SignalCount：数秒内通过结构化与语义谓词实现近似计数

Matteo Agius-D'Arrigo
matteo@outersignal.com
OuterSignal 纽约，纽约州，美国

Zachary Zelner
zach@outersignal.com
OuterSignal 纽约，纽约州，美国

Harrison Ju
harrison@outersignal.com
OuterSignal 纽约，纽约州，美国

Rushab Munot
rushab@outersignal.com
OuterSignal 纽约，纽约州，美国

arXiv:2611.30779v1 [cs.DB] 23 Nov 2026

摘要

许多应用程序会存储带有结构化列和文本文档的记录，并询问筛选集中的多少条记录符合某项判定标准，而这项标准只有阅读文本的人才能判断，例如："有多少名 2 型糖尿病患者年龄在 40 至 70 岁之间，并因血糖持续偏高而停用了二甲双胍？"交互式应用需要在用户输入判定标准时，于几秒内得到这个计数，以便决定是否值得通过批处理流水线物化匹配集，并由大语言模型判定记录。以往针对这类语义谓词的近似查询处理依赖昂贵模型（即预言机）在查询时标注样本，因此每次查询都要等待预言机。我们推出了 SignalCount，这是一个生产系统，可针对包含结构化与语义谓词的 COUNT 查询给出答案，且查询时无需调用预言机，即使判定标准是临时提出的也不例外。其预言机是生产环境中的大语言模型判定器；该判定器记录的判定结果用于离线校准快速分类器，而该分类器本来就在判定器之前运行。查询时，SignalCount 会精确计数筛选集中的记录，并在一次列式扫描中抽取至多 1,000 条记录的均匀样本；随后用分类器为样本评分，并将校准概率的平均值乘以精确计数。其 90% 区间将抽样项与单侧、依赖比率的校准误差边际相结合；这些边际通过跨租户的分割保序预测拟合。在多租户生产部署中，延迟的 p99 为 3.8 秒；即使筛选集规模超过样本量，延迟仍保持稳定。对于未参与拟合的租户，该区间在 92% 的估计中涵盖生产流水线返回的计数；不过，从聚合结果看，估计值偏高约 15%，因为即使校准在单条记录层面近乎无偏，在单项判定标准层面仍可能有偏差。

CCS 概念

• **信息系统** → **数据库查询处理**；在线分析处理引擎； •**计算方法** → 自然语言处理； • **计算数学** →假设检验和置信区间计算。

关键词

近似查询处理、语义谓词、大型语言模型、校准、抽样、置信区间

1 引言

一家研究型医院正在招募糖尿病试验参与者，需要寻找一线药物二甲双胍未能控制其血糖的患者。协调员将患者记录筛选为2型糖尿病、年龄在40至70岁之间、今年就诊且居住在试验地点附近的患者，结果有50,000名患者符合条件。这个过滤条件是精确的，因为其中每一项都对应一个数据列。患者停用二甲双胍的原因并不是数据列中的信息，而是写在临床记录里的；关键词搜索无法区分患者是因副作用停药，还是因为药物失效而停药；而且，没有任何固定分类体系能预见每项试验的所有判定标准。因此，协调员用自然语言写下判定标准，再由一条模型流水线读取每位患者的记录，找出符合条件者。这条流水线适合生成最终名单，但它需要处理整个筛选集，耗时很长，以至于没人愿意运行它来确认判定标准的措辞是否合适。

而且，判定标准第一次往往很难措辞得当（图1）。"使用过二甲双胍"大约匹配其中 9,000 名患者，而协调员本季度只能筛查 3,000 份病历。"停止使用二甲双胍"仍会纳入那些因胃部副作用而停药的患者。"因血糖持续偏高而停止使用二甲双胍"才是试验所需的队列。协调员需要在输入时、几秒钟内看到该判定标准大致选中了多少名患者，以及对这个数字有多大把握，这样才能在缓慢的流水线运行一次之前完成队列塑形。凡是在对文本记录进行计数后需要执行高成本操作的场景中，都会出现同样的循环。保险公司会在派遣查勘员之前询问，有多少份未结案财产索赔"描述的是洪水损害，而不是风灾损害"；营销人员则会在支付营销活动费用之前询问，有多少条客服聊天记录"抱怨的问题正是我们的新产品能够解决的问题"。

SignalCount 回答了这个问题：对一个经过精确过滤且包含一个语义谓词的筛选集进行近似计数，并在几秒内返回一个 90% 区间。对于第三个判定标准，它估计约有 2,000 名符合条件的患者，90% 区间为 418 至 2,808，耗时约 2.5 秒。区间的两端都能为决策提供信息：上限符合筛查容量，而下限仍足以提供填满试验所需的候选者，因此协调员保存了该队列，流水线则找到了所有匹配项。示例中的数字仅用于说明；第5节

评估生产部署。

在 SignalCount 运行的多租户SaaS应用中，保存判定标准会触发一个批处理流水线：该流水线先通过嵌入和分类器过滤器，将精确的筛选集物化，然后再让生产环境中的大语言模型评审器处理最有希望的记录。对于超过 500,000 条记录的筛选集，若以这种方式评审每条记录，成本约为一次估计的 1,100 倍，耗时至少半小时（第5.8节）。

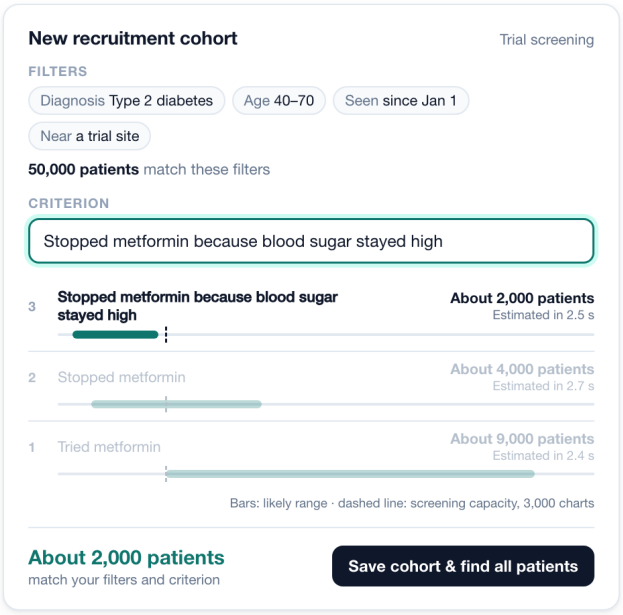

图1：随着用户输入判定标准，逐步形成队列。每个估计值都会在几秒内得出；条形显示其可能范围，以及筛查容量（虚线）。保存后将运行批处理流水线。

现有方法为何需要等待。针对含有昂贵谓词的聚合，既有研究会将一个廉价模型（代理）与一个昂贵模型（预言机）配对，后者的回答定义了真实结果。ABae [20] 使用代理为每条记录评分，按分数将记录分层，然后在查询时让预言机为每层中的一个样本标注。预言机的标注给出估计值及其区间，而代理只负责决定从哪里抽样；后续系统也沿用同一模式 [10, 15, 28]。预测增强推断（PPI） [2] 则对所有记录的代理预测取平均，再用预言机标注的小样本中代理的误差来校正该平均值。无论代理表现多差，这两种方法的区间都能成立，但代价都要在查询时付出：每个新谓词都需要新的预言机标注，而且代理必须处理每条记录。在我们的部署中，预言机是生产环境中的大语言模型评审器，单次调用的 p50 延迟为 3.7 秒，p90 延迟为 7.7 秒。即使样本标注完全并行完成，所需时间也不可能短于最慢的一次调用，超过了 5 秒的预算；此外，代理处理环节所需时间会随筛选集增大。我们依据评审器测得的延迟提出这一论点，并未在生产环境中运行这两种方法。

查询时无需预言机标签。SignalCount 将预言机转移到离线环节。生产流水线已在判定器（Jev，第 3 节）前运行一个快速的非生成式分类器，该分类器会为每条记录的文本符合自然语言标准的可能性打分，并记录判定器所见每条记录的分类器分数和判定结果。分类器就是代理，而这些日志就是无需额外成本即可收集的预言机标签。根据这些判定结果完成校准后，只需利用一个小型均匀样本上的分类器概率，就能直接回答查询。校准结果还必须能迁移到所查询的标准和筛选集，因此区间必须将校准误差考虑在内。

对于已知查询，AlloyDB 将代理训练转移到离线环节 [7],；经过校准的分类器可在每日批处理中针对固定类别测量发生率，无需调用大语言模型 [33]。SignalCount 则能在几秒内，针对查询时提出的任意标准，在经过精确计数的筛选集上完成同样的操作。

要求。 **(R1) 延迟**。 对于从少数记录到租户所拥有的最大规模的筛选集，p99 延迟低于 5 秒；每次查询的模型调用次数和模型成本不会随筛选集规模增长。 **(R2) 经验证的区间**。 每个估计值都附带一个名义90%区间，即使判定标准对应的情况罕见且样本中没有匹配项也是如此；其经验覆盖率会整体测量，并按估计率分组测量。 **(R3) 与生产环境评审器达成一致**。目标值是保存判定标准后，生产流水线标记为匹配的记录数量。

方法。 数据库会精确给出分母，只需估计匹配项所占的比例。一次分析查询会返回筛选集的精确大小，以及一个由种子确定、至多包含 1,000 个记录 ID 的均匀样本。分类器对这些记录的文档进行评分；根据已记录的（分数、判定）配对数据拟合出的保序表，会将每个分数转换为判定器接受该记录的概率；估计值则为这些概率的平均值乘以精确大小。其90%区间会通过两项单侧边界扩大估计值：一项是用分割保序预测拟合、并随估计率平滑变化的校准误差边界，另一项是有限总体抽样项。各统计组件均已得到确立；本文的贡献是将它们组合成一条在查询时无需调用预言机、模型工作量受样本上限约束的路径，并构造一个校准边界跨租户拟合的区间，以及在生产环境中对留出租户测量其覆盖率。

