迈向对基于大语言模型的推荐系统进行更贴近现实的不确定性量化评估

Hao Wu *hwu522@wisc.edu*
威斯康星大学麦迪逊分校

Sipeng Chen *sipengchen2028@u.northwestern.edu*
西北大学

Weikai Li *weikaili@cs.ucla.edu*
加利福尼亚大学洛杉矶分校

Tianyi Chen *tchen523@wisc.edu*
威斯康星大学麦迪逊分校

Qiancheng Yang *qyang333@wisc.edu*
威斯康星大学麦迪逊分校

Hang Zhou *hzh@unc.edu*
统计与运筹学系 &数据与信息科学学院 北卡罗来纳大学教堂山分校

Yizhou Sun *yzsun@cs.ucla.edu*
加利福尼亚大学洛杉矶分校

Wei Wang *weiwang@cs.ucla.edu*
加利福尼亚大学洛杉矶分校

罗晓 *xiao.luo@wisc.edu*
威斯康星大学麦迪逊分校

已在 OpenReview 上评审： https://openreview.net/forum?id=GXDe6REdu3

arXiv:2611.33036v1 [cs.IR] 24 Nov 2026

摘要

大语言模型（LLM）为推荐系统提供了灵活的接口。然而，对其已生成的排序进行可靠的实例级不确定性监测是必要的。在这种场景下，比较不确定性量化（UQ）方法颇具挑战，因为大语言模型排序器缺少标准的物品分类头。其分数可能取决于选项标签、提示、生成流程以及可用的用户与物品证据。现有方法从不同来源推导不确定性，并将其应用于推荐流程的不同阶段。有些方法评估已生成的排序，另一些则利用不确定性改变排序本身，因此难以直接比较。本文提出一种固定排序基准，在统一的输出约定下，将八种异质不确定性量化信号作为记录级监测器进行评估。我们在四个推荐数据集和两种大语言模型骨干上，采用统一的固定排序协议评估这些信号。根据结果，我们总结出三项主要发现。首先，即使面对相同的参考实例，不确定性量化方法的排序也会随骨干模型而变化。其次，在若干评估条件下，与排序对齐的 Top10Mass 能最有效地检测标量 Error@10，体现出有效的与截断点对齐的监测；但在长尾热门度目标上的优势有所缩小。第三，标量不确定性量化信号难以识别哪个对照组具有更大的截断 NDCG@10 差值，而源特定信号（包括扰动排名不稳定性和简单的

列表级距离）则能提供更强的配对对照一致性，体现出它们对两种受控证据源的敏感性。这些发现促使我们将排名失败检测、跨推荐条件的鲁棒性，以及对受控证据变化的响应视为不确定性量化的不同属性来评估。代码和评估工件见我们的代码库：https://github.com/haowu0916edisonwu/recuq-benchmark.git。

1 引言

大语言模型（LLM）通过用自然语言表示用户历史记录、候选项目和任务指令，并要求指令微调模型对候选项进行排序，为推荐系统提供了灵活的接口（Hou 等，2024b）。这种接口支持零样本或少样本推荐，无需针对每个领域额外设置专门的推荐预测头。在本文中，我们聚焦于大语言模型作为重排序器的场景，直接使用已发布的指令微调骨干模型，而不从头训练，也不进行针对推荐任务的微调。然而，流畅且格式规范的排序列表仍可能不可靠。即使模型表现出很强的信心，排名靠前的结果中也可能缺少相关物品。近期研究开始探讨，仅凭平均排序效用是否足以评估基于大语言模型的推荐系统（Jiang 等，2025）。这一更广泛的可靠性视角考察：当置信度不可靠、出现公平性问题，或推荐结果对偏好和呈现偏差敏感时，模型输出是否仍值得信赖（Sah 等，2026；Tang 等，2026）。因此，实际部署需要能够监测每次已发布排序可靠性的信号。

然而，将不确定性量化（UQ）应用于基于大语言模型的排序并非易事。与固定类别预测不同，推荐会生成依赖于用户的排序列表，而其排名第一的k 结果正是监测对象。大语言模型排序器可能输出候选标签或完整的排序列表，其不确定性也可能受到提示、选项标签格式、解码或采样程序，以及上下文中所提供的用户或项目证据的影响。因此，不同的不确定性量化方法可能反映模型行为的不同方面，使其分数难以在统一标准下解读和比较。已有研究从不同角度探讨了不确定性。经典方法关注与预测相关的置信度：要么利用预测概率检测可能出现的错误（Hendrycks 与 Gimpel，2016），要么考察报告的置信度是否与经验正确性相符（Guo 等，2017）。语言模型研究将这一思路从单一概率分数拓展至对自评正确性（Kadavath 等，2022）以及采样响应之间一致性的研究（Wang 等，2022；Kuhn 等，2023）。此外，在推荐领域，不确定性的作用更加广泛：它被用于估计推荐质量（Li 等，2025）、区分推荐不确定性与提示的不确定性（Kweon 等，2025），以及支持重排序或模型优化（Xu 等，2026；Fan 等，2026）。这些方法难以直接比较，因为它们所用的不确定性信号各不相同，而且不确定性介入推荐流程的方式也不一样（Xu 等，2026；Fan 等，2026）。尤其是，用于重排序或模型优化的方法也会改变推荐输出，因此难以将不确定性量化监测质量与底层排序的变化区分开来。为单独评估监测质量，我们因此固定每个已发布排序，并将所有方法映射到相同的记录级输出。这种固定排名设计使不同的不确定性量化信号能够在统一标准下，针对相同的推荐结果进行评估。

基于这一固定排名设计，我们构建了一个用于评估基于大语言模型的推荐中不确定性量化的基准测试。每个20候选项基础排序仅生成一次，并在不确定性量化评估期间保持不变。在统一的输出约定下，我们将八种异质不确定性量化信号作为记录级监测器进行比较。完整的 Llama-3.1-8B-Instruct 评估（Grattafiori 等人，2024年）涵盖四个推荐数据集和四种评估设置：完整证据参考条件（热启动）、用户历史截断（稀疏历史）、受控候选项文本消融（物品文本鲁棒性），以及正向项目属于流行度长尾的热启动记录子集（尾部目标）。我们还在相同的4,000个参考实例上，以非思考模式评估 Qwen3-8B（Qwen 团队，2025年），以考察 UQ方法排序是否能在第二个骨干模型上保持不变。该比较仅限于参考条件，并不能证明广泛的跨模型迁移。该基准定义了两项互补的评估任务。1 **排序失败检测**评估不确定性量化分数能否标记 Error@10 案例，即目标项目未出现在

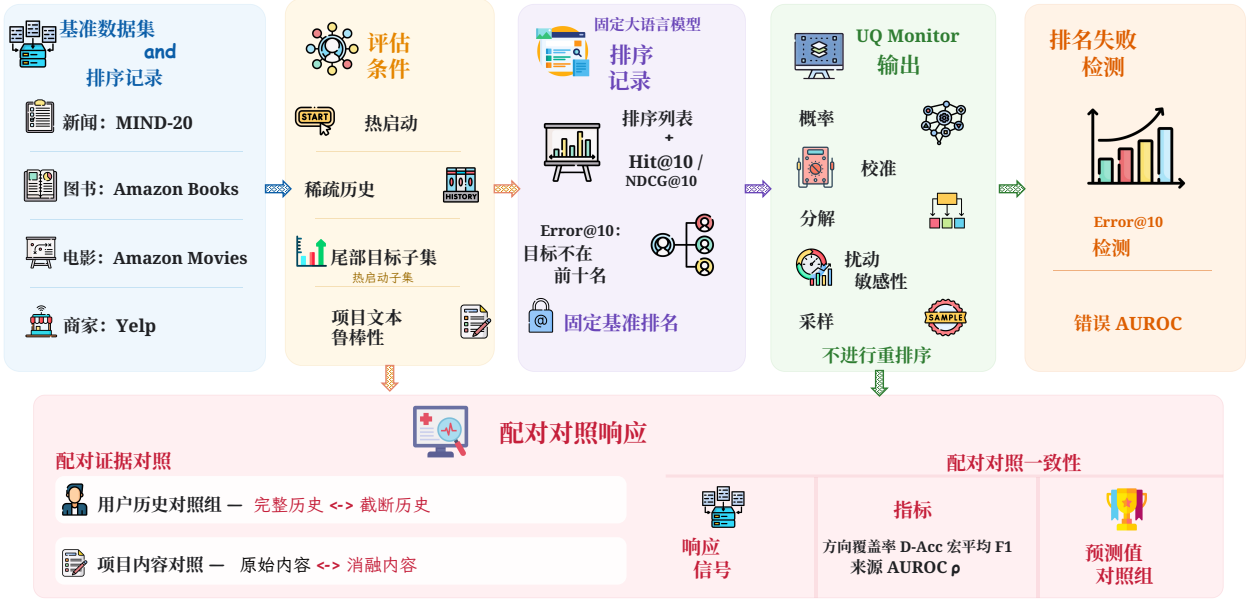

图 1：**基准测试流程**。使用不进行重排序的 UQ 监测器对固定排序记录进行评分，并针对 Error@10 检测和范围限定的配对控制响应一致性进行评估。

固定的前 10 名排序中。2 **配对对照响应一致性**针对同一基础实例比较匹配的用户历史干预和物品内容干预，并测试某种方法对截断 NDCG@10 差值较大的来源是否响应更强。稀疏历史和物品文本鲁棒性分别提供用户历史和物品内容对照。

我们的实验得出了三个主要观察结果。第一，UQ 监测器的相对有效性在不同大语言模型骨干之间并不稳定。第二，与排序对齐的概率质量为检测排序失败提供了强有力的标量信号，但其优势会因评估条件和目标流行度区间而异；这些结果支持针对相应排序失败目标采用排序对齐监控。第三，标量失败检测与配对证据—响应一致性反映了 UQ 的不同属性：在经裁剪的配对控制目标下，源特定信号更适合进行源特定诊断，而简单的非不确定性量化排序距离也表现出同样强的一致性。最后这一发现仅限于两个构造的对照组，并不能证明具有普遍性的来源归因。我们的贡献总结如下：

- 我们定义了面向基于大语言模型的推荐的固定排序不确定性量化监控，并引入了两项评估任务：Error@10 检测，以及针对用户历史和物品内容证据的配对控制响应一致性；
- 我们构建了一个大规模基准测试，在四个数据集上比较八种不确定性量化信号。其中包括完整的 Llama 评估套件，以及在相同 4,000 个实例上、遵循统一记录级契约进行的 Qwen Warm-start 复现；
- 我们为评估基于大语言模型的推荐中的不确定性量化提供了实用指南：排序失败检测和配对控制响应一致性反映了不确定性量化的不同属性；此外，还应在不同骨干模型和推荐条件下评估方法的鲁棒性，因为单一参考设置得出的结论可能无法泛化。

2 相关工作

2.1 用于推荐的大语言模型

基于语言模型的推荐建立在早期将推荐表述为文本生成任务的研究之上。P5 将用户交互、物品信息和不同的推荐任务表示在

统一的文本到文本框架中（Geng 等，2022）。随着指令微调大语言模型的发展，研究人员开始考察通用语言模型是否能够直接执行推荐。常见设置是以自然语言呈现用户历史、候选项目和排序指令，然后要求模型返回一个排序列表（Hou 等，2024b）。后续研究进一步使大语言模型适应推荐数据和项目空间。TALLRec 使用推荐指令对大语言模型进行微调，以预测用户偏好（Bao 等，2023）；BIGRec 生成项目语义表示，并通过两步落地过程将其映射到实际项目（Bao 等，2025）；GenRec 则根据用户和项目信息文本直接生成目标项目（Ji 等，2024）。此外，RecRanker 对大语言模型进行微调，以执行逐点、成对和列表式的前-k 排序任务（Luo 等，2025）。大语言模型也已被集成到多阶段推荐系统中。LlamaRec 首先使用轻量级序列推荐系统检索规模更小的候选集，然后对这些候选项应用大语言模型排序器（Yue 等，2023）。总体而言，这些研究涵盖了在推荐任务中使用语言模型的多种方式，例如任务特定微调、项目落地、直接生成、候选项排序，以及检索与排序集成。近期研究进一步将评估范围扩展到平均排序质量之外，因为较强的总体性能并不能保证对每位用户或每个候选集都能产生可靠结果（Jiang 等，2025）。这一更广泛的可靠性视角涵盖了不同推荐案例中的不确定性和公平性（Sah 等，2026），以及对偏好偏差和呈现偏差的敏感性（Tang 等，2026）。基于这一可靠性视角，我们的研究评估不确定性信号能否在保持推荐系统输出和候选项顺序不变的情况下，识别不可靠的已生成排序。

