因此短篇小说的误判失败率为 0.074（94 个自动判定失败的案例中有 7 个），误通过率为 0.057（106 个自动判定通过的案例中有 6 个）；散文则分别为 0.079（101 个中有 8 个）和 0.101（99 个中有 10 个）。表13 给出了各精炼器的一致性，所有区间彼此重叠；在 GPT-5.5（评审器和顶级模型）上，标注者推翻了其 37 个门控通过结果中的 2 个（§5）。表 14 给出了按门控失败类别划分的书面文本一致性，并像 §5中那样按形式条件化；数值范围为 0.879 至 0.973，而对于主张、探究或限定发生变化的散文，一致性为 0.950（57/60）。

偏好。 任务 A 的偏好行仅汇总独立运行的评审组中的精确分子，以概括一致性。无论对人类还是自动评审器而言，散文偏好都难以判断；评审器在散文上的较低一致性，应结合相应较低的人类之间的一致性来理解。Fleiss' κ 按评审组分别报告，绝不在独立标注的评审组之间合并；每个数值都基于一组较小的共同样本，而每个排行榜则基于数千个配对拟合得出。

指标	值	95% 置信区间	n
门控与人工评审器（合并）对比	0.925	[0.901，0.943]	600
Headline 评审器 vs 个体人类两两一致性	0.526	[0.484、0.568]	540
评审器与留出集多数意见	0.472	[0.424, 0.521]	400
人类与留一法多数表决	0.524	[0.403，0.642]	63
	0.529	[0.451、0.605]	157
Fleiss' κ，音频（5 位读者）	0.030		28
Fleiss' κ，故事（3 位读者）	0.099		20
Fleiss' κ，文章（3 位读者）	0.012		20

表 12：三种形式的汇总人类校准和门控验证。这些是评判审计统计，而非语料库级通过率估计。Fleiss' κ 按评审组分别计算，其中 n 指每位评审组读者都标注的共享比较。

精炼器	一致	n	一致性 [95%CI]
GPT-5.5	61	66	0.924 [0.835，0.967]
Claude Opus 4-6	61	65	0.938 [0.852，0.976]
DeepSeek V4-Pro	60	64	0.938 [0.850, 0.975]
Gemini 3.1 Pro	81	90	0.900 [0.821, 0.947]
Grok 4.3	62	71	0.873 [0.776，0.932]
Gemma 4 26B	64	70	0.914 [0.825，0.960]
Sarvam 105B	64	66	0.970 [0.896，0.992]
Qwen3.5 27B	64	68	0.941 [0.858，0.977]
全部精炼器	517	560	0.923 [0.898，0.943]

表 13：在400个书面案例和160个名册内音频审计案例中，各精炼器与人工标注的门控一致性。其余40个音频案例来自两个不在排行榜中的构建时期系统。

H 提示词堆栈与披露

表 15 列出了九条提示词、它们的共享核心以及形式条件化内容。三种形式共用同一套构建、门控和偏好架构；对于散文形式预期不同之处，则使用明确的体裁卡。完整的各形式提示词文件见发布版本。下文以适用于各种形式的表述概述构建过滤器、门控和标题评判提示词（03、05 和 06），同时说明叙事

形式	失败类别	n	一致性 [95% CI]
故事	事件、结果、时间顺序	44	0.909 [0.788, 0.964]
故事	关系、身份、能动性	40	0.925 [0.801,0.974]
故事	揭示、信息、结局	37	0.973 [0.862,0.995]
故事	对话或母题含义	33	0.879 [0.727,0.952]
文章	主张、探究、限定	60	0.950 [0.863,0.983]
文章	证据、例子、归属	47	0.894 [0.774, 0.954]
文章	推理、反讽、结局	25	0.920 [0.750,0.978]
两者	形式、压缩、其他	3	3/3