贡献。

- **一种受延迟约束的设计**，用于处理带有结构化与语义谓词的 COUNT 查询，其模型工作量受样本上限限制（第3节和第4节）。
- **单侧、依赖比率的校准裕度**，通过跨租户的分割保序预测进行拟合；配合重新拟合的表，相较于单一对称边际，将留出覆盖率从85.5%提高到92.0%，同时具有更窄的中位宽度（第4.4节和第5.4节）。
- **在生产部署中的评估**（第5节）。在数百个租户上对2,119个估计值进行重复租户划分时，名义90%区间在92.0%的留出估计值中覆盖了真实计数。针对租户已保存的筛选器进行的112个估计值中，延迟的p99为3.76秒（图2）；即使面对包含数百万条记录的集合，受样本上限约束的阶段仍保持近乎恒定。

2 问题

数据与查询。一个租户是应用程序的一个客户组织，它有一张记录表：在我们的部署中，每条记录对应一个人；在贯穿示例中，每条记录对应一名患者。范围用于选取查询可能访问的记录，例如去重并应用排除项后的租户记录。每条记录的结构化列存放在列式存储（ClickHouse）中，结构化谓词构成一个编译为 SQL 的过滤条件。对于离线管道已处理的记录，文本文档（在我们的部署中为 JSON）

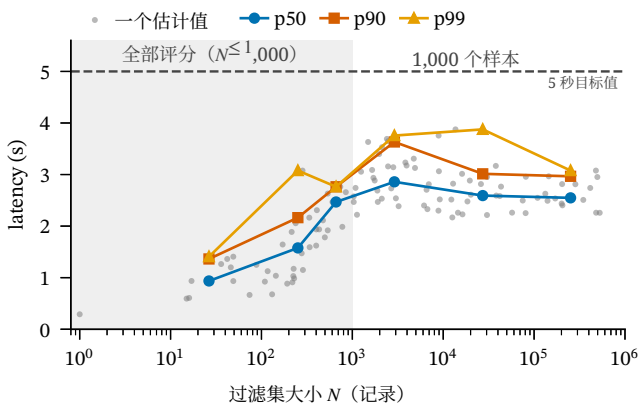

图2：112 个生产环境估计值的延迟与筛选集大小的关系。最多包含 1,000 条记录的集合会被完整评分（阴影部分）。各条线连接每个规模组的百分位数。

存放在以记录 ID 为键的行存储或文档存储中。只有有文档的记录才能由判定器评估，因此筛选集是范围内通过过滤条件且拥有文档的记录，而 N 是其精确大小。该查询是在筛选集上执行的计数，并带有一个语义谓词，即生产环境评审器所使用形式的判定标准。在贯穿示例中，范围是医院的患者记录，过滤条件和判定标准如图1所示。

判定器与目标值。对于某项判定标准，生产流水线会为每条记录给出是或否的判定；用先前研究中的术语来说，其分类器是代理，判定器则是预言机。流水线的最后两个阶段是一个小型分类器和一个大语言模型判定器。分类器会根据该判定标准为每条记录打出 0 到 1 之间的分数；只有分数不低于 0.10 路由阈值的记录才会交由判定器处理，其判定结果为最终结果，其余记录则判为否。更早且成本更低的阶段负责决定哪些记录会交由分类器处理。目标值，即真实计数，是指筛选集中依据已保存标准得到生产判定为是的记录数量；流水线尚未评估的记录按否计（R3）。估算器观察到的是分数，而非判定结果，因此我们将给定分数时每条记录的判定建模为一次概率取决于该分数的抛硬币试验；SignalCount 返回的区间是针对真实计数的预测区间，而非参数的置信区间。

问题陈述。 给定一个租户、一个判定标准、一个过滤条件和一个范围，SignalCount 返回 N，即真实计数的估计值，以及一个标称水平为 90% 的区间，即该区间所声称的水平。其经验 覆盖率是指区间包含真实计数的查询所占的实测比例。该比例应约为 90%，按查询（租户、过滤条件和判定标准）进行统计，其中的随机性来自样本，以及在给定评分条件下的判定结果。p99延迟必须低于 5 秒。每次查询按键获取的行数、模型调用次数和模型成本都必须由与 N 无关的常数限定，这样对范围进行列式扫描就成为唯一一个工作量随数据增长的步骤。我们通过区间的覆盖率以及估计值相对于真实计数的误差来评估系统；系统不输出逐记录标签。

范围。 目标是生产环境评审器，而不是人工标签；评审器与独立标签的一致程度不在本文讨论范围内。校准针对一个分类器和一个评审器；任一者发生变化时，都必须重新拟合表和校准边际。文档在线下生成，因此估计值的新鲜度取决于流水线，并且只涵盖具有文档的记录。SignalCount 针对一个语义谓词、一个分类器族和一个评审器，在一次生产部署中回答计数和比率查询。

3 系统架构

图 3 展示了请求路径，该路径经过三个服务：Web 应用程序的 API 路由、负责文档存储的文档服务，以及评分服务。估计值不会写入任何生产状态，只会写入缓存条目、速率限制计数器以及日志和跟踪记录。只有分析查询会触及整个范围。之后的每个阶段所执行的工作都受样本上限限制，因此其获取的行数、模型调用次数和模型成本不会随 N 增长。

3.1 单次扫描中的计数与采样

估计值需要筛选集的精确大小及其均匀样本。计数必须与租户在过滤器预览中看到的结果一致。单独执行采样查询会再次读取筛选集。直接取存储返回的前几行，会使样本取决于 ID 的排列方式。

因此，我们在计算计数的聚合过程中抽取样本。结构化谓词由过滤器预览所使用的编译器编译为 SQL。单个查询即可返回精确的筛选计数、范围的计数和样本：

```
SELECT count(*) FILTER (WHERE  <过滤条件>) AS N,
count(*) AS N_U, TOP_K_BY(1000, hash(id, <种子>), id)
FILTER (WHERE  <过滤条件>) AS ids FROM
recordsWHERE  <范围>AND  <包含文档>
```

这里的 TOP_K_BY 表示一种有序数组聚合，它会保留哈希值最小的 1,000 条筛选记录，其中哈希使用种子；这就是一个 bottom-k 样本 [9]。如果哈希函数相当于对 ID 进行随机排列，那么该样本就是不放回抽样的均匀样本。当 N 不超过 1,000 时，样本就是整个集合。对于贯穿示例，该查询返回 $N = 50,000$ 个患者 ID。

种子是编译后的查询及其参数的哈希值。因此，重复估计会得到一致结果；编辑过滤条件会生成新的样本；重试也无法通过取平均来消除一次不走运的抽样。均匀性由哈希而非 ID 布局决定。这种设计省去了对筛选集的第二次读取，扫描时只需额外维护一个有界堆。任何支持有序 top-k 或数组聚合的引擎都能做到这一点。

3.2 按键获取文档

文档服务接收判定标准、抽样 ID 和 N。文档是大型 JSON 值，完整样本的一次串行读取将成为整个路径的主要耗时。因此，服务会将这些 ID 划分为多个块，并按键并行读取。每个块都是一次按键进行的只读查找，仅限于请求租户，形式为 WHERE org_id = $1 AND id = ANY($2)。

如果任何数据块失败，整个读取操作都会失败，并被视为可重试。部分样本会偏向于那些已完成的数据块，因此该

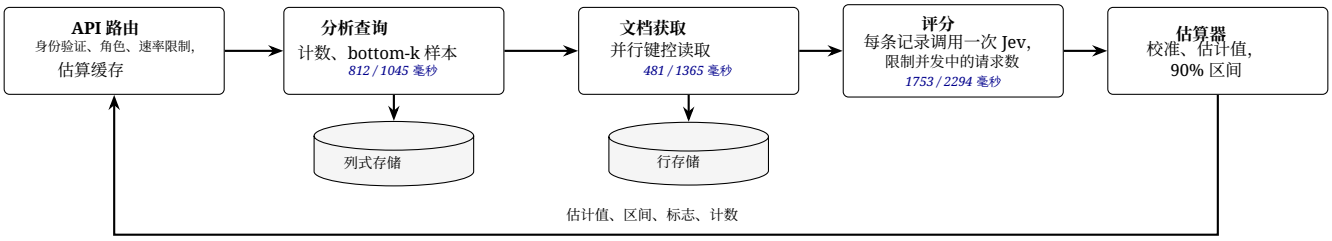

图 3：估计值的请求路径。分析查询之后的每个阶段，其工作量都受 1,000 的样本上限约束。标签给出了延迟工作负载下各阶段的第90百分位 / p99 （第 5.2 节）。

服务永远不会返回部分样本。获取的文档会被打乱，因此请求大小上限保留前缀时，丢弃的是随机子集。

3.3 将评分作为有界扇出

判定器速度太慢，无法在查询时调用。因此，评分服务会使用快速的非生成式分类器为每条抽样记录评分。这种模型能够回答有关输入的结构化问题，而不会生成文本。生产流水线使用的正是 Jev (Typesafe AI)[1]，，以决定哪些记录会送交判定器。嵌入无法取代分类器：它会在判定标准已知之前对每份文档进行编码，因此只能捕捉主题，却会漏掉判定标准所依赖的否定表达和细微差别；此外，不同判定标准之间的相似度没有通用尺度，无法一次性校准。对于每条记录，服务都会根据流水线构造的提示词，针对该记录的文档询问是否满足判定标准，每次请求询问一项判定标准。Jev 会给出 0 到 1 之间的分数，第 4 节将该分数校准为判定器接受该记录的概率。任何能够返回分数的非生成式决策模型都可以替代 Jev，包括自托管开放权重模型，因此评分吞吐量取决于资源配置。