2.2 不确定性量化

不确定性量化提供信号，用于评估模型输出的可靠性（Gawlikowski 等人，2023）。基于概率的方法通过最大置信度、预测熵以及领先候选项之间的差距等量度来概括不确定性（Gawlikowski 等人，2023；Hendrycks 与 Gimpel，2016；Settles，2009）。校准则关注另一个问题：检验报告的置信度是否与实际观察到的正确频率相符（Guo 等，2017）。选择性预测考察随着越来越多的不确定样本被移除，预测风险如何变化（Geifman 与 El-Yaniv，2017）。这些属性彼此相关，但其中一项表现出色，并不意味着其他项也同样出色。对于语言模型，也可以根据生成的回答估计不确定性。先前研究探讨了模型能否评估自身答案的正确性（Kadavath 等，2022），而重复生成则揭示了其输出的稳定程度。自洽性衡量采样输出之间的一致程度（Wang 等，2022），语义不确定性则先按含义对回答分组，再衡量其差异（Kuhn 等，2023）。近期研究还考察了样本一致性与置信度校准之间的关系（Lyu 等，2025；Xiao 等，2025）。然而，这些方法依赖不同的输出和采样过程，因此难以在共同的评估设置下比较其不确定性分数。在推荐领域，不确定性已用于多种目的。Paliwal 等（2024）通过选择性预测、分布外检测和主动学习，研究基于点击率的推荐中的预测不确定性。Angelopoulos 等（2023）在有限样本条件下构建推荐集合并控制错误发现率；Li 等（2025）则利用列表分布不确定性，在没有相关性标签的情况下估计推荐表现。对于基于大语言模型的推荐，Kweon 等（2025）将不确定性分解为推荐组件和提示组件。其他研究则利用不确定性改变推荐流程。GUIDER 根据不确定性调整重排序决策（Xu 等，2026）；不确定性感知的生成式推荐则将不确定性纳入偏好优化和置信度学习（Fan 等，2026）。与利用不确定性改变推荐流程的方法不同，我们的基准测试在排序完成后评估异质不确定性量化信号，保持候选项顺序不变，并将每种方法映射到相同的记录级输出规范，以便直接比较。

2.3 推荐系统的鲁棒性

鲁棒性研究考察推荐系统在数据稀疏、失衡、含噪或受到扰动时能否继续保持可靠（Zhang 等人，2025年）。稀疏的用户历史记录使得准确表示用户偏好变得困难（Adomavicius 和 Tuzhilin，2005）。此外，流行度失衡可能会降低长尾物品的可见度和推荐质量（Abdollahpouri 等人，2019）。在基于大语言模型的推荐中，这些

问题会通过提示中包含的文本证据表现出来，因为缺失的用户历史记录或物品信息可能同时改变最终排序和不确定性分数。输入扰动为研究这种敏感性提供了直接途径。Oh 等人（2022年）衡量了对用户交互施加小幅扰动后推荐列表的变化。受控测试为评估模型行为提供了更广泛的框架。例如，CheckList 使用有针对性的行为测试来揭示总体准确率可能掩盖的失败情况（Ribeiro 等，2020年）。此外，近期关于大语言模型鲁棒性的研究还考察了模型在保持原始含义的受控提示变体下的稳定性（Romanou 等人，2026年）。这些研究表明，单一参考设置下的结果可能无法描述模型在其输入证据发生变化后的行为。推荐研究还考察了哪些输入会导致不确定性。可解释不确定性归因衡量用户历史记录中的单个交互如何影响序列推荐中的预测不确定性（Balsells-Rodas 等人，2024年）。在受控鲁棒性测试的基础上，我们的分析聚焦于两个预先定义的证据来源——用户历史记录和物品内容，并评估不确定性量化响应的变化是否与两个对照组所对应的NDCG@10差异一致。这项范围限定的评估不同于一般的来源归因，也不声称进行了因果识别。

3 基准测试与评估协议

我们的基准测试在固定的大语言模型排序记录上比较异构的不确定性量化（UQ）信号。基础排序仅生成一次，并在整个不确定性量化评估过程中保持不变。图 1 总结了基准测试流程，涵盖从评估条件和固定排序记录，到不确定性量化监测输出、排序失败检测以及配对对照响应分析的各个环节。本节定义问题设置、数据集与评估条件、评估任务、不确定性量化监测信号以及统计协议。

3.1 问题定义

给定一条包含用户历史、20 个候选项目和已解析基础排序的固定大语言模型重排序记录，我们的基准测试将不确定性量化方法评估为已发布排序的记录级监测器，同时保持候选项顺序不变。每种方法都会针对同一基础记录输出一个标量不确定性分数，并且还可以针对用户历史证据和物品内容证据提供配对对照响应分数。这两个对照在基础实例级进行配对：在完整用户历史与截断用户历史，以及原始物品内容与消融后的物品内容两种条件下，对同一个底层推荐案例进行评估。我们的基准测试考察三个问题：（1）不确定性量化分数能否识别 Error@10 案例，即正例物品不在固定的前十名排序中；（2）不确定性量化方法的相对性能是否在不同骨干模型和评估条件下保持稳定；以及（3）特定于对照组的不确定性量化响应，是否与具有较大截断 NDCG@10 差值的来源一致。

对于记录 i，令 $\mathcal{C}_i = \{c_{i1}, \dots, c_{i20}\}$ 表示候选集，$y_i \in \mathcal{C}_i$ 表示正例物品。遵循大语言模型作为重排序器的形式（Hou 等，2024b；Jiang 等，2025），基础大语言模型根据自然语言提示对候选标签进行排序。我们用 π_i 表示解析后的排序，用 $B_i^{(10)}$ 表示其固定的前十候选集。在整个不确定性量化评估过程中，排序 π_i 保持不变，因此所有不确定性量化方法监测的都是同一已发布排序，且不会更改候选项顺序。

3.2 数据集与评估条件

我们使用 Llama-3.1-8B-Instruct（Grattafiori 等人，2024年），按照 Hou 等人（2024b）和 Jiang 等人（2025）的“大语言模型作为重排序器”设置，构建固定的20候选项重排序基准。我们还在相同的4,000个参考实例上评估 Qwen3-8B 非思考模式（Qwen 团队，2025年），以考察方法排序在第二个骨干模型上能否复现。Qwen 仅用于复现相同实例上的参考条件。对于 Amazon Books、Amazon Movies 和 Yelp，我们采用留出最后正反馈评估：将用户最后一次正向交互视为目标项目，并在抽样得到的20项候选集中进行评估。对于 MIND-20，我们使用原始基准提供的基于曝光的候选集，其中每条已记录的曝光都包含展示过的候选新闻，而被点击的项目作为正向目标。记录基于 MIND（Wu 等人，2020年）、亚马逊评论（Hou 等人，2024a）和 Yelp 开放数据集（Yelp 公司）构建。

数据集	领域	基础记录形式	Base	记录
MIND-20	News	点击新闻历史 + 20 候选新闻	1,000	11,000
Amazon Books	图书	用户历史 + 20 书籍候选项	1,000	12,000
Amazon Movies	电影	用户历史 + 20 电影候选项	1,000	10,163
Yelp	商家	用户历史 + 20 商家候选项	1,000	12,000

表1：**基准测试概览**。每个数据集均从1,000个基础实例开始；Llama运行在尾部目标资格筛选前生成45,163条排序记录。

	热启动		稀疏历史		尾部目标子集		物品文本鲁棒性	
数据集	Hit@10	NDCG@10	Hit@10	NDCG@10	Hit@10	NDCG@10	Hit@10	NDCG@10
MIND-20	0.546	0.270	0.557	0.272	0.557	0.278	0.539	0.266
图书	0.697	0.438	0.768	0.530	0.584	0.341	0.717	0.453
电影	0.579	0.316	0.594	0.332	0.566	0.284	0.614	0.341
Yelp	0.605	0.357	0.649	0.411	0.598	0.334	0.614	0.363

表2：**基础大语言模型推荐质量**。不同数据集和评估条件下固定 Llama 排名的 Hit@10 和 NDCG@10。

按照 Jiang 等人（2025）的评估设置，我们从每个数据集的1,000个基础实例开始。我们将原始完整证据记录作为热启动参考条件。稀疏历史保留最多50条最近的正向交互的完整历史（亚马逊和 Yelp 的评分 ≥ 4；MIND 的点击新闻），以及仅保留最近 $k \in \{1,3、5、10\}$ 条交互的四种变体，候选项及其文本保持不变。这些变体保留近期交互并移除较早的交互。物品文本鲁棒性会从全部20条候选描述中移除一个字段，同时保持历史不变：MIND-20 移除标题、摘要或类别；图书移除标题、描述、类别集合或作者；电影移除标题、描述或类别集合；Yelp 移除名称、类别集合、位置或评论摘要（最多三条训练评论，每条240个字符）。只有至少一条候选项具有相应字段时，才会纳入该字段变体。这些条件提供用户历史和物品内容对照组。尾部目标保留热启动记录中目标项目的训练交互计数不高于有训练交互的物品第25百分位数的记录：MIND-$20 \leq 5$、图书和电影 ≤ 3，以及 Yelp ≤ 2。训练交互计数为零的目标也包含在内。MIND 的计数依据训练期间的点击；亚马逊和 Yelp 的计数依据每位用户验证集和测试集物品之前的正向交互。Llama 运行共生成45,163条排序记录。将流行度尾部条件限制为符合条件的1,312条尾部目标记录后，最终评估纳入42,475条记录：4,000条热启动记录、20,000条稀疏历史记录、17,163条物品文本鲁棒性记录，以及1,312条尾部目标记录。表1汇总了各数据集生成的记录。Amazon Movies 生成的变体较少，因为在我们的任务设置中，部分项目文本消融字段不可用。

基础排序质量。表2报告了每种评估条件下固定 Llama 排名的 Hit@10 和 NDCG@10。这些数值描述了不确定性量化方法所监测的推荐记录，并提供了评估不确定性之前的基础排序质量。移除证据后的增益在一定程度上反映了位置排序（候选项顺序或其逆序）的减少，这类排序占 Llama 热启动输出的 46.1%。在非位置排序中，截断主要对图书数据集有帮助，对 Yelp 的帮助略小，这与近期交互具有预测性相一致。项目文本增益取决于字段；对于电影数据集而言，这些增益部分反映了变体构成（附录A.11）。MIND-20 和电影数据集的热启动基础排序相对较弱（表2）。这些固定排序决定了可供监测的 Error@10 记录。错误 AUROC 衡量了在所评估条件下，每个信号区分错误记录与非错误记录的能力（表3）。附录A.11报告了对汇总的非位置 Llama 热启动排名的补充分析。

3.3 评估任务与指标

排序-F **失败检测**。对于每条固定排序记录 i，我们将 Error@10 事件定义为

$$e_i = \mathbf{1}\left[y_i \notin B_i^{(10)}\right], \tag{1}$$

其中，$e_i = 1$ 表示正例物品 y_i 未出现在固定的前十候选集 $B_i^{(10)}$ 中。Error@10 关注单个观测到的正例物品，其范围比总体排序可靠性更窄；我们的监测结论仅针对这一事件。令 $u_i^{(m)}$ 表示不确定性量化方法m产生的标量不确定性分数，数值越大表示估计的不确定性越高。错误 AUROC 是我们的主要指标，用于衡量 $u_i^{(m)}$ 区分 Error@10=1 案例与正例物品仍保留在固定前 10 名排序中的记录的能力（Fawcett，2006）。在共享的热启动记录上，我们还报告错误 AUPRC、Risk@80、AURC 和 E-AURC。当 Error@10 案例分布不均衡时，错误 AUPRC 可补充 AUROC。Risk@80 衡量丢弃不确定性最高的 20% 记录并保留其余 80% 记录后的 Error@10 比率。AURC 汇总所有覆盖率水平下的风险—覆盖率权衡，而 E-AURC 衡量 AURC 超出预言机排序的部分（Geifman 与 El-Yaniv，2017）。对于 Risk@80、AURC 和 E-AURC，数值越低越好。

配对对照响应一致性。我们进一步评估不确定性量化信号对用户历史和物品内容证据受控变化的响应。用户历史对照比较完整历史记录与截断历史记录，物品内容对照则比较原始项目文本与消融后的项目文本。每个带符号的 NDCG@10 差异，均为原始输入的分数减去相应受控变体的平均分数。差异为负表示受控输入下的排序质量更高。令 $N(\cdot)$ 表示 NDCG@10（Järvelin & Kekäläinen，2002）。对于案例 i，截断 NDCG@10 差值为

$$\Delta_{i,u} = \left[N_{i,\text{full}} - \overline{N}_{i,\text{trunc}}\right]_+, \tag{2}$$

$$\Delta_{i,v} = \left[N_{i,\text{orig}} - \overline{N}_{i,\text{abl}}\right]_+, \tag{3}$$