表 14：在涵盖 400 个案例的书面审计中，按失败类别统计的书面门控一致性；仅统计门控为该类别分配了标签的案例。标签为多标签，因此一个案例可能计入多行。

以及文章条件设定；按形式的具体内容列于表 15。

提示词03：源文本与黄金文本构建筛选器（仅限构建时）

判定内容。判断候选草稿 S 是否保留了其黄金源文本 G的核心内容，并在 G 被分离前进行检查。它只评判内容，绝不评判散文质量，也不参与排行榜评审。

音频小说。只要草稿保留了主要事件、人物、关系、揭示信息、对话含义和结局状态，即使文笔平淡也应通过。若删去或改动了重大事件、改变了关系或揭示信息、改动了对话含义、编造了重大内容、改变了结局或悬念结尾、草稿压缩过度以致无法核实，或包含明确或不安全的内容，则判定失败。

短篇小说与散文。默认通过，包括局部省略、一般性压缩、语体更平实以及内容顺序调整。只有当改动影响作品的核心锚点时才判定失败：对于小说，包括核心情境或冲突、决定性事件或因果转折、驱动性身份或关系、核心揭示、推动结果的能动性，或结局；对于散文，包括核心探究或立场、主要论点、关键推理步骤、立场赖以成立的证据，或最终立场。文体形式错误，或大段续写英语或罗马化内容，也判定失败。

输出。 一个通过或失败的判定结果，附带固定的原因标签和简短说明；书面版本还会返回配对证据片段和修复指令，用于重新生成草稿。

提示词05：忠实度门槛（二元判定，在评分前运行）

判定内容。 对源草稿 S 与润色稿 R 进行内容保留检查，仅评判源文本内容，绝不评判散文质量，并决定其是否具备进入排行榜的资格。

叙事体裁。 如果 R 遗漏重要情节点；改变事件、结果或结局；篡改事实、数字或时间线；改变情节关键物件；改变行动者或行动原因；改变角色所说、所知或所作的决定；改变身份或关系；增删揭示内容或转变信息状态；消除或引入歧义；或编造新内容，则判定为失败率。

提示词	共享核心	形式条件化内容
01 大纲	源文本转为风格中立大纲	源形式模式和元数据
02 草稿	大纲转草稿 S	音频、短篇故事或散文形式卡片
03 S-vs-G 筛选	接受草稿 S	形式特定可接受性标准
04 润色稿	保留内容，优化印地语	可表演性、叙事或散文写作技巧
05 门槛	二元 R-与-S 资格	叙事标签与散文论点／证据标签
06 标题评审	整体 A／B／平局偏好	特定文体编辑评分标准
07 J1	地道印地语、准确性、流畅度	不同文体的语体与翻译腔
08 J2	结构与推进	戏剧化或议论文
09 J3	声音与参与感	对话声音或作者声音

表15：九个提示词：共用构建—门控—偏好架构，并在散文形式预期不同时明确提供形式卡片。

散文。 如果 R 改变、遗漏或添加主张；篡改或编造证据、例子或归属；改变限定或主张的力度；或改变推理脉络、反讽或结论，则判定为失败率。

判定规则。 这项改动改变了源文本确立的内容，还是仅仅让同一内容的表达更生动？如果是前者，则判为失败。评审器会给出通过或失败的判定结果，并附上一段简短的证据片段；若判定失败，还会标注一个或多个固定的内容违规标签。

提示06：标题成对评审器（整体评判 A/B/平局，仅散文）

“更好”的含义。 将同一单元的两个版本作为相应体裁的散文来阅读，较优版本就是编辑更愿意发布的版本：印地语自然地道、语体恰当；剧情、论述或反思能够吸引读者；声音也符合体裁。评审器看不到源文本或大纲，只评判写作，不评判内容。

局限。 表达效果必须通过自然的印地语呈现；若某个版本的效果依赖英语仿译或翻译腔，读起来就会显得生硬外来，因此不会因此加分；而华丽词汇若损害可读性，也不构成优势。

输出。 给出一个整体判定结果：A、B 或平局，并附上一句理由。测试框架会运行两种呈现顺序并将结果合并，合并结果用于按形式进行 Bradley–Terry模型拟合。