分类器位于生产流水线的网关之后，估计过程不得使流水线资源枯竭。服务通过进程级限流器限制正在处理的调用，并为每条记录发送一个请求。每次尝试的超时时间为 1.7 秒，最多尝试两次，两次尝试之间有短暂的退避时间。每个服务只接纳数量有限的估计请求，其余请求则会收到重试提示并被拒绝。调用方断开连接时，评分服务会中止尚未完成的调用。因此，已接纳的估计请求具有有界延迟，而被拒绝的请求则会快速失败。

服务通过校准表映射每个分数。它会返回估计值、比率和计数的区间，以及已评分和已丢弃的记录数。当估计率低于 2% 时，它会设置 罕见标志；当没有抽样记录达到路由阈值时，则会设置 低于阈值 标志。该路由会添加 N、范围内的计数以及整个范围的比率。

3.4 缓存与成本

每个未命中的估计值都会针对每条抽样记录进行一次计量分类器调用，而用户编辑判定标准时会重复执行查询。该路由会按租户对估计值进行速率限制。每个响应都会缓存 600 秒，缓存键由租户、应用上下文、标准文本和编译后的查询的哈希组成

[1]https://docs.typesafe.ai/introduction

。因此，对过滤条件的任何修改都会导致缓存未命中。网关不会向估计值调用提供缓存答案，因此每个未命中的估计值都要单独支付其评分成本。因此，估计值的成本是由样本上限决定的固定值；第5.8节将其与评判每条记录的成本进行比较。

3.5 故障处理

每一跳都会根据相同的重试是否可能成功来判定故障。暂时性错误包括占用的槽位、存储、读取或服务超时，以及样本损失过多（第4.5节）。这类错误会返回 503，并附带 retryable: true；如果槽位正忙，则返回带有 Retry-After 的 429。接口会重试这类错误。永久性错误意味着服务之间的契约已发生变化，或凭据有误。这类错误会返回 502，按错误级别记录日志，且绝不会重试。每个请求都有超时限制：上文所述的查询超时，以及 20s 的文档服务超时。

文档服务会根据不变量检查每个评分响应，而任何正确的估计值都必须满足这些不变量。所有字段都必须存在，并且下限、估计值和上限必须在 $[0, N]$ 内按顺序排列。比率必须在 $[0, 1]$，范围内；至少要有一条记录完成评分，丢弃计数也必须为非负数。路由还要求响应回显其进行估计所依据的 N。每个已响应的请求都会记录一条结构化日志行，其中包含各阶段耗时，并且该路径不会记录逐条记录的日志。

4 估计

与 ABae [20]，一样，正确性与效率是分开的。当 N 超过样本上限时，覆盖率取决于校准和样本的均匀性，而延迟则取决于上限。

4.1 校准

单调表将每个分类器评分映射为判定器接受该记录的概率。我们按评分区间进行加权保序回归拟合 [34]，区间宽度为 0.01，范围从 0.10 的路由阈值到 1.00。低于该阈值时，由于判定器从不参与判定，概率为零。该表有 91 个条目，是评分服务中的一个常量。部署校准的表是在主研究（第5节）中，基于同一记录和判定标准对应的查询时分数与生产判定的记录的成对数据拟合而成，未进行模型调用。它取代了采用单一对称边际的初始校准 。

我们使用期望校准误差（ECE）来衡量该表：校准概率与观测接受率之间的差距，按分箱求平均，并根据每个分箱所占记录的比例加权

表1：校准表的期望校准误差。折外：在主研究的配对数据上使用5个租户折重新拟合，每组配对数据均由不包含其租户的拟合表进行评分。底部：初始表在主研究配对数据上的可靠性。

表、配对数据	配对数据	All	≥ 0.10
初始样本内研究	188,757	0.0065	0.0262
初始，主研究	1,660,862	0.0083	0.0275
部署形式，主研究，折外 1,660,862		0.0037	0.0012
校准概率	**记录**	**预测值**	**观测值**
0（分数低于 0.10）	1,373,189	0	0.0042
[0, 0.1)	20,165	0.0736	0.0573
[0.1, 0.2)	50,064	0.1316	0.1043
[0.2, 0.3)	35,125	0.2381	0.2007
[0.3, 0.4)	26,510	0.3435	0.3131
[0.4, 0.5)	20,833	0.4453	0.4186
[0.5, 0.6)	16,938	0.5507	0.5265
[0.6, 0.7)	13,516	0.6593	0.6290
[0.7, 0.8)	15,771	0.7533	0.7241
[0.8, 0.9)	17,285	0.8553	0.8166
[0.9, 1.0]	71,466	0.9770	0.9548

（表1）。对于低于路由阈值的评分，我们使用一个分箱，其概率为零；对于校准概率，则使用10个等宽分箱。低于阈值的分箱包含主研究中82.7%的配对数据，在所有记录中占据主导地位，因此我们还报告了阈值及以上记录的ECE。在主研究的166万条配对数据上按租户进行折外重新拟合后，部署形式在所有记录上的ECE为0.0037，在阈值及以上记录上的ECE为0.0012；相比之下，初始表对应的数值分别为0.0083和0.0275。按记录汇总后得到的较小误差并不能证明计数准确，因为不同标准之间符号相反的误差会相互抵消（第5.5节）。

覆盖率建立在我们所称的校准假设之上。对于筛选集中的每条记录，(a) 判定器以表中给出的查询时分数对应的概率接受该记录；(b) 给定分数后，各判定结果相互独立；(c) 判定结果与抽取了哪些记录无关。该设计并不依赖这一假设完全成立。该表基于多个租户和标准拟合，却应用于一个筛选集中的单项标准；平均校准并不意味着在子集上也校准。因此，校准边界会针对这一假设失效而扩大区间；对于单项标准的误差，其覆盖效果也取决于拟合结果的迁移能力。第5节.4 测量了在留出租户上得到的覆盖率。

4.2 估计

将筛选集中的记录编号为 1 到 N。令 Y_i 在判定器接受记录 i 时为 1，否则为 0；令 p_i 为表中为记录 i的分数给出的校准概率。真实计数 θ是判定结果的和。估计率 $\bar{p}$ 是样本中$\hat{}$ n条已评分记录的 p_i的均值，而估计值θ则是该均值乘以精确大小 N：

$$\theta = \sum_{i=1}^{N} Y_i, \qquad \bar{p} = \frac{1}{n} \sum_{i \in \text{sample}} p_i, \qquad \hat{\theta} = N\bar{p}. \tag{1}$$

服务会将 $\hat{}\,\theta$四舍五入到整数进行报告。这是一种概率式分类计数 [3], ，即在均匀样本上进行分类计数 [11], 的一个变体。

表 2：区间中的量，以及每个量的设定时机。

量	含义	Set
N	筛选集的大小	每次查询，精确计数
n	已评分记录，最多 1,000 条	每次查询
Y_i, θ	对记录 i 的判定（1 或 0）；真实计数 ，它们的和	查询时未知
p_i	记录 i的校准概率	每条记录，从表中读取
$\bar{p},\ \hat{}\,\theta$	估计率；估计值 $N\bar{p}$	每个查询，从样本中
V, W	方差：p_i；平均值：$p_i(1-p_i)$	每个查询，从样本中
T	抽样项	每个查询，来自 N, n, V, W
表	评分到概率，91 个条目	离线，由所有租户共享
f_L, f_U	低因子和高因子（表3）	离线，跨租户；读取时间为估计率
1.645, 2.303	正态分布第95百分位数；$-\ln 0.10$	常数

根据校准假设，在给定评分的条件下，该估计值的期望误差为零，因为 θ 和 θ 的期望值相同，即筛选集上 p_i 的和。对于 $\hat{}\,\theta$，均匀样本的均值是集合均值的无偏估计值 [8, 定理 2。1], 因此，$\bar{p}$ 的平均值为 $\frac{1}{N}\sum_{i=1}^{N} p_i$。对于 θ，该假设的 (a) 部分指出，每个 Y_i 以 p_i的概率为 1。该估计不需要记录级准确性，但需要筛选集上的校准（第 5.5 节），并且要计入评分前移除的每条抽样记录。每个查询的种子都是固定的，因此覆盖率是针对查询而言的，而非针对同一查询的重复实验而言。在贯穿示例中，1,000 名抽样患者中有 50 名的校准概率为 0.8，其余患者的分数低于阈值。均值为 0.04，估计值为 2,000 名患者。

4.3 区间

名义90%区间在每一侧漏掉真实计数的概率约为5%。估计值与真实计数之间存在两种误差来源，需要采用不同的工具。抽样误差源于评分 n 条记录，而非全部 N；教科书式调查统计给出了其大小，且随着 n 增大而减小。校准误差则源于将基于多个租户和标准拟合的表应用于单个判定标准；无论评分多少条记录，它都相同，而且没有公式可以确定其大小，因此我们在真实计数已知的过去判定标准上测量它。我们假设表格正确，以此推导下文中的抽样项 T来衡量第一种误差；对于表格不正确的情况，则用校准边际来衡量第二种误差。表2列出了所用的各项数量。

如果这张表没问题，还有两件事仍然取决于运气。假设校准假设成立，估计值仍会与真实计数不同，原因有两个，且彼此无关。

第一个原因是，只有一部分样本会被评分。用 μ 表示筛选集中所有 N条记录的 p_i 平均值。若对每条记录都评分，得到的估计值将是 $N\mu$。但服务得到的是 $N\bar{p}$，而 $\bar{p}$ 与 μ之间的差异，取决于抽中了哪些记录。