其中，$[x]_+ = \max(x,0)$，u 表示用户历史，v 表示物品内容。$N_{i,\text{full}}$ 和 $N_{i,\text{orig}}$ 分别表示完整历史记录设置和原始项目文本设置下的 NDCG@10，而 $N_{i,\text{trunc}}$ 和 $N_{i,\text{abl}}$ 则分别表示相应截断历史记录变体和项目文本消融变体的平均值。因此，$\Delta_{i,u}$ 和$\Delta_{i,v}$ 分别是用户历史对照和物品内容对照的截断 NDCG@10 差值。$\Delta_{i,u} = \Delta_{i,v}$ 的案例不纳入主导控制评估。对于其余案例，操作性主导控制标签为

$$d_i = \arg\max_{s \in \{u,v\}} \Delta_{i,s}. \tag{4}$$

该标签为案例 i 选择截断 NDCG@10 差值较大的来源。在 4,000 个 Llama 案例中，历史/物品效应为负的比例分别为 49.5%/38.8%；截断操作额外产生 1,798 个平局，并保留其余非平局案例的用户/物品标签。互补的有符号分析选择未截断效应较大的来源（3,617 个非平局案例），并使用未裁剪标量响应（附录 A.10）。令 $r_{i\,u}^{(m)}$ 和 $r_{i\,v}^{(m)}$ 分别表示方法 m 的用户历史和物品内容响应分数。采用显式控制特定响应的方法直接使用这些分数。对于仅标量方法，则根据相应配对对照下不确定性的增加量推导两种响应，详情见附录 A.4。对于一对非平局响应，方法 m 预测

$$\hat{d}_i^{(m)} = \arg\max_{s \in \{u,v\}} r_{i,s}^{(m)}. \tag{5}$$

因此，预测会选择方法 m 响应较大的那个对照；响应分数相等则视为平局预测。我们报告方向覆盖率、主导控制准确率、宏平均 F1、来源 AUROC，以及与 NDCG@10 差异的斯皮尔曼秩相关系数。方向覆盖率衡量方法产生非平局响应的频率。其余指标衡量方法响应与两个构造对照下相对 NDCG@10 差异的一致程度。

3.4 UQMonitoring 信号

第 3.3 节根据通用标量不确定性分数和控制特定响应分数定义了评估目标和指标。接下来，我们将介绍用于生成这些量的八种不确定性量化信号。这些信号涵盖基于概率的置信度、校准、分解、扰动敏感性和采样（Gawlikowski 等人，2023年；Kweon 等，2025年）。所得的排序级信号可能同时反映

偶然不确定性和认知不确定性，包括用户或项目证据中的歧义性或不完整性，以及对模型、提示或采样程序的敏感性。我们在共享协议下评估其观察到的监测行为，而不强行进行偶然—认知分解。所有信号均监测同一固定基础排序。详细的分数定义、方向约定和实现参数见附录 A.3。

概率与校准信号。 LabelProb-Max 根据有效候选标签$A–T$ 中下一个词元的最大概率推导不确定性（Hendrycks 与 Gimpel，2016）。LabelProb-Entropy 衡量同一选项标签分布的离散程度（Shannon，1948；Settles，2009）；LabelProb- Margin 则衡量概率最高的两个标签之间的歧义性（Scheffer 等，2001；Settles，2009）。温度缩放标签概率对该分布进行事后标量校准（Guo 等，2017）。LabelProb-Top10Mass 根据分配给固定基础前十集合的概率质量推导不确定性，从而使概率质量截断值与受监测的 Error@10 事件相匹配。Top10Mass 将单字母提示中的选项标签分布与给出的列表式排名相结合，因此在一定程度上衡量两种引导格式之间的跨引导一致性。附录 A.12 将其与根据原始列表式排名提示词重建的教师强制评分进行比较。

分解与采样信号。 分解式大语言模型不确定性量化遵循 Kweon 等（2025年）的做法，使用候选项顺序变体来区分推荐和提示不确定性组件。自洽性为每条记录抽取五个随机排序，并测量其前十不一致性，这遵循了将输出一致性用作不确定性证据的先前研究（Wang 等，2022；Xiao 等，2025年）。

成对控制扰动信号。 扰动排名不稳定性衡量在构造的用户历史和物品内容对照组下，固定的前十名集合变化有多大，这遵循基于扰动的行为评估方法（Ribeiro 等，2016；2020）。除了标量不稳定性分数外，该方法还为成对控制评估保留了单独的用户历史和项目内容响应。除基础排序外，四种标签概率信号共用一次前向传播；温度缩放重复使用这次前向传播，并增加一次离线校准拟合。分解式大语言模型不确定性量化使用五次前向传播，自洽性使用五次采样生成，而扰动排名不稳定性平均使用 7.3 次受控贪心生成结果。Llama 评估重复使用为鲁棒性条件生成的受控排名（附录 A.14）。

3.5 统计协议

每个评估条件的总体分数通过汇总该条件下所有合格记录计算得出。热启动置信区间采用按数据集分层、跨方法配对的 1,000 次自助法重采样。稀疏历史、物品文本鲁棒性和尾部目标子集采用按数据集分层的 200 次配对自助法重采样。对于包含同一基础实例多个变体的条件，这些变体会一起进行重采样。实现细节、工件对齐、解析检查和覆盖率统计见附录A。

4 实验与分析

我们围绕八项观察组织固定排序评估。观察1考察骨干模型敏感性，观察2–7考察排序失败检测，观察8考察配对控制响应一致性。图 2总结了主要趋势。其他诊断结果见附录A。

4.1 实验设置

完整的 Llama-3.1-8B-Instruct 评估套件包含 42,475 条符合条件的排序记录：4,000 条热启动记录、20,000 条稀疏历史记录、尾部目标子集中的 1,312 条记录，以及 17,163 条物品文本鲁棒性记录。Qwen3-8B 非思考模式评估仅涵盖相同的 4,000 个热启动实例，用于检验热启动方法排序能否在第二个骨干模型上复现。热启动是参考条件。稀疏历史和物品文本鲁棒性分别提供用户历史和物品内容对照。尾部目标条件通过根据目标流行度筛选热启动记录构建。置信区间和

	标签概率 Max	标签概率 Top10Mass	标签概率 熵	标签概率 间隔	温度 缩放	分解 大语言模型不确定性量化	扰动 排名不稳定性	自我- 一致性
Llama-3.1-8B-Instruct								
MIND-20	0.476	**0.528**	0.486	0.482	0.477	0.484	0.517	**0.528**
图书	0.607	**0.673**	0.601	0.583	0.608	0.620	0.631	0.570
电影	0.449	**0.579**	0.448	0.463	0.449	0.511	0.525	0.529
Yelp	0.551	**0.686**	0.552	0.540	0.551	0.556	0.525	0.521
总体	0.523	**0.628**	0.526	0.519	0.524	0.546	0.542	0.533
Qwen3-8B 非思考模式								
MIND-20	0.475	0.486	0.476	0.472	0.492	**0.518**	0.462	0.441
图书	0.446	0.458	0.449	0.443	0.558	**0.576**	0.509	0.493
电影	0.439	0.426	0.440	0.437	0.478	**0.498**	0.413	0.372
Yelp	0.333	0.427	0.337	0.328	0.487	**0.567**	0.400	0.363
总体	0.436	0.447	0.442	0.432	0.538	**0.559**	0.437	0.381

表3：**不同数据集和骨干模型上的热启动错误 AUROC**。 两个骨干模型的结果均基于相同的 4,000 个热启动实例计算。粗体和下划线分别表示每行中表现最佳和次佳的方法。

方法	错误 AUPRC ↑	风险@80 ↓	AURC ↓	E-AURC ↓
Llama-3.1-8B-Instruct				
LabelProb-Max	0.407	0.387	0.373	0.283
LabelProb-Top10Mass	**0.509**	**0.355**	**0.298**	**0.208**
LabelProb-Entropy	0.409	0.388	0.373	0.282
LabelProb-Margin	0.401	0.392	0.371	0.281
温度缩放标签概率	0.408	0.385	0.373	0.283
分解式大语言模型不确定性量化	0.427	0.386	0.355	0.264
扰动排名不稳定性	0.435	0.373	0.372	0.282
自洽性	0.408	0.385	0.365	0.274
Qwen3-8B 非思考模式				
LabelProb-Max	0.321	0.391	0.386	0.308
LabelProb-Top10Mass	0.319	0.409	0.372	0.295
LabelProb-Entropy	0.329	0.389	0.384	0.307
LabelProb-Margin	0.318	0.392	0.389	0.312
温度缩放标签概率	0.393	**0.357**	0.325	0.247
分解式大语言模型不确定性量化	**0.402**	**0.357**	**0.307**	**0.230**
扰动排名不稳定性	0.352	0.369	0.423	0.346
自洽性	0.329	0.383	0.448	0.370

表3：**热启动在互补可靠性指标上的结果**。两个骨干模型均在相同的 4,000 个实例上报告错误 AUPRC、Risk@80、AURC 和 E-AURC。粗体和下划线分别表示每个骨干模型区块中的最佳和次佳方法。

合并后的分数遵循第3.5节。附录A报告了更多实现、覆盖范围和评估器的细节。

4.2 热启动

热启动作为参考条件，原始用户历史记录和项目描述均保持不变。为考察骨干模型敏感性，我们使用非思考模式下的 Llama- 3.1-8B-Instruct 和 Qwen3-8B，对相同的 4,000 个热启动实例进行评估。表3报告了不同数据集和骨干模型上的错误 AUROC，而图 2 左侧面板总结了各方法的总体排序。

观察***1***：不确定性量化方法的排序取决于骨干模型。Top10Mass 在 Llama 上排名第一，但在 Qwen 上降至第三；分解式大语言模型不确定性量化则从第二升至第一，温度缩放从第六升至第二。此外，Top10Mass 与分解式大语言模型不确定性量化之间的错误 AUROC 间隔发生逆转，在 Llama 上为 $+0.082$，在 Qwen 上则为 -0.112。由于评估实例和协议均保持不变，这一逆转表明，不确定性量化监测器的相对效果可能高度依赖骨干模型。若干

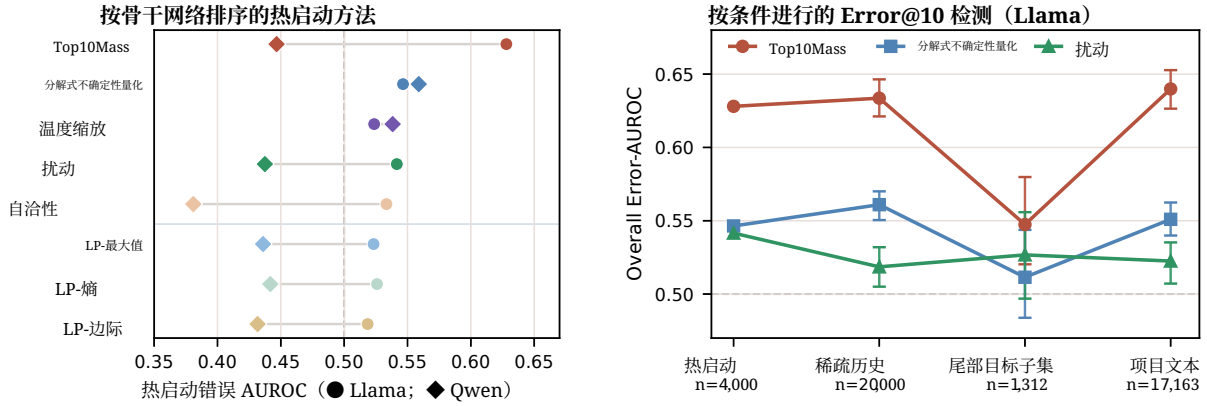

图 2：**特定于骨干模型和条件的排名失败检测**。不同骨干模型上的热启动错误 AUROC，以及不同评估条件下 Llama 的汇总错误 AUROC。

数据集	标签概率 Max	标签概率 Top10Mass	标签概率 熵	标签概率 间隔	温度 缩放	分解 大语言模型不确定性量化	扰动 排名不稳定性	自我-一致性
MIND-20	0.488	0.515	0.485	0.493	0.487	0.484	**0.516**	0.509
图书	0.610	**0.696**	0.608	0.581	0.611	0.630	0.558	0.584
电影	0.455	**0.575**	0.454	0.476	0.455	0.515	0.519	0.533
Yelp	0.596	**0.694**	0.594	0.571	0.596	0.586	0.506	0.549
总体	0.536	**0.634**	0.536	0.532	0.536	0.561	0.519	0.536