I 诊断词表

人工评审员在评分前汇编并筛选候选措辞，形成固定的注册表。程式化短语注册表收录了音频小说的 272 个经审阅表层形式，以及书面形式的 274 个；额外的两个书面形式表层属于一个修辞模板类别。仿译注册表收录了 120 个经审阅表层形式，适用于全部三种形式。匹配采用针对轻度规范化天城文的精确子串匹配：FPB 统计每 1,000 个类词元中的不同匹配表层形式数量（各输出比率的均值），仿译率则为至少出现一次匹配的输出占比。精确匹配仿译检测可能会漏掉采用其他措辞表达的仿译词（Rahman等人，2026年），因此这些仿译率可能会低估仿译现象。额外的两个书面形式表层形式是文章修辞模板，不适用于音频小说，因此音频 FPB 使用 272 项表层登记表（表 16）。完整词汇表及每个表层形式的实证例句见发布版本；有代表性的仿译词见表 3。表17和18 提供了更多定性示例：有代表性的程式化短语族（其中突出显示了反复出现的短语），以及被门控捕获、经人工确认的忠实度偏移最小对。

Form	类别	来源	占比
音频	奇幻或修仙	335	27.9%
音频	剧情、家庭或社会	293	24.4%
音频	惊悚或悬疑	195	16.2%
音频	浪漫小说	133	11.1%
音频	犯罪或悬疑小说	119	9.9%
音频	超自然，历史，or 其他	125	10.4%
故事	一般文学短篇故事	358	71.6%
故事	寓言或民间故事	117	23.4%
故事	儿童故事	23	4.6%
故事	实验性故事	2	0.4%
文章	文化反思	209	59.7%
文章	文学反思类	111	31.7%
文章	个人反思类	17	4.9%
文章	文学评论	9	2.6%
文章	议论性或说明性	4	1.1%

表19：不同散文形式的源文本来源多样性。百分比按各散文形式分别计算。音频小说类别为类型家族；音频长尾行汇总了超自然、历史、动作、科幻和神话类项目。

J 来源多样性与可复现性

来源。 音频小说来源是从私人当代语料库中选取、经专业编辑的剧集；短篇小说取自印地语维基文库、BHAAV语料库（Kumar 等人，2019）和 Hindi Discourse（Dhanwal 等人，2020），文学散文则取自印地语维基文库。版权

A 面板。程式化短语负担（每 1,000 个词）

模型	音频	故事	文章
GPT-5.5	3.39	1.05	0.05
Claude Opus 4-6	4.42	1.45	0.04
DeepSeek V4-Pro	3.54	0.87	0.04
Gemini 3.1 Pro	3.15	1.04	0.06
Gemma 4 26B	4.28	0.87	0.03
Sarvam 105B	2.78	0.71	0.04
Grok 4.3	2.73	0.66	0.07
Qwen3.5 27B	3.51	0.80	0.05

B 面板。仿译输出命中率

模型	音频	故事	文章
GPT-5.5	77.3%	13.6%	2.9%
Claude Opus 4-6	81.2%	18.2%	1.1%
DeepSeek V4-Pro	70.8%	15.2%	2.0%
Gemini 3.1 Pro	82.1%	25.4%	2.9%
Gemma 4 26B	77.5%	24.2%	1.4%
Sarvam 105B	59.3%	13.2%	0.9%
Grok 4.3	43.5%	11.6%	0.9%
Qwen3.5 27B	76.3%	15.0%	1.1%

表 16：跨形式词汇诊断。A 面板：每 1,000 个类词元的程式化短语负担（每条输出的不同表面形式比率的平均值）。B 面板：至少包含 120 个经审查的仿译表面形式之一的输出百分比；三种形式使用完全相同的仿译列表。这些是描述性诊断，而非排名输入项。音频小说的程式化表达注册表包含 272 个表面形式，书面形式包含 274 个；多出的两个表面形式是仅适用于书面形式、不适用于音频的修辞模板。