第二，概率并非判定。对于满足 $p_i = 0.8$ 的记录，每五次中有四次会被接受，但任何一条这样的记录要么被接受，要么不被接受。因此，即使是 $N\mu$，也就是接受的记录的期望数量，也会与实际数量 θ 不同；正如抛掷一枚正面朝上的概率为五分之四的硬币 100 次，正面朝上的次数也很少恰好是 80 次。

估计值的误差是这两个差距之和，因为 $N\mu$ 会抵消：

$$\hat{\theta} - \theta = \underbrace{\left(N\bar{p} - N\mu\right)}_{\substack{\text{sample gap: which} \\ \text{records were drawn}}} + \underbrace{\left(N\mu - \theta\right)}_{\substack{\text{verdict gap: how} \\ \text{the verdicts fell}}}. \tag{2}$$

每个差值的大小。 我们用方差（标准差的平方）来衡量随机误差的大小。首先，假设每个概率都是已知的，用全部 N条记录表示每个方差。其中涉及的两个量 V_{all}和 W_{all}在查询时是未知的，下一段将用它们在样本中的值替代这些量。

样本差值。 从总体中随机抽取的 n 个值的平均值，其方差等于总体方差除以 n。设 V_{all} 为 p_i 在全部 N条记录上的方差。由于样本是不放回抽样地从有限集合中抽取的，根据 $1 - n/N$ [8, 定理 2.2]，方差还要乘以一个系数：样本覆盖集合的大部分时，剩下的随机性很小；覆盖全部集合时，则没有任何随机性。样本差值是平均值误差的 N 倍，而将某个量乘以 N会使其方差乘以 N^2。因此，样本差值的方差为 $N^2(1 - n/N)\ V_{\text{all}}/n$。

判定差值。 这个差值为 $\sum_{i=1}^{N}(p_i - Y_i)$，每条记录对应一项。若判定结果以概率 p_i 为 1，则其方差为 $p_i(1 - p_i)$；对于无论如何都确定的记录，方差为零，而对于概率为 0.5 的记录，方差最大，为 0.25。根据假设的 (b) 部分，各判定结果相互独立，独立量的方差可以相加。设W_{all}为所有 N条记录中 $p_i(1 - p_i)$ 的平均值，那么 N方差之和为 $N\,W_{\text{all}}$。这个差值与样本无关：即使对每条记录都进行了评分，它依然存在，因为判定结果仍然未知。

根据 (c) 部分，抽取了哪些记录并不能说明判定结果如何。因此，两个差值相互独立，其方差可以相加：

$$\begin{aligned} \operatorname{Var}(\hat{\theta} - \theta) &= N^2\left(1 - \frac{n}{N}\right)\frac{V_{\text{all}}}{n} + N\,W_{\text{all}} \\ &= N^2\left[\left(1 - \frac{n}{N}\right)\frac{V_{\text{all}}}{n} + \frac{W_{\text{all}}}{N}\right]. \end{aligned} \tag{3}$$

第二行只是将 N^2 从第一行中提取出来。

从方差到抽样项。 公式 3 无法按原式计算： V_{all} 和 W_{all} 是基于全部 N条记录计算的，而实际只有样本经过评分。在样本上计算的同两个量，即 V（除数为 $n - 1$）和W，其平均值与它们相等 [8, 定理 2。4], 因此可以用它们代替。将它们代入公式 3 并开平方，即可得到误差的标准差。误差是许多微小贡献之和，因此近似服从正态分布 [12],，而一个正态量

有 5% 的时间会比均值高出 1.645 个标准差以上，而低于均值同样多的情况也会有 5%。因此，抽样项为 1.645 个标准差：

$$T = 1.645\,N\sqrt{\left(1 - \frac{n}{N}\right)\frac{V}{n} + \frac{W}{N}}. \tag{4}$$

如果该表正确，真实计数约有 90% 的时间会落在估计值的 T 范围内。

在贯穿示例中，样本包含 50 个 0.8 的概率和 950 个零，因此 $V = 0.0304$ 和 $W = 0.008$。样本

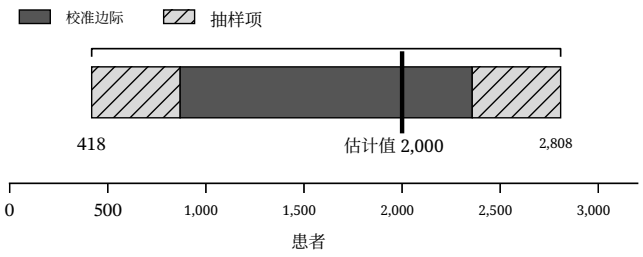

图4：贯穿示例中围绕2,000名患者估计值的90% 区间：校准边际（低侧更宽）以及两侧各自的抽样项。在2,390的宽度中，边界约占1,490，抽样项约占900；约115名患者的下限不起作用。

差值的方差为 $50{,}000^2 \times 0.98 \times 0.0304/1{,}000$，约为 74,500，标准差为 273 名患者。判定差值的方差为 $50{,}000 \times 0.008 = 400$，标准差为 20 名患者。两者合计方差约为 74,900，标准差为 274 名患者，因此 $T = 1.645 \times 274$，约为 450 名患者。样本差值占方差的 99.5%。

如果表是错的：校准边际。抽样项假定表是正确的。对于某个判定标准，情况可能并非如此：表中可能给该判定标准的记录赋予了概率，例如高出三分之一。无论样本量多大，估计值也会随之高出三分之一。这种误差是一个因子，而不是一个固定数值；增加评分记录并不会使其缩小。因此，校准边际会将估计值乘以一个因子：如果该判定标准在表中的数值偏高，则乘以低于1的低因子；如果偏低，则乘以高于1的高因子。没有任何公式能够给出这些因子。第4.4节利用真实计数已知的过往判定标准来拟合这些因子，而它们取决于估计率。

区间。 区间先对每一侧应用校准边际，然后应用抽样项（图4）：

$$\begin{aligned} \text{low} &= \text{estimate} \times \text{low factor} - T, \\ \text{high} &= \max\left(\text{estimate} \times \text{high factor} + T,\ 2.303\,N/n\right). \end{aligned}$$

这里的估计值是四舍五入前的 ˆ θ。下限向下舍入并截断为零，上限向上舍入并截断为 N，因此区间始终包含估计值。下限 $2.303\,N/n$ 确保仅有少量弱匹配的样本，其上限不低于一个包含 n 条记录且没有匹配的样本所支持的边界（第4.5节）。

在贯穿示例中，估计率为 4% 时，第4.4节的因子在表3各中心点之间进行插值，得到低因子为

0 43 1 18 2 000 × 0.43 − ，高因子为 。
450 2 000 × 1.18 + 450
低端为 ，高端为 ，；使用未舍入因子时，区间为 418 至 2,808 位患者。

4.4 拟合边际

初始校准为每个估计值扩大了一个对称边际。在留出租户上，其覆盖的估计值少于名义水平所声称的数量。未命中呈单边分布：几乎全是

高估，真实值低于下限，而且这些情况出现在估计比率较高时（第5.3节）。

因此，部署的校准边际是单侧的，并且取决于估计率。在估计率下，将 f_L 写作低因子，将 f_U 写作高因子。用符号表示，在四舍五入和截断之前，第4.3节中的区间为

$$\begin{aligned} \text{low} &= \hat{\theta}\, f_L - T, \\ \text{high} &= \max\left(\hat{\theta}\, f_U + T,\; 2.303\, N/n\right). \end{aligned} \tag{5}$$

一次过去的运行需要什么。校准运行是一次过去的估计，其真实计数 θ 已知。对于每次运行，我们会询问：要使区间达到 θ，恰好需要扩大多少。在低侧，低端 $\hat{\theta} f_L - T$ 必须小于或等于 θ。令其等于 θ 并求解该因子，即可得到该次运行的所需低因子。高侧与此相同，只是符号相反：

$$f_L^{\text{need}} = \min\left(1,\; \frac{\theta + T}{\hat{\theta}}\right), \qquad f_U^{\text{need}} = \max\left(1,\; \frac{\theta - T}{\hat{\theta}}\right). \tag{6}$$

两者都使用该次运行自身的抽样项 T。因子为1表示不进行扩大。只有当估计值超过真实计数的幅度大于抽样项时，运行才需要小于1的低因子；只有当估计值低于真实计数的幅度大于抽样项时，才需要大于1的高因子。当下限已经达到 θ 时，所需高因子同样为1。例如，一次运行的估计值为2,000，抽样项为450，真实计数为1,000，则需要低因子1，450/2，000 = 0.725：缩放至1,450后，估计值与真实值相差不超过450。实际上，我们会搜索一个精细的因子网格（步长约为0.3%），并应用四舍五入；其结果与公式6一致，误差不超过网格步长。

从多次运行到一个因子。 我们将运行分为四个估计率组 （表3）。在每个包含 m 次运行的组内，我们按照所需扩大量从少到多排列运行，并取位置为 $\lceil 0.95 \ (m+1) \rceil$ 的运行所需的因子。该组约95%的运行所需的扩大量不超过这一数值，因此约有5%的运行会在该侧被遗漏。这就是 [24] 在每侧遗漏率为5%时所采用的拆分共形规则，因此当运行可交换时，双侧覆盖率约为90%。来自同一租户的运行并不可交换，而插值会保留每组规则，因此覆盖率是通过测量得到的，而非得到保证。该规则不对校准误差的分布作任何特定假设。