表 5：**稀疏历史下的不确定性量化结果**。在 Llama-3.1-8B-Instruct 上，用户历史截断条件下 Error@10 检测的错误 AUROC。

Qwen 监测器的错误 AUROC 值也低于 0.5，这表明在此设置下，其预定义的不确定性分数与 Error@10 呈反向关联。附录 A.7 提供了额外的分数方向以及错误与正确诊断。我们进一步使用互补可靠性指标评估相同的热启动记录。表4 报告了两个骨干模型的错误 AUPRC、Risk@80、AURC 和 E-AURC。附录表 11 进一步检查了监测截断值与概率质量截断值之间的对齐是否与 Error@K 检测相关。

观察 ***2***：在 ***TopKMass*** 系列中，使截断值与 ***Error@K*** 匹配可以提升检测效果。我们比较了 TopKMass 变体，其中截断值为 $K = 5$、10、15，以考察截断对齐情况。Top5Mass、Top10Mass 和 Top15Mass 分别在 Error@5、Error@10 和 Error@15 上表现最佳。与最强的不匹配截断相比，它们的优势分别为 0.022、0.012 和 0.036 AUROC（附录表 11）。在 Llama 热启动中，在相同的单字母分布下，使用已发出的前 10 项集合，相比探针自身的前 10 项集合，可使错误 AUROC 提高 0.103。在原始排序提示下重建的首槽位前 10 项质量分数比跨诱导 Top10Mass 分数低 0.022（附录 A.12）。该已发出集合的比较支持与截断点对齐的监测。

4.3 稀疏历史

稀疏历史通过截断可用的用户历史，同时保持候选集和排序协议不变，来评估不确定性量化监测。表5报告了数据集级别和汇总错误 AUROC，图2右侧面板总结了 Llama 的总体结果。实例聚类置信区间见附录表 12。

观察 ***3***：在稀疏历史下，***Top10Mass*** 仍是最强的汇总监测指标。Top10Mass 达到 0.634 的汇总错误 AUROC，比分解式大语言模型不确定性量化高 0.073，且两者的 95% 实例聚类自助法置信区间不重叠。在图书、电影和 Yelp 上，Top10Mass 同样排名第一。与此同时，在 MIND-20 上，Top10Mass 仅比扰动排名不稳定性低 0.001。由于 Top10Mass 在四个数据集中的三个上领先，并且在第四个数据集上几乎并列第一，因此其汇总

数据集	n	标签概率 Max	标签概率 Top10Mass	标签概率 熵	标签概率 间隔	温度 缩放	分解 大语言模型不确定性量化	扰动 排名不稳定性	自-一致性
MIND-20	870	0.463	0.503	0.477	0.471	0.464	0.483	**0.525**	0.505
图书	221	0.575	**0.643**	0.588	0.560	0.578	0.587	0.598	0.492
电影	129	0.522	0.529	0.513	**0.540**	0.522	0.482	0.411	0.485
Yelp	92	0.651	**0.687**	0.608	0.675	0.647	0.630	0.558	0.585
总体	1,312	0.503	**0.547**	0.510	0.511	0.504	0.511	0.527	0.506

表6：**Llama-3.1-8B-Instruct 上尾部目标子集不确定性量化结果**。在 1,312 条热启动记录上进行 Error@10 检测的错误 AUROC，这些记录的目标项目属于数据集特定的流行度长尾分桶。

数据集	标签概率 Max	标签概率 Top10Mass	标签概率 熵	标签概率 间隔	温度 缩放	分解 大语言模型不确定性量化	扰动 排名不稳定性	自一致性- 一致性
MIND-20	0.488	**0.521**	0.492	0.490	0.488	0.484	0.519	0.502
图书	0.577	**0.663**	0.567	0.566	0.578	0.597	0.586	0.553
电影	0.463	**0.618**	0.452	0.476	0.462	0.509	0.507	0.532
Yelp	0.564	**0.677**	0.554	0.548	0.563	0.559	0.506	0.519
总体	0.537	**0.640**	0.534	0.530	0.537	0.551	0.523	0.523

表 7：**Llama-3.1-8B-Instruct 上的物品文本鲁棒性**。 在受控候选项目文本字段消融下进行 Error@10 检测的错误 AUROC。

这种优势并非由单一领域驱动。这一模式表明，即使可用的用户历史证据减少，与排序对齐的 Top10Mass 信号对于 Error@10 检测仍然具有信息量。

4.4 尾部目标子集

尾部目标子集评估正例物品落入数据集特定的流行度长尾分桶中的热启动记录。其中包含 1,312 条记录，各数据集保留的记录数有所不同。表6 报告了数据集级别和汇总错误 AUROC，图 2 右侧面板总结了汇总结果。置信区间见附录表 12。

观察 ***4***：流行度长尾目标削弱了 ***Top10Mass*** 的热启动优势。Top10Mass 的汇总错误 AUROC 达到 0.547，比其热启动分数低 0.081。其相对于扰动排名不稳定性的汇总领先幅度仅为 0.020，且二者的实例聚类置信区间存在重叠。这些结果表明，Top10Mass 显著的热启动优势并不能普遍延伸到流行度长尾目标。

观察 ***5***：尾部目标不确定性量化的表现因数据集而异。 数据集级别的最佳方法因领域而异。扰动排名不稳定性在 MIND-20 上表现最佳，LabelProb-Margin 在电影上表现最佳，而 Top10Mass 在图书和 Yelp 上仍然表现最佳。因此，解读汇总估计时应结合数据集级别的结果，因为 MIND-20 占保留记录总数 1,312 条中的 870 条。

4.5 项目文本鲁棒性

项目文本鲁棒性用于评估在受控的候选项目文本字段消融条件下的不确定性量化监测，同时保持其余排序协议不变。表 7 报告了数据集级别和汇总错误 AUROC。图 2 右侧面板总结了 Llama 的总体结果。相应的项目内容 NDCG@10 差异则在配对对照分析中单独评估。

观察 ***6***：在项目文本消融条件下，***Top10Mass*** 仍是表现最强的汇总监测指标。Top10Mass 的汇总错误 AUROC 达到 0.640，比分解式大语言模型不确定性量化高 0.089，且两者的 95% 实例聚类自助法置信区间不重叠。

观察 ***7***：***Top10Mass*** 的项目文本鲁棒性在各数据集上表现一致。Top10Mass 在全部四个数据集上均排名第一，分别领先最强替代方法：在图书数据集上领先 0.066，在电影数据集上领先 0.086，在 Yelp 上领先 0.113。MIND-20 是唯一差距接近的情况，其领先扰动排名不稳定性 0.002。这一模式表明，汇总后的优势得到多个领域的支持，并非由某个表现出色的单一数据集所驱动，

方法	信号	方向覆盖					有符号效应 D-Acc	
分解式大语言模型不确定性量化	标量响应	0.487	0.267	0.328	0.521	0.071		**0.548**
LabelProb-Max	标量响应	0.482	0.262	0.331	0.520	0.054	0.570	0.535
LabelProb-Top10Mass	标量响应	0.656	0.374	0.447	0.531	0.092	0.571	0.531
LabelProb-Entropy	标量响应	0.478	0.250	0.301	0.517	0.060	0.571	0.539
LabelProb-Margin	标量响应	0.500	0.268	0.347	0.518	0.040	0.551	0.527
温度缩放标签概率	标量响应	0.479	0.263	0.332	0.521	0.056	0.572	0.535
自洽性	标量响应	0.619	0.316	0.377	0.503	-0.006	0.529	0.506
扰动排名不稳定性	源评分	**0.987**	**0.709**	**0.712**	**0.675**	**0.456**	0.482	0.496
前五名杰卡德距离	排名距离	0.977	0.700	0.706	0.671	0.440	0.475	0.498
Kendall距离	排名距离	0.997	0.725	0.725	0.692	0.498	0.482	0.504
RBO 距离（$p = 0.9$）	排名距离	1.000	0.746	0.744	0.692	0.494	0.497	0.505
多数类	参考	1.000	0.502	0.334	0.500	–	0.575	0.500

表 8：**Llama-3.1-8B-Instruct 上的配对控制响应**。裁剪指标使用 1,819 个无并列案例；有符号指标使用 3,617 个案例。粗体表示每列中最佳的不确定性量化信号。

尽管优势幅度因领域而异，但在受控的项目文本消融设置下，优势方向始终一致。

4.6 配对对照响应

配对控制评估比较用户历史对照中的完整历史记录与截断历史记录，以及物品内容对照中的原始项目文本与消融后的项目文本。我们评估了 1,819 个截断 NDCG@10 差值不相等的案例。表 8 报告了全部五项一致性指标。此外，表 8 还报告了有符号效应一致性，以及简单的非不确定性量化排序距离基线。附录 A.6 提供了更多可视化结果和敏感性分析。如第 3.3 节所定义，本分析仅限于这两个构造的对照组。

观察***8***：配对对照一致性反映了来源特定的排序敏感性。在全部五项裁剪响应指标上，扰动排名不稳定性均领先于不确定性量化信号。与 Top10Mass（最强的标量响应）相比，其方向覆盖率提高了 0.331，来源 AUROC 提高了 0.144。当将评估限制在两个截断 NDCG@10 差值之间的绝对差距大于 0.10 的案例时，主导控制准确率提高了 0.066，来源 AUROC 提高了 0.064，而方向覆盖率仅提高了 0.002（表 13）。扰动排名不稳定性使用了定义配对控制标签的相同用户历史对照和物品内容对照。这种对齐可能使该方法在此任务上占优。标量不确定性量化方法根据单个不确定性分数的变化推导两种响应。当将分析限制在两个对照的 NDCG@10 差值均严格为正的案例时，同样的定性结论仍然成立。与扰动排名不稳定性相比，Kendall 和 RBO 各自将裁剪后的源 AUROC 提高了 0.017，支持来源特定的排序敏感性。对于有符号效应，排序距离信号在来源 AUROC 上仍接近随机水平。最佳标量信号比多数类基准高出 0.004 的 D-Acc。

4.7 案例研究与诊断检查

► **案例示例**。 图 3 将亚马逊图书的一个示例中的标量误差监测与配对对照响应联系起来。移除类别元数据后，目标项移出固定的前 10 名排序，NDCG@10 下降 0.631，Top10Mass 不确定性上升 0.676。物品内容的扰动响应比用户历史的响应高 0.583。物品内容的 NDCG@10 差异比用户历史的差异高 0.475。因此，该案例具体展示了观察8中的整体配对对照模式：较强的扰动响应与截断 NDCG@10 差值较大的来源相符。

► **合理性与组成部分检查**。 表14和表15报告了对全部42,475条符合条件的Llama记录进行的合理性与组成部分检查。随机评分和解析失败指标仍接近随机水平，而预言器达到1.000。Top10Mass在错误AUROC上比LabelProb-BaseTop1高出0.050。分解提示和自洽性 top-1 组成部分仍接近随机水平。