短语类别	润色稿示例（突出显示重复出现的短语）
对话/声音标签	उसकी आवाज़ काँपी, "मैं जानता हूँ वह कौन है।"
对话/声音标签	“没关没系，别这么想。”帕兰贾伊没有**缓下语气清晰地说道**。
身体反应描写	他**咬紧牙关**猛地一用身站了起来。
身体反应描写	他的**手指在颤抖**。
感官／氛围填充	……自行身后一片**沉重而令人窒息的寂静**。
感官／氛围填充内容	**城市的灯光影在渐渐**消退……
寂静／紧张填充内容	小屋里面**一片死寂，仿佛**死神正屏住呼吸一般。
寂静／紧张填充内容	就在那一刻，腰间笼罩上一阵新的、**沉重的沉默**。
一般情绪叙述	那个男人的声音里此刻**明显透出慌张**。
一般情绪叙述	……但她的声音里她**想要说的话，却因**多年的责任而显得苍白无力。

表17：程式化短语负担：模型输出中高于人类基线反复出现的固定短语族示例，其中重复出现的短语已突出显示（真实模型输出）。图中按模型量化5；这是描述性特征，并非排名依据。

源短语 (S)	改写短语 (R)	忠实性问题
...明天早ᴗ上之前回来不了ᵕ。	...明天早ᴗ上之前就能回来ᵕ。	否定反转。
现在休ᵕ息 吧	现在休息 吧	指令改变，从从步行变为休息。
他ᵕ ᶠ让人送去了 是ᵕ	他ᵕ是ᶠ ᶠ让人送去了 将	事实变为推测。
正在思考自己的下一步行动ᶠ ᶠ	也许 自己的下一ᶠ 步正想着ᶠ呢	陈述的事实变成了可能性。
कपूᵕ尔家族的继承人	唯一ᴗ继承人ᵕ （无依据地补充）ᵕ 家族	凭空添加了“唯一继承人”这一细节。
我的ᵕ意见क्म ज़्य更有分量ᵕ	我的ᵕ意见क्म 也同样有分量ᵕ	更大的权威性被改成了同等的权威性。
村庄里最ᵕ 大 ᴗ ज़ुर्ग	生命有ᵕ 一百二十二年 的ᵕ历程ᴗ。ᵕ	虚构的具体年龄。
ṡ迪普（控制的 ᶠ 负责者）	他ᶠ和ᵕ迪普 ᶠ 一起ᵕ 飞了 ᶠ 多少越ᵕ 。ᶠ थीं	虚构的共同历史。
（她喜欢ᴗ日出）	ᶠ 愤怒在她ᵕ 心中 多年ᵕ 燃烧不ᶠ 息	编造了一个多年的反故事。
揪着耳朵蹲起 ᵕ ठक	来回 挣扎 ᵕ 觉得，ᵕ से व्य运动	具体的 惩罚变成了含糊的动作。

表18：门控捕捉到的忠实度偏移（经人工确认的最小配对）：绿色部分为源文本片段，红色部分为改写中发生变化的片段。偏移现象遍及所有模型，包括排名靠前的模型，因此门控会检查每一条输出。所有文本均由模型生成。

以及发布版本均在条目层级处理，部分材料受到限制（伦理声明）。

构建过滤器。 提示词03（附录 H）将每个候选草稿与其源文本进行比较 G在 G 脱离后，依据形式条件化的措辞进行比较。统计同时通过严格长度和脚本检查的候选草稿后，过滤器拒绝了 1,612 份音频小说草稿中的 307 份（19.0%）、624 份短篇小说草稿中的 124 份（19.9%），以及 394 份文章草稿中的 44 份（11.2%）。被拒绝的书面草稿会从留出池中换用另一生成器重新生成；13.8% 的短篇小说源文本和 12.0% 的文章源文本需要这样处理；音频草稿被拒绝的源文本则会替换为另一集。音频拒绝项带有多标签原因标签：重大事件发生变化（占拒绝项的 38%）、人物或关系发生变化（37%）、结局或悬念结尾发生变化（24%）、重大事件缺失（24%）、对话含义发生变化（21%），以及新增事件或揭示（17%）。该过滤器为自动运行，且本身未经人工审查。