每个因子位于其所属组的中心，即该组边界在对数尺度上的中点（最低组为0.3%，因为它没有下边界）；在各中心之间，服务会进行插值，因此边界不会在组边缘处突然跳变。它在线性插值中使用估计率的对数来计算每个因子的对数$_{10}$，并在最外侧中心之外保持因子不变。两种尺度都采用对数形式，因为这两个量都是按比率进行比较：1%到10%与10%到100%的距离相同，而因子2与1的距离和因子1/2与1的距离相同。贯穿示例中的4%比率，在该尺度上从2.24%的中心到10%的中心之间走过了39%的距离，因此其低因子也从0.47到0.38之间走过39%的距离，得到0.43。我们会检查边界相对于估计率是否为非递减。最低组自身的高因子接近1，因此将其提高到基于所有运行共同拟合的高因子。因子基于主研究各次运行的折外估计进行拟合，按租户划分折，并使用不包含该折数据的表对每次运行进行估计。两侧之所以不同，是因为在低比率和中等比率下，估计值偏高。

表3：按估计率组划分的已部署边际因子。

估计率	低于1%	1–5%	5–20%	高于20%
组中心	0.3%	2.24%	10%	44.7%
低因子 f_L	0.81	0.47	0.38	0.53
高因子 f_U	1.31	1.11	1.30	1.53

使用根据查询时分数重新拟合的表，这些校准边际将留出覆盖率从85.5%提高到92.0%，同时中位宽度更窄（第5节4）。

4.5 稀有标准、小型集合和损失

零事件界限。 对于稀有判定标准，样本中通常没有达到或超过阈值的记录。此时估计值和抽样项都为零，而将区间设为从零到零，会宣称一种样本无法支持的确定性。空样本并不能说明没有记录匹配。它说明匹配过于稀少，不会在 n 次抽样中出现，这也就为比率设定了上限。

假设筛选集中有 r 的比例符合条件。随机抽取一条记录时，它不匹配的概率为 $1 - r$。独立抽取两条记录时，它们都不匹配的概率为（$1 - r$）（$1 - r$），抽取 n 条记录时则为（$1 - r$）^{n}。比率越高，样本为空的可能性就越低：当比率为1%时，抽取1,000条记录却没有匹配项的概率为0.99_{1000}，约为23,000次中有1次，因此在样本为空之后，1%的比率并不可信。我们将空样本仍有10%概率出现时对应的比率作为上限，因为任何更高的比率都会使我们观测到的结果发生概率低于这一水平：

$$(1-r)^n = 0.10 \;\Longrightarrow\; r = 1 - 0.10^{1/n} \le \frac{2.303}{n}. \tag{7}$$

最后一步用一个更简单、但绝不会更小的值替代精确解：2.303 是 10 的自然对数，而对于 $n = 1,000$，两者都是 0.23%。在贯穿示例中的 50,000 名患者里，这相当于 115 名患者，也就是图4中提到的下限。不放回抽样会使抽到空样本的概率略低于 $(1-r)^n$，因此这个界限偏向安全一侧。这就是 90% 置信水平下的“三法则” [13]。随后，服务返回一个估计值，下限为零，上限为 $\max(0.01,\; 2.303/n)$，并将其限制在 1 以内。它会设置稀有标志和低于阈值标志，因此界面可以显示“少于”这个上限。对于 n 大于 230 的情况，0.01 下限起主导作用，因此该界限的置信水平高于 90%，且越接近样本上限，置信水平越高；对于更小的 n，它达到的是单侧 90% 置信水平，而非其他区间的 95%。只有当每条得到阳性判定的记录在查询时的评分也都不低于阈值时，该界限才会覆盖真实计数。几乎所有记录都符合这一条件（第5节.7）。

小型筛选集。当 N 不超过样本上限时，整个筛选集都会被评分（即进行普查），抽样项的第一部分便会消失。此时，区间反映的是判定变异性、边界，以及比率较小时的下限。文档服务将 N 设为读取的已列出记录数，因此，在两次读取之间离开筛选集的记录不会被计入。

最小样本与损失。 服务会拒绝评分记录少于 100 条的估计值，或者在 N 低于 100 时，拒绝少于 N 条记录的估计值。记录可能在评分前丢失：行不再完整，

以及请求大小上限。文档服务会在抽取样本时统计这些损失。被拒绝的分类器请求只会丢弃其对应的记录。两个限制均适用，各为样本的20%。如果超出分类器大小限制的文档超过该限制，则估计值会被拒绝，并标记为不可用的永久性错误。如果总损失（包括分类器的瞬时错误）超过该限制，则估计值会被拒绝，并标记为性能下降的瞬时错误。被丢弃的记录视为随机缺失。

5 生产部署评估

我们先评估延迟（R1），再评估区间的覆盖率及估计值在留出租户上的偏差（R2、R3）、基线、稀有标准和成本。

5.1 部署与工作负载

部署。 SignalCount 已在一个生产环境应用程序中全面推出，该应用程序为数千个租户提供服务，并以此支持用于受众细分的自定义筛选器和标准。该路由运行在无服务器函数中，两个服务则运行在容器中。本节中的每个估计值均由作者通过该部署，基于真实租户的数据得出。

工作负载。 延迟工作负载 包含 112 个估计值，针对一小部分租户的非空已保存筛选器进行计算，每个筛选器均使用其所属租户保存的判定标准，提交间隔为 1.5 秒。样本内研究 包含 244 个估计值，涉及来自不到 25 个租户的 142 个筛选集，且这些数据均纳入初始校准的边际拟合。其估计值中，125 个使用租户自己的已保存筛选器，51 个使用不带筛选条件的完整范围，68 个使用其他租户保存的筛选器。主研究 给出了我们的主要准确度结果，研究对象为未参与任何先前拟合的租户。我们汇总了每个活跃租户的所有活跃已保存筛选器，并计算了每个符合条件的判定标准对应的真实计数：该标准处于活跃状态、由生产流水线评估，且创建时间至少早于运行日期 30 天。对于由四个真实匹配率组和四个规模组构成的每个分层，我们最多抽取 150 个（筛选集、判定标准）配对数据；每个租户在每个分层中最多抽取 6 组配对数据，每个筛选器最多抽取 3 个判定标准。我们排除了样本内研究涉及的所有租户，并增加了 60 组真实计数为零的配对数据。在数百个租户中共提交了 2,123 个估计值，其中 2,119 个得到了响应：2,059 个对应正的真实计数，60 个对应零。另有 300 个估计值（294 个得到响应）在 10 个抽样种子下重复计算了其中 30 个筛选集；各抽样种子所得估计值的中位变异系数为 0.083。主研究中的筛选集最多包含 857,662 条记录，样本内研究中则最多包含 170 万条记录。表4汇总了这三种工作负载。

此处的覆盖率含义。 部署的表和边界是在主研究上拟合的，因此第5节.4在200个租户划分上重新运行完整的拟合过程，并在两个拟合过程之外的租户上分别进行测试。我们报告总体经验覆盖率、按估计率划分的经验覆盖率以及按侧划分的经验覆盖率。这是对留出租户和标准取平均后的结果，并不构成逐查询保证。它同样是近似的：跨租户的运行至多是可交换的；此外，与边界的拟合一样，该研究按结果进行分层，而不是遵循生产环境查询的实际构成。

真实计数与可复现性。 真实计数（第2节）来自生产环境判定结果，无需进行新的模型调用，输入判定标准就是已保存标准本身。在23组核查样本对中，生产标签率与同一评审器判定结果的另一份记录中的评审器接受率之比的中位数为0.975。

表4：三种工作负载。大小和谓词行数均按每个发出的筛选集计数一次。真实计数行数统计已回答的估计值，并排除样本内研究中的23个核查样本对。

	延迟	样本内	Main
租户	少于25	少于25	数百
已发出/已响应的估计值	112 / 112	244 / 244	2,123 /2,119
筛选集	112	142	1,687
标准	14	136	1,766
每个租户的估计值，中位数 / 最大值	7 / 20	13 / 19	1 / 19
N，中位数	2,435	5,561	2,362
N，第90百分位	144,538	324,327	111,031
N，最大值	527,197	1,697,655	857,662
阳性时的真实率，中位数 / 第90百分位	4.8 / 77.8%	4.1 / 37.1%	2.8 / 30.9%
真实计数为 0 的估计值	4	12	60
结构化谓词，0 / 1 / 2+	12 / 55 / 33%	23 / 43 / 34%	38 / 49 / 14%
（分数，判定）配对数据	–	188,757	1,660,862
判定为阳性	–	11.7%	8.9%

分类器网关记录每次分类器调用。将已部署估计器逐行移植后在这些分数上运行，对于112/112次延迟运行、244次样本内运行中的243次，以及2,119/2,119次留出运行，均能准确复现估计值和上下界。因此，我们在记录的分数上离线评估下面的每个校准变体。由于数据包含租户记录、标准和文档，因此无法发布。

失败。 所有研究中发出的 2,779 个估计值里，有 2,769 个得到响应，且未发生瞬时错误或超时。在其余 10 个案例中，分类器拒绝了某条记录的请求，导致整个估计值失败；已部署服务会丢弃此类记录并将其计数（第4.5节）。

5.2 延迟

尾延迟。 我们测量从判定标准到估计值的延迟：通过只读代理发出的分析查询，加上文档服务调用的耗时；后者会计时其获取和评分阶段。延迟始终低于 5 秒的目标值：在延迟工作负载中，从判定标准到估计值的延迟 p99 为 3.76 秒（图2）。在每个百分位数上，评分都是耗时最长的阶段；在全部 112 次运行中，评分的 p50 为 1483 毫秒，而总耗时为 2491 毫秒；图 3 给出了各阶段的 p90 和 p99。在超过 1,000 条记录的 71 个筛选集中，评分的中位数耗时为 1.6 秒。受样本上限约束的阶段保持稳定。在主研究及其重复实验中的 2,413 次文档服务调用里，该调用的 p50 为 2.1 秒，p90 为 2.6 秒。其 p99 为 4.5 秒，最大值为 7.2 秒；单独看评分，其 p99 为 3.5 秒，因此较长的尾延迟是由分类器造成的。