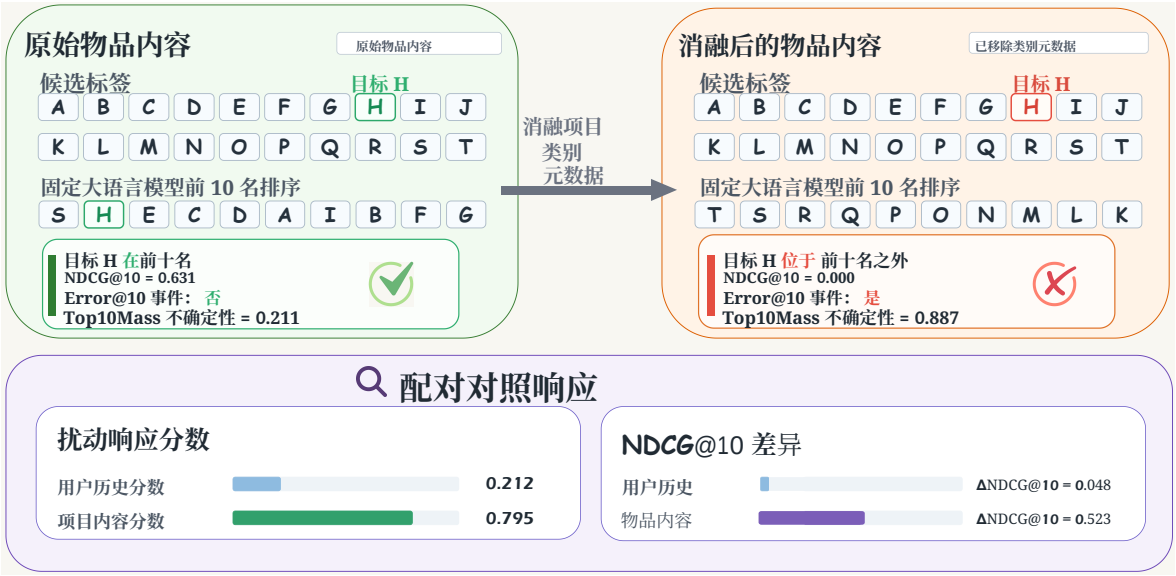

图 3：**物品内容案例**。 移除类别元数据会使目标移出固定的前十名；对于物品内容，扰动响应和 NDCG@10 差异都更大。

5 结论

我们提出了一个用于基于大语言模型的推荐中不确定性量化的固定排序基准测试，其中不确定性量化方法监测相同的已生成排序，而不改变候选项顺序。在四个数据集上，该基准测试评估排序失败检测、不同骨干模型和评估条件下不确定性量化方法性能的稳定性，以及针对用户历史和物品内容证据的配对对照响应一致性。结果表明，在所有设置下，没有一种不确定性量化排序能够保持一致。在若干Llama条件下，Top10 Mass实现了最强的汇总 Error@10 检测性能，体现出有效的与截断点对齐的监测能力；但在尾部目标子集上，其优势有所减小。在Qwen上进行相同的热启动评估会产生不同的方法排序。对于配对控制，扰动排名不稳定性与相对截断 NDCG@10 差值的一致性强于标量不确定性量化信号；简单的列表级距离也与这些截断差值高度一致，从而支持来源特定诊断。这些结果表明，不确定性量化质量具有任务依赖性。对排序失败检测有效的标量监测器，未必是受控证据响应一致性中信息量最大的监测器。实际监测结论仅限于针对单个已观测正例物品的 Error@10。经目标骨干模型和推荐条件验证的不确定性量化监测器，可以将高不确定性请求路由至备用推荐系统。留出数据上的风险–覆盖率曲线可以指导阈值选择。Risk@80和AURC描述的是保留的大语言模型输出中的排序错误。对完整路由策略进行在线 A/B 测试，可以衡量点击率和转化率。通过固定已生成的排序，并使所有方法对齐到相同的记录，我们的基准测试能够直接比较这些差异，而不会将监测质量与重排序行为混为一谈。这种分离为在不同推荐条件下评估不确定性量化方法提供了共同基础。

6 致谢

本研究部分得到了 NSF 2616633、NSF 2303037、NSF 2312501、NSF 2531008、NSF 2553619、NSF 2617441、NSF 2609582、美国国防高级研究计划局 HR001126CE054、美国国立卫生研究院 OT2OD038003、美国国立卫生研究院 U54OD036472、美国国立卫生研究院 U54DK097771、Optum、NEC、SRC JUMP 2.0 中心、CDSC 行业合作伙伴、亚马逊研究奖、Snapchat 和 Google Gifts 的部分资助。本研究的计算资源部分由北卡罗来纳大学教堂山分校通过其人工智能加速计划提供支持。Hang Zhou 部分获得了北卡罗来纳大学教堂山分校 SDSS 计算数据科学与人工智能云试点计划的支持。

参考

Himan Abdollahpouri、Robin Burke 和 Bamshad Mobasher。通过个性化重排序管理推荐系统中的流行度偏差。*arXiv*预印本 *arXiv:1901.07555*，2019年。Gediminas Adomavicius 和 Alexander Tuzhilin。迈向下一代推荐系统：现状综述及可能的扩展。*IEEE*《知识与数据工程汇刊》，17(6):734–749，2005年。Anastasios N Angelopoulos、Karl Krauth、Stephen Bates、Yixin Wang 和 Michael I Jordan。具有无分布可靠性保证的推荐系统。载于面向应用的保形预测与概率预测，第175–193页。PMLR，2023年。Carles Balsells-Rodas、Fan Yang、Zhishen Huang 和 Yan Gao。序列推荐中的可解释不确定性归因。载于第*47*届国际 *ACM SIGIR* 信息检索研究与开发会议论文集，第2401–2405页，2024年。Keqin Bao、Jizhi Zhang、Yang Zhang、Wenjie Wang、Fuli Feng 和 Xiangnan He。Tallrec：一种有效且高效的调优框架，用于使大语言模型与推荐相匹配。载于第*17*届 *ACM* 推荐系统会议论文集，第1007–1014页，2023年。Keqin Bao、Jizhi Zhang、Wenjie Wang、Yang Zhang、Zhengyi Yang、Yanchen Luo、Chong Chen、Fuli Feng 和 Qi Tian。用于推荐系统中大语言模型的双步落地范式。*ACM*《推荐系统汇刊》，3(4):1–27，2025年。Emily M Bender 和 Batya Friedman。自然语言处理的数据声明：迈向缓解系统偏差并推动更好的科学研究。《计算语言学协会汇刊》，6:587–604，2018年。Chenxiao Fan、Chongming Gao、Yaxin Gong、Haoyan Liu、Fuli Feng 和 Xiangnan He。不确定性感知生成式推荐。*arXiv*预印本 *arXiv:2602.11719*，2026年。Tom Fawcett。ROC分析导论。模式识别快报，27(8):861–874，2006年。Jakob Gawlikowski、Cedrique Rovile Njieutcheu Tassi、Mohsin Ali、Jongseok Lee、Matthias Humt、Jianxiang Feng、Anna Kruspe、Rudolph Triebel、Peter Jung、Ribana Roscher 等人。深度神经网络不确定性综述。人工智能评论，56(Suppl 1):1513–1589，2023年。Timnit Gebru、Jamie Morgenstern、Briana Vecchione、Jennifer Wortman Vaughan、Hanna Wallach、Hal Daumé III 和 Kate Crawford。数据集数据表。《*ACM*通讯》，64(12):86–92，2021年。Yonatan Geifman 和 Ran El-Yaniv。深度神经网络的选择性分类。神经信息处理系统进展，30，2017年。Shijie Geng、Shuchang Liu、Zuohui Fu、Yingqiang Ge 和 Yongfeng Zhang。推荐即语言处理（RLP）：统一的预训练、个性化提示与预测范式（P5）。载于第*16*届 *ACM* 推荐系统会议论文集，第299–315页，2022年。Aaron Grattafiori、Abhimanyu Dubey、Abhinav Jauhri、Abhinav Pandey、Abhishek Kadian、Ahmad Al-Dahle、Aiesha Letman、Akhil Mathur、Alan Schelten、Alex Vaughan 等人。Llama 3 模型家族。*arXiv*预印本 *arXiv:2407.21783*，2024年。Chuan Guo、Geoff Pleiss、Yu Sun 和 Kilian Q Weinberger。现代神经网络的校准。载于国际机器学习会议，第1321–1330页。PMLR，2017年。Dan Hendrycks 和 Kevin Gimpel。用于检测神经网络中错误分类样本和分布外样本的基线方法。*arXiv*预印本 *arXiv:1610.02136*，2016年。Yupeng Hou、Jiacheng Li、Zhankui He、An Yan、Xiusi Chen 和 Julian McAuley。连接语言与物品以实现检索和推荐。*arXiv*预印本 *arXiv:2403.03952*，2024a。

Yupeng Hou、Junjie Zhang、Zihan Lin、Hongyu Lu、Ruobing Xie、Julian McAuley 和 Wayne Xin Zhao。大语言模型是推荐系统的零样本排序器。载于欧洲信息检索会议，第 364–381 页。Springer，2024b。Kalervo Järvelin 和 Jaana Kekäläinen。基于累积增益的信息检索技术评估。*ACM信息系统汇刊 (TOIS)*，20(4):422–446，2002。Jianchao Ji、Zelong Li、Shuyuan Xu、Wenyue Hua、Yingqiang Ge、Juntao Tan 和 Yongfeng Zhang。Genrec：用于生成式推荐的大语言模型。载于欧洲信息检索会议，第 494–502 页。Springer，2024。Chumeng Jiang、Jiayin Wang、Weizhi Ma、Charles L. A. Clarke、Shuai Wang、Chuhan Wu 和 Min Zhang。超越效用：评估大语言模型作为推荐系统。载于*ACM Web*会议*2025*论文集。ACM，2025。doi: 10.1145/3696410.3714759。Saurav Kadavath、Tom Conerly、Amanda Askell、Tom Henighan、Dawn Drain、Ethan Perez、Nicholas Schiefer、Zac Hatfield-Dodds、Nova DasSarma、Eli Tran-Johnson 等。语言模型（大多）知道自己知道什么。*arXiv*预印本 *arXiv:2207.05221*，2022。Lorenz Kuhn、Yarin Gal 和 Sebastian Farquhar。语义不确定性：用于自然语言生成不确定性估计的语言不变性。*arXiv*预印本 *arXiv:2302.09664*，2023。Wonbin Kweon、Sanghwan Jang、SeongKu Kang 和 Hwanjo Yu。基于大语言模型的推荐中的不确定性量化与分解。载于*ACM Web*会议*2025*论文集，第 4889–4901 页，2025。Quentin Lhoest、Albert Villanova Del Moral、Yacine Jernite、Abhishek Thakur、Patrick Von Platen、Suraj Patil、Julien Chaumond、Mariama Drame、Julien Plu、Lewis Tunstall 等。Datasets：一个自然语言处理社区库。载于*2021*年计算语言学经验方法会议论文集：系统演示，第 175–184 页，2021。Jiayu Li、Ziyi Ye、Guohao Jian、Zhiqiang Guo、Weizhi Ma、Qingyao Ai 和 Min Zhang。推荐系统是否具有自我认知？通过模型不确定性进行无标签的推荐性能估计。*arXiv*预印本 *arXiv:2507.23208*，2025。Sichun Luo、Bowei He、Haohan Zhao、Wei Shao、Yanlin Qi、Yinya Huang、Aojun Zhou、Yuxuan Yao、Zongpeng Li、Yuanzhang Xiao 等。RecRanker：通过指令微调将大语言模型用作 Top-k推荐排序器。*ACM*信息系统汇刊，43(5):1–31，2025。Qing Lyu、Kumar Shridhar、Chaitanya Malaviya、Li Zhang、Yanai Elazar、Niket Tandon、Marianna Apidianaki、Mrinmaya Sachan 和 Chris Callison-Burch。利用样本一致性校准大语言模型。载于*AAAI*人工智能会议论文集，第 39 卷，第 19260–19268 页，2025。Sejoon Oh、Berk Ustun、Julian McAuley 和 Srijan Kumar。推荐系统的排序列表对交互扰动的敏感性。载于第*31*届*ACM*信息与知识管理国际会议论文集，第 1584–1594 页，2022。Charul Paliwal、Anirban Majumder 和 Sivaramakrishnan Kaveri。推荐系统的预测相关性不确定性。载于*ACM Web*会议*2024*论文集，第 3900–3909 页，2024。Joelle Pineau、Philippe Vincent-Lamarre、Koustuv Sinha、Vincent Larivière、Alina Beygelzimer、Florence d’Alché Buc、Emily Fox 和 Hugo Larochelle。提升机器学习研究的可复现性（NeurIPS 2019 可复现性计划报告）。机器学习研究期刊，22(164):1–20，2021。Qwen 团队。Qwen3技术报告。*arXiv*预印本 *arXiv:2505.09388*，2025。Marco Tulio Ribeiro、Sameer Singh 和 Carlos Guestrin。“为什么我应该相信你？”解释任意分类器的预测。载于第*22*届*ACM SIGKDD*知识发现与数据挖掘国际会议论文集，第 1135–1144 页，2016。

Marco Tulio Ribeiro、Tongshuang Wu、Carlos Guestrin 和 Sameer Singh。超越准确率：使用 CheckList 对自然语言处理模型进行行为测试。载于第*58*届计算语言学协会年会论文集，第 4902–4912 页，2020。

Angelika Romanou、Mark Ibrahim、Candace Ross、Chantal Shaib、Kerem Oktar、Samuel J Bell、Anaelia Ovalle、Jesse Dodge、Antoine Bosselut、Koustuv Sinha 等。Brittlebench：通过提示敏感性量化大语言模型鲁棒性。*arXiv* 预印本 *arXiv:2603.13285*，2026年。

Chandan Kumar Sah、Xiaoli Lian、Li Zhang、Tony Xu 和 Syed Shazaib Shah。基于大语言模型的推荐系统中的不确定性与公平性意识。*arXiv* 预印本 *arXiv:2602.02582*，2026年。

Tobias Scheffer、Christian Decomain 和 Stefan Wrobel。用于信息抽取的主动隐马尔可夫模型。载于智能数据分析国际研讨会，第309–318页。Springer，2001年。