长度。 长度按类似词语的词元计数；由于天城文在八个系统中的分词方式各不相同，因此不列出模型分词器的计数。音频小说草稿平均为 1,468 个词元，短篇小说草稿平均为 1,186 个，散文草稿平均为 1,067 个，表 20 列出了每种精炼器的输出长度相对于草稿长度的比例。相较于 G，短篇小说的逐篇草稿与源文本长度平均比为 0.657，散文为 0.664；语料库总长度之比则更低（分别为 0.490 和 0.461），因为草稿长度受到上限约束，因此较长的源文本按比例压缩得更多。

精炼器	音频故事	文章
GPT-5.5	1.598 1.156 1.099	
Claude Opus 4-6	1.119 0.963 0.987	
DeepSeek V4-Pro	1.102 1.017 1.026	
Gemini 3.1 Pro	1.661 1.109 1.117	
Grok 4.3	0.775 0.839 0.909	
Gemma 4 26B	1.169 1.046 1.039	
Sarvam 105B	1.003 0.955 1.009	
Qwen3.5 27B	1.142 1.007 1.014	
平均初稿（词元）	1,468 1,186 1,067	

表20：按精炼器和形式划分的平均润色稿长度与平均初稿长度之比，以类词元计。

设置。基准测试为每个模型和阶段采用一种固定且有文档记录的设置；表21 总结了温度设置。

阶段	温度
大纲提取	0.1
初稿生成与润色稿	0.65
构建过滤器	0.0
门控与评估裁判	在支持的情况下为 0.0

表21：各阶段的温度设置。GPT-5.5 适配器省略不受支持的温度字段。

模型文档。主要技术报告、模型卡或发布公告记录了以下构建模型：Gemini 3.1 Flash-Lite（GoogleDeepMind，2026a）、GPT-5.4 Mini（OpenAI，2026b）、Claude Sonnet 4.5（Anthropic，

2025年）、Gemini 3 Flash（Google DeepMind，2025年）、Kimi K2（Kimi Team，2025年）、GLM-5（GLM-5-Team，2026年）、MiniMax M2.7（MiniMax，2026年）和 Mistral Medium 3.5（Mistral AI，2026年）；以及接受评估的系统 GPT-5.5（OpenAI，2026a）、Claude Opus 4-6（Anthropic，2026年）、Gemini 3.1 Pro（Google DeepMind，2026b）、DeepSeek V4（DeepSeek-AI，2026年）、Gemma 4（Gemma Team，2026年）、Sarvam 105B（Sarvam Foundation Models Team，2026年）、Qwen3.5 27B（Qwen Team，2026年）、Grok 4.3（xAI，2026年）和 Kimi-K3（Kimi Team，2026年）。Kimi K2 的引文记录了所请求的 K2.6 路线所使用的模型族。

多样性。 这 1,200 个音频小说来源涵盖 82 部剧集（每部剧集的剧集数中位数为 14），分布在大约十一个类型家族中：奇幻与修炼占 27.9%，家庭与社会剧占 24.4%，惊悚与悬疑占 16.2%，浪漫小说占 11.1%，犯罪与推理占 9.9%，此外还有少量超自然、历史和神话题材。70.8% 的作品为当代题材，每集平均包含 4.1 个场景和 7.3 个人物；所有此类元数据均为近似值，提取自大纲阶段。表 19 展示了这一分类，同时列出书面形式源文本的多样性：500 篇短篇小说和 350 篇散文各自涵盖多种子类型，并给出了每种形式内部的占比。场景数和人物数的直方图以及完整表格见发布版本。