筛选集的大小。 筛选集超过样本大小后，延迟不会随筛选集增大而增加（图2）。少于 1,000 条记录时，会对整个集合进行评分，因此延迟会随集合增大而增加。对于 1k–10k、10k–100k 和超过 100k 条记录的集合，p99 分别为 3.76、3.88 和 3.08 秒；工作负载中最大的筛选集包含 527,197 条记录。对于超过 1,000 条记录的 71 个集合，N 与延迟的秩相关为 -0.27。对于更大的集合，在超过 1,000 条记录的 179 个样本内估计中，文档服务调用在每个规模组中的第90百分位均为 2.6 至 3.0 秒，最多涉及 170 万条记录。

租户数据的大小。分析扫描是唯一随范围而非 N 增长的阶段；在测量的

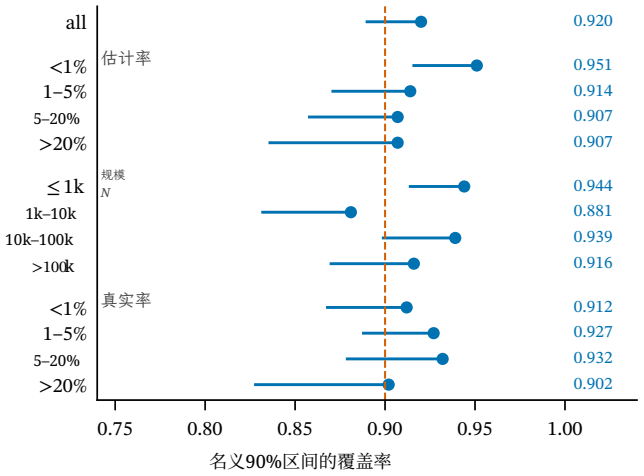

图5：在200次重复租户划分中，留出租户的覆盖率（按组别划分），针对部署校准：均值（点及数字）和第5百分位数（线段末端）。

范围内，它耗时不到一秒。针对最大租户的 132 个已保存的筛选器进行测试，范围最多包含 170 万条记录，分析查询的 p99 为 860 毫秒（运行 3 次取最小值，包含代理跳转）。在最大范围内，仅范围边界处理（去重和排除项，不含过滤条件）耗时 251 毫秒，而对已解析的数组字段应用过滤条件耗时 766 毫秒，因此评估该过滤条件约占扫描耗时的三分之二。该扫描是常规的筛选计数，采用有界的 top-k，因此其扩展性与列式存储一致，而其他所有阶段的耗时均受样本上限约束。

5.3 使用初始校准的样本内结果

样本内研究展示了校准如何出现未命中，以及为何需要留出租户。在样本内，初始校准覆盖了 209 个正真值估计中的 191 个，即 91.4%（租户聚类自助法 [86.7%，95.7%]）。但这些租户无一例外，且其中 121 个（租户、判定标准）配对数据中的 82 个，都参与了其校准边际的拟合。基于样本内租户拟合、并在发出任何留出估计值之前固定的一张表，仅覆盖了 87.0% 的留出正向真实值估计。

未命中都出现在同一侧。18 次未命中中，有 17 次是高估，真实值低于下限：下侧未命中率为 8.1%，而上侧未命中率为 0.5%。这些未命中落在两个较高估计率组中，其覆盖率分别为 85.5% 和 81.8%。校准边际是实现大部分覆盖率的关键。去掉它（两个因子均设为 1）后，覆盖率降至 54.3%；在完整评分的 55 个筛选集上，覆盖率则降至 70.9%，这些筛选集没有抽样误差。在其他租户折上拟合的分割共形边际得到 89.1% [84.3%，92.6%]，，逐折覆盖率介于 75.4% 至 95.5% 之间。

5.4 留出准确度

留出租户上的覆盖率。 在留出租户上，部署校准的平均表现达到名义水平。我们在主研究租户中随机进行 200 次 40/30/30 划分，并在每次划分上重新执行整个拟合过程。表格在第一部分上拟合，边界在

表5：按折外评估的设计备选方案。上：在主研究上的区间，按租户进行交叉拟合。下：基于208个正向真实值样本内估计的点估计，使用复现得分、5个租户折和初始校准；按真实率计算和之比。

区间	覆盖率	最差组			宽度
全局对称边际	90.0%			–	1.79
全局单侧配对	90.3%			80.7%	–
各组单侧	90.5%			90.4%	1.50
期望最大化先验偏移 +边界	≤86.3%			–	7.63
点估计	**中位误差**	**<1%**	**1–5%**	**5–20%**	**>20%**
平均（SignalCount）	29.6%	1.45	1.58	1.12	0.92
ACC	52.5%	0	0.50	1.06	1.05
PACC	83.8%	0	0.63	1.00	1.07
计数 ≥ 0.35（拟合）	21.6%	–	–	–	–

第二部分上计算，第三部分用于测试，因此测试租户不会参与任何一项拟合（图5）。平均覆盖率为 92.0%，标准差为 1.7 个百分点，第 5 百分位数为 89.0%。除记录数为 1k–10k 的筛选集外，各组均值均不低于 90.2%；该组为 88.1%。最低的第 5 百分位数分别为该组的 83.2% 和真实率高于 20% 时的 82.8%。

从初始校准到部署校准。 依赖比率的单侧边界解释了相较于初始校准的大部分提升。初始校准使用了同一形式的表，该表根据流水线自身在四周内记录的 340,683 组（分数、判定）配对数据拟合而成。其唯一的对称边际在所有比率下均采用 1.43 的高因子和 0.70 的低因子，相当于实际接受数与预测接受数之逐准则自然对数比率的 1.645 个稳健标准差（各为 0.216）。该比率基于 645 组（租户、准则）数据计算，其拟合样本包含样本内研究中的所有租户。在相同的 200 次划分中，初始校准覆盖了 85.5% 的留出估计值。部署校准使用查询时分数重新拟合该表，并采用表3中的边界。它将覆盖率提高到了 92.0%，且每个估计比率、规模和真实比率分组的覆盖率均不低于初始校准。其宽度中位数也更窄：为真实计数的 1.59 倍，而初始校准为 1.64 倍。仅重新拟合该表并沿用初始边际，在下文的交叉拟合中覆盖率仅达到 86.5%。

边际形式的选择。 按组拟合的单侧边际在每个估计率组中的覆盖率为 90.4% 至 90.5%，中位宽度为 1.50；相比之下，对称边际的中位宽度为 1.79（表5）。我们通过在租户间进行5折交叉拟合来比较不同形式：表和每个边际都在其他折上拟合。单一的全局单侧边际对在最高估计率组中的覆盖率仅为 80.7%。然而，在重叠的校准集上拟合的逐组分位数会使逐组覆盖率接近恒等关系，因此，上述重复划分是我们采用的留出测量方法。部署形式增加了对数比率插值、最低组上边际的下限，以及一个基于折外估计值的边际、在所有租户上拟合的表（第4.4节）。

5.5 剩余偏差

点估计的残余偏差。 在估计率为 1–5% 时尤为明显。在多次划分中，

估计值之和是真实计数之和的 1.15 倍（第5至第95百分位数为 1.07 至 1.27）。按估计率组划分，中位数带符号相对误差分别为 +19.1%、 +29.5%、 +11.6% 和 −6.0%。下界裕度根据该误差拟合（第4.4节），因此覆盖率得以保证。第5.6节将该误差与偏差校正基线方法进行比较。

逐记录和逐标准校准。偏差源于校准误差在各标准间的变化，而非记录层面的拟合。对记录取均值后，在其他数据上拟合的校准几乎无偏。在主研究的 166 万组（查询时分数，生产判定）配对数据中，初始校准给出的概率总和是肯定判定数量的 1.01 倍。高于阈值时，每个概率分箱都高估了 0.02 至 0.04；而分数低于阈值的记录按零计，却占肯定判定的 3.95%。将初始校准应用于整个主研究一次后，对于匹配筛选集不到 1% 的标准，总和之比为 2.26；对于超过 20% 的标准，则为 0.90。按估计率分箱后，估计率为 1–5% 时比值为 1.42，5–20% 时为 1.34，超过 20% 时为 0.93，因此这并非向均值回归。在部署校准下进行折外评估时，真实率低于其表格拟合汇总率的运行结果会被高估（带符号误差中位数为 +28.7%，总和之比为 1.38）。高于该汇总率的运行结果则被低估（−7.1%，0.91）。运行结果比率与汇总率之比的对数，与其估计值和真实值之比的对数之间，斯皮尔曼相关系数为 −0.46。这正是先前研究所预测的向汇总率收缩现象 [4, 22]，，不过仅凭这一点无法区分先验偏移与其他原因。总和之比会使最大的筛选集占据更大权重。

先验偏移校正。 标准的先验偏移校正会在两个方向上都过度校正，因此我们不会部署它。它通过期望最大化（EM）重新估计每个查询的先验，并据此重新缩放校准概率 [29]。在主研究中按租户进行交叉拟合后，它使总和之比达到 0.96，但绝对误差中位数增加了一倍以上（66.7%，而非 28.7%）。在最低估计率组中，总和之比从 1.09 降至 0.05（中位有符号误差 −100%）；在最高组中，则从 0.89 升至 1.15。使用保序边际时，其覆盖率最高为 86.3%，中位宽度是真值的 7.63 倍（表5）。

未见标准。 部署校准也涵盖了拟合过程中从未见过的标准。所有经评估的标准都至少在 30 天前保存，因此在拟合窗口期间已经存在。不过，在留出正向真实值估计中，有 249 个（涉及 124 个租户）的租户和标准文本是此前拟合从未见过的。在预先确定的分析中，按租户进行五折拟合并重新运行其拟合过程后，部署校准对其中 92.0% [87。9%、95.5%]进行了折外覆盖。其较低和较高漏检率分别为 1.6% 和 6.4%。