Burr Settles。主动学习文献综述。2009年。

Claude Elwood Shannon。通信的数学理论。贝尔系统技术期刊，27（3）：379–423，1948。

Zichen Tang、Zirui Zhang、Qian Wang、Zhenheng Tang、Bo Li 和 Xiaowen Chu。你的大语言模型推荐系统代理值得信赖吗？大语言模型的推荐很容易受到偏差（偏好）的攻击。*arXiv* 预印本 *arXiv:2603.17417*，2026。

Xuezhi Wang、Jason Wei、Dale Schuurmans、Quoc Le、Ed Chi、Sharan Narang、Aakanksha Chowdhery 和 Denny Zhou。自洽性提升语言模型中的思维链推理。*arXiv* 预印本 *arXiv:2203.11171*，2022。

Fangzhao Wu、Ying Qiao、Jiun-Hung Chen、Chuhan Wu、Tao Qi、Jianxun Lian、Danyang Liu、Xing Xie、Jianfeng Gao、Winnie Wu 等。MIND：用于新闻推荐的大规模数据集。载于第*58*届计算语言学协会年会论文集，第3597–3606页，2020年。

Quan Xiao、Debarun Bhattacharjya、Balaji Ganesan、Radu Marinescu、Katsiaryna Mirylenka、Nhan H Pham、Michael Glass 和 Junkyu Lee。不确定性量化中大语言模型的一致性假设。*arXiv* 预印本 *arXiv:2506.21849*，2025年。

Cai Xu、Xujing Wang、Ziyu Guan、Wei Zhao 和 Meng Yan。GUIDER：面向基于大语言模型的推荐系统的不确定性引导的动态重排序。载于 *AAAI*人工智能会议论文集，第40卷，第16049–16057页，2026年。

Yelp 公司。Yelp 开放数据集。https://business.yelp.com/data/resources/open-dataset/。访问日期：2026年5月26日。

Zhenrui Yue、Sara Rabhi、Gabriel de Souza Pereira Moreira、Dong Wang 和 Even Oldridge。LlamaRec：使用大语言模型进行排序的两阶段推荐。*arXiv* 预印本 *arXiv:2311.02089*，2023年。

Kaike Zhang、Qi Cao、Fei Sun、Yunfan Wu、Shuchang Tao、Huawei Shen 和 Xueqi Cheng。鲁棒推荐系统：综述与未来方向。*ACM*计算调查，58(1):1–38，2025年。

可复现性与制品详情

本附录介绍复现第3节所定义的固定排序基准所需的实现细节。我们将说明制品契约、骨干模型与采样设置、UQ分数定义、配对控制响应构建、覆盖率检查以及补充诊断。

字段	描述
模式 –版本	工件模式版本。
记录_ID	基础记录和不确定性量化行的稳定连接键。
方法	不确定性量化方法标识符。
方法_类别	概率、校准、分解、敏感性或采样。
不确定性_分数	标量分数；数值越大表示越不确定。
置信度_分数	可选的置信度对应项。
来源_分数	可选的配对对照用户/项目响应分数。
组件	可选的特定方法的诊断值。
元数据	可选的运行设置和解析器状态。

Table 9：**不确定性量化产物模式**。 与记录对齐的不确定性量化输出所需的字段和可选字段ts.

A.1 工件与 EvaluatorContract

该基准将固定的基础推荐输出与不确定性量化输出分开存储。基础预测表包含解析后的20候选项排序、记录级推荐指标、解析状态，以及二元标签为_error_at_10。每个不确定性量化制品都通过record_id与相应的基础记录对齐。Error@10 标签仅供评估器使用，不用于构建UQ分数。

表 9 总结了共享的不确定性量化产物模式。

标量评估器将每个不确定性量化产物与固定的基础记录表连接起来，并报告用于主要 Error@10 分析的错误 AUROC。它还会生成错误 AUPRC、Risk@80、AURC 和 E-AURC，用于补充性的可靠性分析。结果可以按方法、评估条件和数据集进行汇总。成对对照评估器使用第 3.3 节定义的操作性标签，计算方向覆盖率、主导控制准确率、宏平均 F1、来源 AUROC 和斯皮尔曼秩相关系数。附录 A.6–A.8 中的诊断基于现有工件运行，不需要额外的 LLM 推理。

A.2 骨干模型与采样配置

Llama-3.1-8B-Instruct 提供完整的评估套件。Qwen3-8B 在共享的 4,000 条热启动记录上以非思考模式进行评估。本次复现中的 Qwen 生成是确定性的。模型标识符、可用的版本信息、数值精度、随机种子和生成参数均存储在运行元数据中。扰动排名不稳定性可能使用辅助排序作为监控输入；这些辅助排序不构成额外报告的评估条件。自洽性为每条记录以温度 0.9 和 top-p 0.95 抽取五个随机排名。Llama 的采样排序输出的平均解析成功率为 0.995。由于自洽性需要重复生成，其推理成本高于单次通过的标签概率信号。

A.3 不确定性量化信号定义详解

设记录 i 包含 $K = 20$ 个具有有效选项标签的候选项

$$\mathcal{A}_i = \{A, \ldots, T\}. \tag{6}$$

所有标量不确定性量化分数的方向均设为数值越大表示不确定性越高。

选项标签分布。 对于基于概率的方法，我们使用单字母选择提示，并读取有效选项标签对应的下一个词元的对数几率。令 $\mathcal{T}(a)$ 表示标签 a 支持的词元形式，包括不带空格和带前导空格的形式。我们定义

$$z_i(a) = \max_{t \in \mathcal{T}(a)} \text{logit}_i(t), \tag{7}$$

其中， $z_i(a)$ 是选择支持度最高的词元形式后分配给选项标签 a 的逻辑值。然后，我们对有效标签进行归一化：

$$p_i(a) = \frac{\exp z_i(a)}{\sum_{b\in\mathcal{A}_i} \exp z_i(b)}. \tag{8}$$

因此， $p_i(a)$ 是 LabelProb 系列所使用的选项标签概率。

概率评分。LabelProb-Max 使用

$$u_i^{\max} = 1 - \max_{a\in\mathcal{A}_i} p_i(a). \tag{9}$$

当没有任何单个有效选项标签获得较高概率时，该分数较高。
设 $\mathcal{A}_i^{(10)}$ 表示与固定候选集 $B_i^{(10)}$ 对应的选项标签。LabelProb-Top10Mass 使用

$$u_i^{\text{top10}} = 1 - \sum_{a\in\mathcal{A}_i^{(10)}} p_i(a). \tag{10}$$

该分数衡量固定基础前十集合之外的概率质量。它不使用相关性标签，也不会修改基础排序。

LabelProb-Entropy 使用归一化熵，

$$u_i^{\text{ent}} = -\frac{1}{\log K} \sum_{a\in\mathcal{A}_i} p_i(a) \log p_i(a), \tag{11}$$

当概率质量在有效标签之间更加分散时，该值会增大。LabelProb-Margin 使用反向的前二名间隔，

$$u_i^{\text{margin}} = 1 - \left(p_{i,(1)} - p_{i,(2)}\right), \tag{12}$$

其中 $p_{i,\ (1)}$ 和 $p_{i,\ (2)}$ 分别是最大的和第二大的选项标签概率。

温度缩放标签概率。 温度缩放会对选项标签分布应用一个标量温度 $T > 0$ ：

$$p_i^{(T)}(a) = \frac{p_i(a)^{1/T}}{\sum_{b\in\mathcal{A}_i} p_i(b)^{1/T}}. \tag{13}$$

我们使用正选项标签 a_i^+ 通过交叉验证拟合 T：

$$\mathcal{L}(T) = -\sum_i \log p_i^{(T)}(a_i^+). \tag{14}$$

校准不确定性评分为

$$u_i^{\text{temp}} = 1 - \max_{a\in\mathcal{A}_i} p_i^{(T)}(a). \tag{15}$$

按基础实例标识符划分为五个折叠，以避免相关变体之间的数据泄漏。Llama拟合得到的各折叠温度为1.199–1.245，Qwen为8.563–8.717；各折叠的特定值存储在运行元数据中。表10报告了在两个骨干模型上采用相同三个固定温度时的四个LabelProb分数。Top10Mass–分解式大语言模型不确定性量化反转在测试温度范围内依然存在。在$T = 8.62$下，相较于$T = 1$，Qwen的LabelProb-Max和LabelProb-Entropy在错误 AUROC上分别提升0.102和0.157。Llama的分数在这些设置下变化不大。

分解式大语言模型不确定性量化。 基于分解的大语言模型推荐不确定性量化，我们创建 $M = 5$个候选项顺序变体，并将其选项标签分布对齐回原始候选项。令 $p_{im}(c)$表示在提示变体 m下、关于原始候选项 c 的对齐分布，并定义

$$\bar{p}_i(c) = \frac{1}{M} \sum_{m=1}^{M} p_{im}(c). \tag{16}$$

	Llama-3.1-8B-Instruct			Qwen3-8B 非思考模式		
分数	T=1	T=1.23	T=8.62	T=1	T=1.23	T=8.62
LabelProb-Max	0.523	0.524	0.525	0.436	0.441	0.538
LabelProb-Entropy	0.526	0.526	0.526	0.442	0.451	0.599
LabelProb-Margin	0.519	0.519	0.519	0.432	0.434	0.458
LabelProb-Top10Mass	0.628	0.627	0.623	0.447	0.448	0.444
分解式大语言模型不确定性量化		0.546			0.559	

表10：**共享固定温度下的热启动错误 AUROC**。 相同的 T 应用于两个骨干模型。 $T = 1.23$ 和 $T = 8.62$ 分别表示针对 Llama 和 Qwen 拟合得到的平均温度。分解式大语言模型不确定性量化作为未改变的参考进行报告。

使用归一化熵 $H(\cdot)$，我们计算

$$U_i^{\text{total}} = H(\bar{p}_i), \tag{17}$$

$$U_i^{\text{rec}} = \frac{1}{M}\sum_{m=1}^{M} H(p_{im}), \tag{18}$$

$$U_i^{\text{prompt}} = U_i^{\text{total}} - U_i^{\text{rec}}. \tag{19}$$

主要基准测试分数为

$$u_i^{\text{decomp}} = U_i^{\text{total}}. \tag{20}$$

在表 15 的组件诊断中，保留推荐和提示组件。

扰动排名不稳定性。 对于控制条件 q，令 $R_i^{\text{rest}}{}_{,\,q}$ 表示参考排序，令$R_i^{\text{pert}}{}_{,\,q}$ 表示相应的扰动后排名。令 $T_k(R)$表示排序 R 的前-k 候选集。我们将前-k 杰卡德不稳定性定义为

$$d_{i,q} = 1 - \frac{\left|T_k(R_{i,q}^{\text{rest}}) \cap T_k(R_{i,q}^{\text{pert}})\right|}{\left|T_k(R_{i,q}^{\text{rest}}) \cup T_k(R_{i,q}^{\text{pert}})\right|}. \tag{21}$$

$d_{i,\,q}$ 越大，表示在控制条件 q 下，前-k 推荐集的变化越大。我们使用 $k = 10$，并将标量扰动分数定义为可用的特定控制条件下不稳定性的平均值，

$$u_i^{\text{pert}} = \frac{1}{|\mathcal{Q}_i|}\sum_{q\in\mathcal{Q}_i} d_{i,q}, \tag{22}$$

其中，$\mathcal{Q}_i$ 是记录 i 的可用成对控制条件集合。产物字段 source_scores 存储了成对控制评估所使用的独立用户历史和物品内容响应。

自一致性排名分歧。 基础排序保持不变，同时该方法抽取S 个额外的随机排序 $\{R_{is}\}_{s=1}^{S}$。自一致性不确定性是抽样排序之间两两前-k 杰卡德分歧的平均值：

$$u_i^{\text{sc}} = \binom{S}{2}^{-1}\sum_{s<t}\left(1 - \frac{|T_k(R_{is}) \cap T_k(R_{it})|}{|T_k(R_{is}) \cup T_k(R_{it})|}\right). \tag{23}$$

我们使用 $k = 10$ 和 $S = 5$。数值越大，表示采样得到的前十项集合之间的一致性越低。样本分歧的两两评分用于衡量模型排名分布的离散程度。基准锚定分歧用于衡量已发布排序在该分布中的非典型程度。附录 A.13 评估基础排序与五个样本中每个样本之间的平均前 10 个 Jaccard 距离，以及这五个样本与基础排序之间的两两评分。

分数	Error@5	Error@10	Error@15
Top5Mass	**0.656**	0.609	0.583
Top10Mass	0.634	**0.628**	0.592
Top15Mass	0.624	0.616	**0.628**