5.6 基线与消融实验

量化基线。 在三种无阈值估计量中，平均校准概率的中位数误差最低（表5）。调整后的分类计数法（ACC） [11] 利用阈值的真实率和假正率校正经阈值处理的计数。其概率版本（PACC） [3] 同样校正校准概率的均值。我们在样本内研究的5个租户折上拟合这两种方法（ACC 的阈值为0.35），并将其样本外点估计与初始校准下的平均结果进行比较。对于真实率低于1%的所有估计值，两种方法都会将结果截为零；在1–5%时则会低估真实率。在真实率高于5%时，两种方法的聚合估计值与真实值相差不超过8%。低于5%时，平均法会高估，但其绝对误差中位数较低：分别为52.7%和34.1%，而另一组为63.7%至100%。

样本量。 中位数误差由校准而非抽样决定。对上述 149 次样本内运行中记录的、超过 1,000 条记录的分数进行离线子采样，将样本量缩减至 $n = 500$ 条已评分记录后，绝对误差中位数为 29.5%，而在 $n = 1,000$ 条记录时为 29.3%。评分时间的第90百分位从 1,788 毫秒降至 1,286 毫秒，因此 1,000 条记录的上限仍符合延迟目标。较大的样本量是为了获得更窄的区间：在大型数据集上，抽样项按 $1/\sqrt{n}$ 缩小，零事件界限 $2.303/n$ 减半（第4.5节）。

阈值化。 对达到或超过阈值的分数进行计数，而非取平均，在基于其他折拟合出的阈值处，绝对误差中位数较低（阈值为 0.35 时为 21.6%）；但误差取决于阈值（阈值为 0.1 时为 157.5%，阈值为 0.9 时为 88.3%），且区间覆盖不足：覆盖率为 0.889，而平均法为 0.913；真实率高于 20% 时，覆盖率为 77.5%。因此，SignalCount 对校准概率取平均。

5.7 稀有标准

该区间对稀有标准同样成立。在真实计数为零的 60 组留出配对数据中，初始校准的区间有 57 组包含 0（95.0%）。在部署校准下，估计率最低的组平均覆盖率为 95.1%（图5）。在样本内研究中，零事件界限在 7 个估计值中触发，这些估计值的真实计数介于 2 至 71 条记录之间，并覆盖了全部这些记录。该界限假设匹配记录的分数不低于阈值，而实际几乎都如此：在主研究中，137 万条抽样记录中，分数低于阈值的记录有 99.58% 得到了负面的生产判定。

最低组的高因子选择（第4.4节）并非这一结果的决定因素：在 200 个租户划分中分别重新选择该因子后，该组的留出覆盖率为 95.2%，而采用部署时的选择则为 95.1%。

5.8 与逐条判定的成本比较

一次估计的成本由样本上限设定的固定值决定，而穷举评分的成本会随筛选集线性增长。在研究开始前的三天内，分类器网关记录了 140 次估计执行，每次处理 1,000 条记录。一次估计使用约 200 万个提示词令牌（上四分位中位数，范围为 170 万至 250 万），即每次调用约 2,000 个。穷举方案会在筛选集的每条记录上运行分类器，并在达到或超过阈值的每条记录上运行判定器；在样本内可靠性数据中，这类记录占 20.0%。我们均按照模型标价计费，其中分类器两种方案均采用实测的每条记录令牌数，判定器则采用两天生产调用中的每次调用平均令牌数。对于 1,000 条记录，穷举评分的成本约为一次估计的 2.1 倍；差异来自判定器。对于 10,000 条记录，成本为 21 倍；对于 100,000 条记录，为 210 倍；对于超过 500,000 条记录的筛选集，则为 1,100 倍。对于最大的筛选集，仅判定器在其峰值观测生产吞吐量下就需要 30.9 分钟，而在其中位数吞吐量下则需要 83.8 小时。

6 经验教训

按租户留出。在拟合校准的租户上测得的覆盖率，高估了新租户实际获得的覆盖率；对于我们的初始校准，高估幅度约为六个百分点（第5.4节和第5.3节）。

对于跨租户共享的校准，留出的单位应为租户，而且测试需要涵盖许多租户。

按判定标准划分的偏差。在记录上汇总后几乎无偏的校准，按判定标准划分时仍可能存在偏差（第5节5.5）：对校准概率取平均的计数会继承表格拟合所依据的判定标准的先验 [22, 33], , 因此按记录汇总的可靠性图无法证明其无偏，必须针对每个判定标准，在正在估计的筛选集上进行检查。

单侧、依赖比率的校准边际。我们的校准误差既不对称，也不会在不同比率下保持不变（第5节5.4），而按估计率组拟合的校准边际同时解决了这两个问题。覆盖率主要来自校准边际，因为已知仅基于二项分布的区间对于分类并计数会出现覆盖不足 [22], , 而将样本量从500增加到1,000几乎没有改变误差（第5节5.3和5.6）。

先验偏移校正。针对这种偏差的标准做法，即对每个查询的先验进行EM重新估计 [29], , 修正了我们的聚合结果，却使绝对误差中位数增加了一倍以上（第5节5.5）。EM假设每个判定标准具有相同的类内分数分布，但当判定器读取每条记录的文本时，这一假设未必成立 [4]；两个方向上的过度校正与该机制一致。偏差校正上线前，应按估计率组进行检查，因为它所修正的聚合结果可能掩盖两个方向上的误差。

半结构化谓词。语义谓词一旦受到样本上限的约束，增长的成本就来自扫描，而扫描中涉及读取 JSON 文本内部字段的结构化谓词（第 5.2 节）。过滤条件读取的字段可作为候选字段，从半结构化数据源中提取并物化为类型化列。

适用性。该方法需要对结构化谓词进行精确计数、为语义谓词提供快速分类器，以及记录分类器分数与昂贵模型判定的成对数据。模型级联（即由廉价模型先筛选行，再交由昂贵的大语言模型处理）是常见做法 [7, 25, 26]。这种级联已经为每一行计算了代理分数，并为每一行计算了预言机判定；只需记录这组成对数据，就满足了校准所需条件。

7 相关工作

近似查询处理。基于采样的近似查询处理通过在线聚合 [14]、存储样本 [1, 5] 以及生产环境中的查询时采样器 [17]，以带有误差界的方式回答聚合查询。Chaudhuri 等人指出，近似查询处理的研究成果尚未进入实际产品 [6]。在近似查询处理中，扫描是成本高昂的部分；而在 SignalCount 中，成本高昂的是语义谓词，因此扫描会精确计算结构化谓词，样本仅用于语义谓词。

基于代理的查询处理。 Kang等人使用低成本代理模型来减少对昂贵预言机的调用，用于在提供保证的情况下进行聚合、选择以及构建语义索引 [18, 19, 21]。ABae [20]和 InQuest [28] 对每条记录运行代理模型，根据其分数进行分层，并在查询时让预言机为样本标注。他们的区间不需要对代理模型作任何假设，但每次查询都必须等待预言机；SignalCount则以校准假设（第4.1节）换取免除这些标签。BlazeIt的查询重写是少数几种与我们一样、在查询时无需调用预言机的系统之一：它在数据摄取时为数据添加标签，并根据这些标签界定误差。但它为每个对象类别训练专用网络，并假设不存在漂移；相比之下，我们的标准是在查询时提出的任意文本，而我们的校准边际是在跨租户数据上拟合，并在留出租户上测量的。Chung等人使用基于大语言模型标签离线训练的代理模型，替换 BigQuery 和 AlloyDB 中的 LLM 过滤条件，以处理已知查询 [7]。代理过滤器会遍历每一行，却不会为聚合提供任何区间；SignalCount支持任意标准，并且只对一个有上限的样本进行评分。

语义算子和大语言模型驱动的引擎。LOTUS [27] 和 BARGAIN [35]保证相对于金标准模型的输出质量；Cortex AISQL [25]则在生产环境中运行代理-预言机级联，并为过滤条件的输出提供精确率-召回率保证。对于聚合结果，UQE [10] 基于嵌入分层样本中的大语言模型标签估计 COUNT，但不提供区间；ThalamusDB [16]和 OLLA [15] 则对多模态数据和文本上的聚合结果给出界限；这三种方法都会在查询时调用昂贵模型。SemStats [32] 研究的问题与此最为接近：它估计非结构化数据上的语义过滤器基数 [30, 36] 。大语言模型会为每个语料库构建索引，并将每个查询映射到该索引上，随后由判定器为样本贴标签。与所有这些方法不同，SignalCount 添加精确的结构化谓词，不会在查询时进行传统的大语言模型调用，不会为每个语料库构建索引，并会返回一个覆盖率经过测量的区间。

基于模型预测的统计估计。预测增强推断（PPI） [2] 将模型预测与少量金标准标签结合，得出有效区间，即使模型存在偏差也不例外。对于 SignalCount，这些标签将是判定器对查询样本给出的判定结果，这会使判定器重新进入查询路径；我们将此留待未来工作（第 8 节）。标签偏移下的 PPI [2] 和经过校正的大语言模型判定器估计值 [23] 会复用来自其他分布的标签，但需要对每条记录运行模型，并假设已知偏移模型。在机制上最接近的是 Xu 等人：他们将按类别分类器校准到生产环境中的大语言模型标注器，并在不再调用大语言模型的情况下报告发生率 [33]。他们的类别是固定的，并在每日批处理中评分；SignalCount 则支持查询时提出的任意标准，由所有租户共享一个校准器，对经过过滤且已精确计数的数据集进行处理，耗时仅数秒。Wilkins-Reeves 等人通过从过往领域中学习得到的偏差分布，扩大代理区间 [31]；我们通过跨租户的分割保序预测拟合校准边际，并在留出租户上测量其覆盖率。用量化术语来说，我们的点估计是校准概率分类计数 [3, 4]。