表11：**热启动 Error@K/TopKMass 对齐**。共享的 4,000 条 Llama 热启动记录上的错误 AUROC。

A.4 配对控制响应构造

对于仅标量方法 m，配对控制响应分数源自每个对照组下不确定性的增加：

$$r_{i,u}^{(m)} = \left[\bar{u}_{i,\text{trunc}}^{(m)} - u_{i,\text{full}}^{(m)}\right]_+, \tag{24}$$

$$r_{i,v}^{(m)} = \left[\bar{u}_{i,\text{abl}}^{(m)} - u_{i,\text{orig}}^{(m)}\right]_+. \tag{25}$$

这里，$[x]_+ = \max(x,0)$ 以及柱状图表示对可用变体取平均值。因此，$r_{i\,u}^{(m)}$,衡量截断用户历史后方法-m 不确定性的增加，而 $r_{i\,v}^{(m)}$,衡量消融物品内容后的增加。

提供显式有限用户历史和物品内容来源_分数的方法直接使用这些值。操作性主导控制标签 d_i 和方法预测 $\hat{}$ $d_i^{(m)}$ 分别定义于公式（4）和（5）中。方法响应相同时，预测为平局；对于截断 NDCG@10 差值相同的情况，则不纳入主导控制评估。该解释仅限于用户历史控制和物品内容控制内的一致性。

A.5 覆盖率与健康检查

每个经评估的 Llama UQ工件都包含全部 45,163 行生成的排序，连接覆盖率为 100%。将流行度尾部条件限制为 1,312 条符合条件的尾部目标记录后，最终评估中保留 42,475 条记录：4,000 条热启动记录、20,000 条稀疏历史记录、17,163 条物品文本鲁棒性记录，以及 1,312 条尾部目标记录。每个 Qwen复现实验工件都包含相同的 4,000 行热启动记录。

所有经评估的工件均报告未匹配记录_标识符为零，模式错误为零。基础记录使用唯一标识符。格式错误的生成排序会以确定性方式修复：通过追加缺失标签补全了 337 条 Llama 记录；两条 Qwen Warm-start 输出在重试失败后，则回退到原始候选项顺序。

A.6 补充鲁棒性检查

匹配的Error@K/TopKMass。 表11比较了共享的 4,000 条 Llama 热启动记录上的 Top5Mass、Top10Mass 和 Top15Mass 与 Error@5、Error@10 和 Error@15。在 TopKMass 系列中，对于每个受监测的 Error@K 事件，截止点对齐分数均为最高。相较于最强的不匹配截止点，对齐分数在 Error@5 上使 AUROC 提高 0.022，在 Error@10 上提高 0.012，在 Error@15 上提高 0.036。

实例聚类置信区间。稀疏历史和物品文本鲁棒性包含源自同一基础实例的多个变体。因此，其自助法重采样以数据集×实例_id簇为采样单位，并在方法之间进行配对。尾部目标子集为每个实例保留一条记录。表12报告了条件级分析中使用的200次重采样置信区间。

NDCG@10差异敏感性。 表13将成对控制评估限制在截断用户历史与物品内容NDCG@10差异相差超过所述阈值的案例中。随着

条件	记录	簇数	Top10Mass AUROC [95% CI]	最佳替代方案 [95% 置信区间]
稀疏历史	20,000	4,000	0.634 [0.621，0.646]	分解：0.561 [0.550，0.570]
尾部目标子集	1,312	1,312	0.547 [0.520，0.580]	扰动：0.527 [0.497，0.556]
物品文本鲁棒性	17,163	4,000	0.640 [0.626, 0.653]	分解：0.551 [0.540, 0.562]

表12：**实例聚类配对自助法区间**。聚合 Llama 错误 AUROC 在非参考评估条件下的 95% 置信区间。

Gap	n	方向覆盖率	D-Acc 来源-AUROC		ρ
0.00	1,819	0.987	0.709	0.675	0.456
0.05	1,522	0.988	0.739	0.701	0.477
0.10	1,124	0.989	0.775	0.739	0.498

表 13： **NDCG@10差值间隔阈值下的扰动响应** **s.**

最小差距增加，扰动排名不稳定性仍保持近乎完整的方向覆盖率，同时其主导控制准确率、来源 AUROC和排序相关性均有所提高。

在 0.10 的阈值下，相较于完整的非平局集合，来源 AUROC 提高了 0.064，主导控制准确率提高了 0.066，排序相关性提高了 0.042，而方向覆盖率仅变化了 0.002。该敏感性分析仍仅针对两个构造的对照组。由于截断可能会将负的 NDCG@10 差异设为零，我们进一步将分析限制在截断前用户历史和物品内容 NDCG@10 差异均为正的 863 个案例中。扰动排名不稳定性仍是最强的方法，其主导控制准确率为 0.714，来源 AUROC 为 0.671；在完整的非平局集合中，这两个数值分别为 0.709 和 0.675。该正-正比较涵盖表 8 中的八种不确定性量化信号。

A.7 合理性检查

我们纳入了三个不可部署的参考对象，以验证 Error-AUROC评估器的方向和实现。这些检查不纳入基准排名。

三个参考对象均涵盖 42,475 条符合评估条件的 Llama 记录。随机评分和解析失败指标仍接近随机水平，而预言器证实，较大的不确定性与 Error@10 事件相一致。

低于随机水平的AUROC诊断。 所有不确定性量化分数均保持预定义的方向，即数值越大表示名义不确定性越高。在 Qwen Warm-start 上，多个 Error-AUROC 低于 0.5 的方法，对 Error@10 案例赋予的平均不确定性也低于正确案例。例如，Top10Mass 的平均不确定性分别为 0.0847 和 0.1943%，而 Self-consistency 分别为 0.1849 和 0.2707。相比之下，分解式大语言模型不确定性量化的 Error-AUROC 为 0.559，在 Error@10 案例和正确案例上的平均不确定性分别为 0.5192 和 0.4917。这一模式与以下解释一致：在这些设置中，低于随机水平的 AUROC 值反映了名义不确定性信号与 Error@10 之间的反向关联。对于两个骨干模型，单字母分数使用标签 A–T 的下一个词元的对数几率，以及受支持的裸词元和带空格前缀的词元形式，随后在 20 个标签上进行归一化（附录 A.3）。Qwen 探针在 81.9% 的记录中将最大概率赋予标签 A。相较于已发布集合上的单字母 Top10Mass 分数，使用探针自身的前 10 项集合可使 Error-AUROC 提高 0.121。使用在排序提示下重构的首槽位概率可使其提高 0.071（附录 A.12）。对不同候选项顺序下的探针分布取平均，可使其提高 0.190。这些比较表明，结果对监测集合、概率来源和候选项顺序具有敏感性。

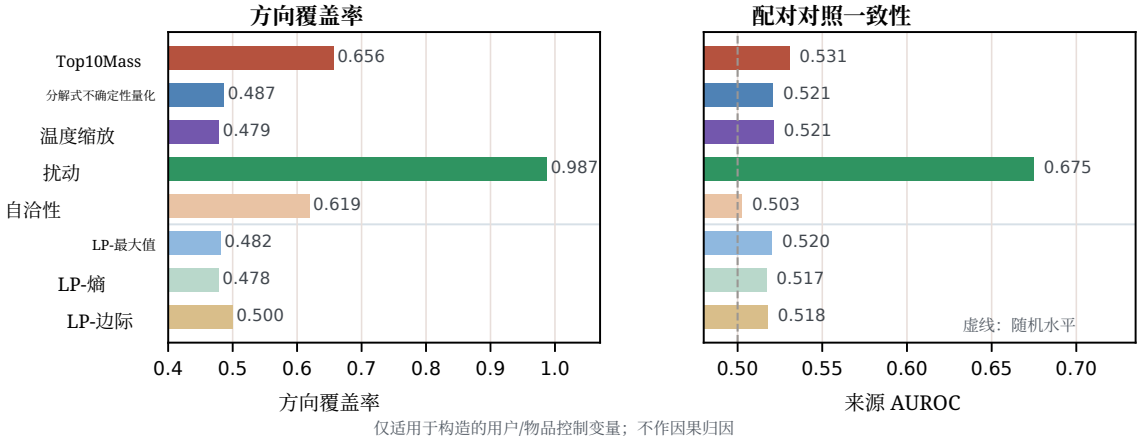

图 4：**Llama-3.1-8B-Instruct 上的配对对照响应一致性**。用户历史和物品内容控制变量的方向性覆盖率和来源 AUROC。

参考	错误 AUROC	覆盖率
随机得分	0.495	1.000
解析失败指标	0.501	1.000
预言机 Error@10	1.000	1.000

表14：**Error-AUROC 的合理性检查**。随机和解析失败参考仍接近随机水平；预言器达到预期上限。

检查	错误 AUROC
LabelProb-BaseTop1	0.584
LabelProb-Top10Mass	0.634
分解 推荐组件	0.542
分解 提示组件	0.483
分解 提示比例	0.478
自一致性 top-1 分歧	0.508
SC top-1 熵	0.509
SC top-1 FSD	0.507

表 15：**推导出的排序对齐与组件检查**。汇总 42,475 条记录的错误 AUROC 符合评估条件的 Llama 记录。

A.8 排序对齐与组件检查

我们在不进行额外 LLM 推理的情况下，从现有工件中推导出组件级诊断。Top10Mass 被纳入作为排序对齐参考项；其余各行分别拆分出基础记录的 top-1 概率、分解组件以及自一致性 top-1 汇总。

在汇总组件评估中，Top10Mass 将错误 AUROC 从基础记录 BaseTop1 的 0.584 提升至 0.634，增益为 0.050。分解式大语言模型不确定性量化中的推荐组件达到 0.542，其提示组件达到 0.478–0.483，而自洽性 top-1 组成部分达到 0.507–0.509。这些数值为主要标量结果提供了背景信息，同时未增加新的 LLM 推理。

A.9 发布的工件与数据卡

匿名化制品包围绕固定的基础预测、记录对齐的不确定性量化输出、评估表、诊断表、清单文件、模式、运行元数据和重建说明进行组织。

（*a*）影响方向：正向／负向／零（%）

数据集	n	用户历史	物品内容
MIND-20	1,000	39.9/44.8/15.3	34.7/33.2/32.1
图书	1,000	27.7/57.9/14.4	39.0/45.2/15.8
电影	1,000	39.0/47.4/13.6	30.7/34.4/34.9
Yelp	1,000	29.1/47.7/23.2	37.1/42.2/20.7
Llama，全部	4,000	33.9/49.5/16.6	35.4/38.8/25.9
Qwen，全部	4,000	28.8/40.0/31.2	28.1/34.6/37.2

(b) 平局和主导来源标签计数

数据集	完全平局	通过截断添加	截断标签（用户/项目）	带符号标签（用户/项目）
MIND-20	97	429	289/185	435/468
图书	87	461	187/265	318/595
电影	66	464	247/223	440/494
Yelp	133	444	183/240	343/524
Llama，全部	383	1,798	906/913	1,536/2,081
Qwen，全部	918	1,582	830/670	1,415/1,667

表1：**带符号的 NDCG@10 差异及截断的影响**。数据集行使用 Llama；最后一行给出汇总的 Qwen 结果。效果方向百分比按每行的全部案例计算，并四舍五入至小数点后一位。标签计数根据相应的标签定义排除平局。

完整的按数据集和条件划分的错误 AUROC 明细以及各方法的配对控制结果，均以机器可读的 CSV 工件形式提供。诊断文件还包括配对控制混淆摘要、逐实例 NDCG@10差异、合理性检查、组件检查和覆盖率报告。

遵循数据文档编制和可复现性实践（Bender 和 Friedman，2018年；Gebru 等人，2021年；Pineau 等人，2021年；Lhoest 等人，2021年），匿名化制品包包含评估器脚本、架构文档、清单文件、运行元数据、重建说明和数据卡备注。数据卡记录了固定排序监控基准测试的源数据集、评估条件、记录数量、预期用途和已知局限。原始条目文本的再分发遵循 MIND、亚马逊和 Yelp 的许可条款。若无法再分发原始文本，则在许可允许的情况下，发布内容将包括标识符、哈希值、排序、分数和重建脚本。

A.10 有符号 NDCG@10差异与排序-响应基准

有符号分析是对公式（2）–（5）中的截断评估以及附录 A.6 中正-正分析的补充。其用户历史和物品内容 NDCG@10差异分别为 $N_{i,\,\text{full}} - N_{i,\,\text{截断}}$ 和 $N_{i,\,\text{orig}} - N_{i,\,\text{消融}}$。有符号标签选择较大的差异。根据评估器的平局约定，差异相等的情况会被排除。负差异表示受控变体的平均 NDCG@10 高于原始输入下的 NDCG@10。

对于有符号评估，标量响应为 $\bar{u}_i^{(m)}{}_{,\,\text{截断}} - u_i^{(m)}{}_{,\,\text{完整}}$ 和 $\bar{u}_i^{(m)}{}_{,\,\text{消融}} - u_i^{(m)}{}_{,\,\text{原始}}$。两种评估中的特定来源排序距离使用相同的定义。响应分数相等时，遵循附录 A.4 中的平局预测约定。多数类基准在每个评估集内预测出现频率更高的来源标签。

表16(a)报告效应符号，面板(b)报告平局和主导来源标签计数。图5展示了有符号效应的分布。在所有数据集和两种骨干模型中，裁剪都保留了每个仍未平局案例的用户/物品标签。它对评估集成员资格的影响是产生了更多平局。有符号的 Llama 标签更常以物品内容为主导。

非 UQ 基线对相同的原始–受控排序对采用前五项Jaccard、归一化 Kendall 或 RBO 距离。前五项Jaccard 使用 1 减去前五项交集与并集之比。归一化

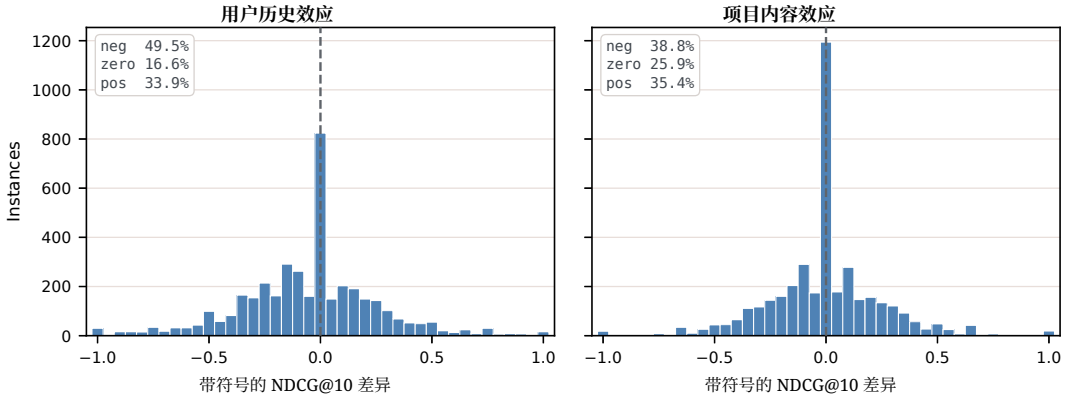

图5：**Llama 上带符号的 NDCG@10 差异分布**。负差异表示与原始输入相比，受控变体的平均 NDCG@10 更高。

(*a*) *Llama* 在各数据集上的带符号标签结果

数据集	n	多数 D-Acc	分解 D-Acc	分解 源 AUROC	PRI 源 AUROC	RBO 源 AUROC
MIND-20	903	0.518	0.525	0.520	0.487	0.507
图书	913	0.652	0.650	0.561	0.497	0.503
电影	934	0.529	0.531	0.526	0.494	0.494
Yelp	867	0.604	0.615	0.558	0.492	0.500

(b) 裁剪后的源*AUROC*与 *Qwen* 检验

信号	Llama ($n = 1,\ 819$)	Qwen ($n = 1,\ 500$)
扰动排名不稳定性	0.675 [0.661, 0.689]	0.634
Kendall距离	0.692 [0.678, 0.707]	0.639
RBO 距离	0.692 [0.677, 0.706]	0.674

表17：**补充配对控制结果**。面板（a）中的 Decomp. 表示采用未裁剪标量响应的分解式大语言模型不确定性量化；PRI 表示扰动排名不稳定性。面板（b）中的括号给出 95% 置信区间。Qwen诊断使用了相同热启动基础实例的辅助控制排名。

Kendall 表示完整排名中不一致项目对所占的比例。RBO 距离使用 1 减去加权前缀重叠比例，其中权重为在深度 p^{d-1} 、 d、 $p = 0.9$ 处的权重，并通过将这些权重在 20 个排名位置上的总和进行归一化。每个源响应都会在该源可用的受控变体上，对其成对距离取平均值。在两个骨干模型上，前十项Jaccard 得出的源评分与扰动排名不稳定性完全相同。

表17(a)报告了 Llama 上数据集特定的带符号结果；面板(b)报告了两个骨干模型上的裁剪后的源AUROC。Llama 的裁剪区间采用按数据集分层的自助法重采样。对于 Llama，扰动排名不稳定性的带符号的源 AUROC 的 95% 置信区间为 $[0.484, 0.509]$。排序距离显示出较强的裁剪后一致性，以及接近随机水平的带符号源判别。数据集特定的 PRI 和 RBO 结果也呈现出后一种模式。分解式大语言模型不确定性量化的 D-Acc 在这些数据集上始终接近多数类基准。这些比较刻画了排序对两个受控证据源的敏感性。

A.11 证据移除与鲁棒性设置详情

位置性输出遵循候选项顺序或其逆序。它们占 Llama 热启动排名的 46.1%（7.1% 为正序，39.0% 为逆序），占 Qwen 热启动排名的 48.9%（全部为正序）。在 Llama 上，这些输出的按数据集计算的 Error@10 为 0.46–0.51。表 18(a) 报告了以下情况下位置性占比的下降：

（*a*）位置输出占比（%）

骨干模型	数据集	受控输入	原始	受控
Llama	图书	最近一次交互	29.6	7.1
Qwen	Yelp	评论摘要已移除	46.9	10.1

(b) Llama 上非位置性输出的 *NDCG@10*

数据集	受控输入	原始	受控
图书	历史记录截断	0.508	0.556–0.599
Yelp	历史记录截断	0.465	0.488–0.505
图书	作者移除	0.508	0.547
Yelp	评论摘要移除	0.465	0.570

（*c*）*Llama Movies* 上的字段可用性和原始输入质量

		原始 NDCG@10	
字段	符合条件的实例	符合条件	剩余
标题	587	0.381	0.224
描述	576	0.384	0.223

表18：**证据移除诊断**。 原始设置使用完整历史和原始项目文本；受控设置使用该行指定的输入变更。面板 (b) 以各条件的非位置性输出为条件；范围涵盖四种历史截断变体。面板 (c) 比较不同字段可用性子集下原始输入的质量。

将图书历史记录限制为最近一次交互，并在 Qwen 上移除 Yelp 评论摘要。在汇总后的非位置性 Llama 热启动子集中，Top10Mass 的错误 AUROC 达到 0.612。该子集不包括遵循候选项顺序或其逆序的排序。

表18(b)报告了每个评估条件下非位置性输出中的NDCG@10。历史截断提升了 Books 和 Yelp 上的条件质量；MIND-20 和电影未显示出相应增益。这些比较评估了保留最近一次交互的较短历史记录。项目文本增益具有字段特异性，正如移除 Books 作者信息和 Yelp 评论摘要所示。

表18(c)分别报告了符合各字段消融条件的电影实例和其余实例的原始输入 NDCG@10。符合条件的子集原始 NDCG@10 已经较高。因此，字段可用性会通过变体构成影响 Item-text总体结果。这些诊断结果补充了表2中的条件级结果。

对于第3.2节中的设置，稀疏历史包括完整历史记录和四种截断变体，因此每个实例有五条记录，总计20,000条记录。Item-text变体要求至少一个候选项包含被移除的字段。表18(c)中的Movies标题和描述覆盖率遵循此规则。尾部截断点根据有训练交互的物品计算，而符合条件的目标也包括训练交互数为零的目标。MIND-20/图书/电影/Yelp的尾部数量分别为870/221/129/92；870个MIND-20目标中，有856个的训练交互数为零。

A.12 同提示词逐项评分

我们针对每个骨干模型上的相同4,000条热启动记录，在原始逐项排序提示词下重建概率。提示词哈希与基础运行一致。存储输出采用教师强制方式生成；我们在输出标签槽位读取有效标签分布，使用与LabelProb相同的不带空格和带空格前缀的词元形式，同时读取输出词元对数概率。已发布排序始终保持不变。

首位前十质量是 1 减去指定的前十标签上的首位概率质量。首位最大概率使用 1 减去最大有效标签概率。序列NLL@10 对前十个发出的标签的负对数概率求和。探针自身集合对照将前十质量计算应用于

分数	概率来源 / 集合	Llama	Qwen
LabelProb-Top10Mass	单字母提示 / 发布的前十项	0.628 [0.611，0.644]	0.447 [0.430，0.464]
探针自身前10项质量	单字母提示 / 探针前10项	0.525 [0.508，0.544]	0.568 [0.550，0.585]
首位前十质量	排序提示 / 发布的前十项	0.606 [0.589，0.624]	0.518 [0.500，0.535]
首位最大概率	排序提示	0.527 [0.509，0.545]	0.486 [0.469，0.505]
序列NLL@10	排序提示	0.399 [0.381，0.416]	0.306 [0.290，0.322]

表19：**同提示词与跨诱导得分**。热启动错误 AUROC 及其 95% 置信区间。排序提示词得分使用存储输出的教师强制重建。

(a) Llama 错误 *AUROC*

变体	热启动	稀疏历史	尾部目标 项目文本	
仅样本	0.533 [0.515，0.552]	0.536	0.506	0.523
基准锚定	0.555 [0.538，0.573]	0.552	0.524	0.554
样本加基准	0.553 [0.536，0.571]	0.555	0.523	0.550

(b) Qwen Warm-start 热启动错误 *AUROC*

变体	热启动
仅样本	0.381 [0.362，0.399]
基准锚定	0.365 [0.347，0.383]
样本加基准	0.376 [0.356，0.394]

表20：**与基础排序的自洽性**。错误 AUROC；热启动条目包含 95% 置信区间。Qwen 在热启动条件下进行评估。

单字母分布中概率最高的十个标签。表19使用热启动评估器和1,000次按数据集分层的配对自助重采样报告这些分数。

这两种比较的方向取决于骨干模型。对于 Qwen，使用已发布集合会使错误 AUROC 相较于探针自身的集合降低 0.121；在排序提示下重建首位前十质量，则会使其相较于跨诱导 Top10Mass 提高 0.071。在两种骨干模型上，序列NLL@10 按预先定义的方向计算均低于随机水平。

这些概率是对存储输出进行的教师强制重建。Llama 的贪心下一词元一致率为 94.4%，Qwen 为 99.5%。使用原始 transformers 5.7.0 版本和 KV 缓存解码时，精确贪心再生成与经检查的 50 个 Llama 输出中的 19 个相匹配。

A.13 基于基础排序的自洽性

我们使用现有的五个抽样排序，评估公式（23）之后所述的两种包含基准的变体。基准锚定分歧取已发布排序与每个样本之间前 10 个 Jaccard 距离的平均值。样本加基准则将原始成对计算应用于这六个排序。表20报告了原始分数和两种包含基准的分数。

相较于仅样本自洽性，这两种变体在 Llama 上分别使热启动错误 AUROC 提高 0.022 和 0.020，在 Qwen 上分别使其降低 0.016 和 0.005。在 Llama 的每种条件下，两种变体的分数均低于 Top10Mass；在 Qwen Warm-start 上，两者均低于分解式大语言模型不确定性量化。

A.14 推理成本

表21将额外的模型调用与基础排序调用分开。四种标签概率信号共用一次前向传播；温度缩放复用其输出，并使用带标签校准数据拟合一个一维温度。分解式大语言模型不确定性量化使用候选顺序前向传播，而自洽性使用采样生成。扰动排名不稳定性复用基准测试中受控的 Llama 排序。其受控生成要求为：
MIND-20 需要 7 次调用，图书和 Yelp 各需要 8 次，电影需要 5–7 次。

信号	额外调用次数 每条记录	Type		Llama （45,163）	Qwen （4,000）
LabelProb 系列	1（共享）	前向传播		8小时02分钟	1小时28分钟
温度缩放标签概率	0（复用输出）	离线拟合		—	—
分解式大语言模型不确定性量化	5	前向传播		22小时11分钟	3小时26分钟
自洽性	5	采样生成		121小时53分钟	13小时25分钟
扰动排名不稳定性	平均7.3	受控 次数	贪心生成-	复用17小时43分钟†	
基础排序	1次基础调用（参考）	贪心生成		24小时32分钟	2小时36分钟

表21：**推理要求和记录的实际耗时总计**。时间涵盖所示的记录集。
†Qwen 的扰动时间涵盖 41,163 个受控排名。

所有运行均使用单块 A100 80GB。实际耗时总计反映了特定于方法的批处理设置下的运行成本，仅供参考。Llama 计时集包括尾部目标资格筛选前生成的记录；Qwen 受控排序的总计则在表21中单独列出。