8 结论

SignalCount 能在数秒内回答带有结构化与语义谓词的 COUNT 查询，而无需在查询时调用预言机。它会准确计算筛选集，并在同一次列式扫描中抽取均匀样本。随后，它使用经过校准的快速分类器对样本进行评分，该分类器依据生产环境评审器记录的判定结果进行校准；针对校准误差，再通过单侧、依赖比率的保序边际扩大区间。在生产环境中，延迟工作负载的 p99 为 3.76 秒。在留出租户上，名义90%区间在 92.0% 的估计值中覆盖了真实计数。

还剩下三个方向。利用样本中少量判定器判定结果进行预测驱动校正 [2]，即可消除校准迁移这一假设。对语义匹配中的结构化列求和或取平均值

可通过同一样本上的比率估计量实现。对表的副本按租户和记录 ID 的固定哈希排序后进行有界时间采样器采样，即可限制最大范围的扫描量。

参考文献

[1] Sameer Agarwal、Barzan Mozafari、Aurojit Panda、Henry Milner、Samuel Madden和Ion Stoica。2013。BlinkDB：针对超大规模数据、具有有界误差和有界响应时间的查询。载于《欧洲计算机系统会议（EuroSys）论文集》。29–42。[2] Anastasios N. Angelopoulos、Stephen Bates、Clara Fannjiang、Michael I. Jordan和Tijana Zrnic。2023。预测增强推断。《科学》382，6671（2023），669–674。[3] Antonio Bella、Cèsar Ferri、José Hernández-Orallo和María José Ramírez- Quintana。2010。通过概率估计器进行量化。载于IEEE国际数据挖掘会议（ICDM）。737–742。[4] Dallas Card和Noah A. Smith。2018。从标注中估计比例时校准的重要性。载于2018年计算语言学协会北美分会：人类语言技术会议（ NAACL-HLT）论文集。

1636 1646 10 18653 1 18 1148 – 。doi: 。/

[5] 2007
v /N - Surajit Chaudhuri、Gautam Das和Vivek Narasayya。 。用于近似查询处理的优化分层抽样。《ACM数据库系统汇刊》

32, 2, 9 2007 10 1145 1242524 1242526 文章（）。doi: 。/ 。

[6] 2017
Surajit Chaudhuri、Bolin Ding和Srikanth Kandula。 。近似查询处理：没有灵丹妙药。载于2017年ACM数据管理国际会议（SIGMOD）论文集。

511 519 10 1145 3035918 3056097 – 。doi: 。/

[7] 。Yeounoh Chung 、 Rushabh Desai 、 Jian He 、 Yu Xiao 、 Thibaud Hottelier 、 Yves- Laurent Kom Samo、Pushkar Khadilkar、Xianshun Chen、Sam Idicula、Fatma Özcan、Alon Halevy和Yannis Papakonstantinou。2026。成本与延迟降低100倍：使用轻量级代理模型进行AI查询近似的性能分析。《ACM数据管理会刊》4，3（2026），23 10 1145 3802002 页。doi: 。/ William G. Cochran。 。

[8] 1977
《抽样技术》（第3版）。John Wiley & Sons。[9] Edith Cohen和Haim Kaplan。2007。使用Bottom-k草图汇总数据。载于ACM分布式计算原理研讨会（PODC）论文集。225–234。[10] Hanjun Dai、Bethany Yixin Wang、Xingchen Wan、Bo Dai、Sherry Yang、Azade Nova、Pengcheng Yin、Phitchaya Mangpo Phothilimthana、Charles Sutton和Dale Schuurmans。2024。UQE：非结构化数据库查询引擎。载于《神经信息处理系统进展》（NeurIPS）。arXiv:2407.09522。[11] George Forman。2008。通过分类量化计数和成本。《数据挖掘与知识发现》17，2（2008），164–206。[12] Jaroslav Hájek。1960。有限总体简单随机抽样的极限分布。《匈牙利科学院数学研究所出版物》5（1960），361–374。[13] James A. Hanley和Abby Lippman-Hand。1983。如果一切顺利，是否就万事大吉？解读零分子。《美国医学会杂志》249，13（1983），1743–1745。[14] Joseph M. Hellerstein、Peter J. Haas和Helen J. Wang。1997。在线聚合。载于ACM SIGMOD 数据管理国际会议论文集。171–182。[15] Chao Hui、Weizheng Lu、Yanjie Gao、Lingfeng Xiong、Yunhai Wang和Yueguo Chen。2026。面向非结构化文本分析的大语言模型驱动的在线聚合。arXiv:2603.08443。已被DASFAA 2026接收。[16] Saehan Jo和Immanuel Trummer。2024。ThalamusDB：多模态数据的近似查询处理。《ACM数据管理会刊》

2, 3, 186 2024 10 1145 3654989 文章（）。doi: 。/ Srikanth Kandula、

[17] Kukjin Lee、Surajit Chaudhuri和Marc Friedman。2019。在微软生产环境大数据集群中近似查询的实践经验。《VLDB基金会会刊》12，12（2019），2131–2142。[18] Daniel Kang、Peter Bailis和Matei Zaharia。2019。BlazeIt：为基于神经网络的视频分析优化声明式聚合和限制查询。《VLDB基金会会刊》13，4（2019），533–546。[19] Daniel Kang、Edward Gan、Peter Bailis、Tatsunori Hashimoto和Matei Zaharia。2020。使用代理实现有保证的近似选择。《VLDB基金会会刊》13，12（2020），1990–2003。doi:10.14778/3407790.3407804[20] Daniel Kang、John Guibas、Peter Bailis、Tatsunori Hashimoto、Yi Sun和Matei Zaharia。2021。使用高开销谓词加速近似聚合查询。《VLDB基金会会刊》14，11（2021），2341–2354。[21] Daniel Kang、John Guibas、Peter Bailis、Tatsunori Hashimoto和Matei Zaharia。2022。TASTI：面向非结构化数据上基于机器学习的查询的语义索引。载于ACM SIGMOD 数据管理国际会议论文集。

1934 1947 10 1145 3514221 3517897 – 。doi: 。/ 。

[22] 2018
Katherine A. Keith和Brendan O’Connor。 。用于推断文档类别流行率的不确定性感知生成模型。载于2018年自然语言处理实证方法会议（*EMNLP*）。4575– 4585。doi:10.18653/v1/D18-1487[23] Chungpa Lee、Thomas Zeng、Jongwon Jeong、Jy yong Sohn和Kangwook Lee。2026。如何正确报告基于大语言模型的评审评估。载于国际机器学习会议（*ICML*）论文集。arXiv:2511.21140。[24] Jing Lei、Max G’Sell、Alessandro Rinaldo、Ryan J. Tibshirani和Larry Wasser- man。2018。用于回归的分布无关预测推断。美国统计学会杂志 113，523（2018），1094–1111。[25] Paweł Liskowski、Benjamin Han、Paritosh Aggarwal、Bowei Chen、Boxin Jiang、Nitish Jindal、Zihan Li、Aaron Lin、Kyle Schmaus、Jay Tayade、Weicheng Zhao、Anupam Datta、Nathan Wiegand和Dimitris Tsirogiannis。2026。Cortex AISQL：面向非结构化数据的生产级 SQL 引擎。载于*ACM SIGMOD* 数据管理国际会议（工业界专题）伴随论文集。400– 412。doi:10.1145/3788853.3803093[26] Paweł Liskowski和Kyle Schmaus。2026。面向语义SQL的流式模型级联。arXiv:2604.00660 [cs.DB].[27] Liana Patel、Siddharth Jha、Melissa Pan、Harshit Gupta、Parth Asawa、Carlos Guestrin和Matei Zaharia。2025。语义算子及其优化：借助LOTUS实现具有准确度保证的大语言模型驱动的数据处理。*VLDB*基金会会刊 18，11（2025），4171–4184。doi:10.14778/ 3749646.3749685[28] Matthew Russo、Tatsunori Hashimoto、Daniel Kang、Yi Sun和Matei Zaharia。2023。加速非结构化数据流上的聚合查询。*VLDB*基金会会刊 16，11（2023），2897–2910。[29] Marco Saerens、Patrice Latinne和Christine Decaestecker。2002。根据新的先验概率调整分类器输出：一种简便方法。神经计算 14，1（2002），21–41。[30] Matthias Urban、Vu Huy Nguyen、Gabriele Sanmartino、Paolo Papotti和Carsten Binnig。2026。图像数据语义过滤器的选择率估计。arXiv:2606.04610。[31] Steven Wilkins-Reeves、Alexandra N. M. Darmon和Deeksha Sinha。2026。随机分布偏移下基于代理推断的估计值水平调整。arXiv:2605.06484。[32] Shihui Xu、Jiayi Wang和Guoliang Li。2026。弥合差距：非结构化数据语义查询的基数估计。*ACM*数据管理会刊

4, 3, 147 2026 26 10 1145 3802024 文章（），

[33] 2026
页码。doi: 。/ Zehao Xu、Tony Paek、Kevin O’Sullivan和Attila Dobi。 。全局校准，处处测量：扩展A/B实验中的大语言模型流行率测量。arXiv:2602.16111。[34] Bianca Zadrozny和Charles Elkan。2002。将分类器分数转换为准确的多分类概率估计。载于*ACM SIGKDD*知识发现与数据挖掘国际会议论文集。694–699。[35] Sepanta Zeighami、Shreya Shankar和Aditya Parameswaran。2025。降低成本，不牺牲准确度：有保证的大语言模型驱动的数据处理。*ACM*数据管理会刊

3, 6 2025 10 1145 3769776 ( )。doi: 。/

[36] Andreas Zimmerer、Claudius Kühn、Yang Li、Mihail Stoian、Renata Borovica- Gajic和Andreas Kipf。2026。SemCEB：语义算子的基数估计基准。arXiv:2606.23081。