ByteDance | Seed

# FutureX-Pro：将未来预测拓展至高价值垂直领域

字节跳动 Seed、香港科技大学、佐治亚理工学院、斯坦福大学、普林斯顿大学

完整作者名单见贡献部分

arXiv:2601.12259v1 [cs.AI] 18 Jan 2026

## 摘要

在 FutureX 的基础上，该基准测试为通用未来预测建立了实时基准测试。本报告介绍了 **FutureX-Pro**，包括 **FutureX-Finance**、**FutureX-Retail**、**FutureX- 公共卫生**、**FutureX-NaturalDisaster** 和 **FutureX-Search**。这些项目共同构成一个专门框架，将智能体未来预测拓展至高价值垂直领域。尽管通用型智能体在开放域搜索方面表现出色，但其在资本密集型和安全关键行业中的可靠性仍缺乏充分研究。FutureX-Pro 聚焦于四个对经济和社会至关重要的垂直领域：金融、零售、公共卫生和自然灾害。我们对智能体大语言模型（LLM）开展基准测试，评估其在入门级但基础性预测任务中的表现，任务范围涵盖市场指标和供应链需求预测，以及流行病趋势和自然灾害追踪。我们沿用 FutureX 的无污染实时评估流程，评估当前最先进水平的智能体大语言模型是否具备工业部署所需的领域 grounding。研究结果揭示了通用型推理能力与高价值垂直应用所要求的精准度之间的性能差距。

**联系邮箱：** Jiashuo Liuliujiashuo77@gmail.com，Wenhao Huang huang.wenhao@bytedance.com**项目主页：** https://futurex-ai.github.io/

## 1 引言

大语言模型智能体的发展使系统能够收集信息并对未来进行推理。我们之前的工作 FutureX [18], 评估了这些能力在广泛通用主题上的表现，并引发了广泛讨论，获得了超过一亿次曝光。然而，智能体智能的真正考验并不在于通用任务，而在于那些决定经济与社会稳定的领域。**FutureX-Pro**通过将未来预测拓展至高价值垂直领域，弥补了这一关键空白，明确地将评估重点从通用广度转向高风险深度。

FutureX-Pro 聚焦于四个垂直领域，选择它们不仅是因为其复杂性，更是因为它们对全球系统具有深远影响：

- 金融（市场效率）：金融市场代表着全球经济的脉搏。准确预测企业收入或宏观经济指数等入门级指标的能力，是理解经济健康状况的基础。该领域要求严格的定量推理能力，以及从高频市场信号中滤除噪声的能力。
- 零售（资源优化）：零售预测能提升全球供应链的效率。预测产品需求和销售趋势，对于减少库存浪费和优化资源

分配至关重要。该领域考验智能体解读消费者行为和动态市场趋势的能力。

- 公共卫生（社会韧性）：在全球卫生挑战频发之后，追踪传染病的能力是社会韧性的基石。准确预测流行病指标有助于支持预警系统和公共卫生应急准备。在此领域评估智能体，可以验证其在生物监测场景中的潜在可靠性。

- 自然灾害（环境安全）：随着气候波动加剧，预测环境灾害（例如台风、洪水）的发生及其影响，已成为至关重要的安全要求。该领域的成功有赖于对气象数据的解读，以预测风险，从而作为衡量智能体在高风险安全场景中实用性的代理指标。

尽管近期的智能体大语言模型擅长综合多样化的开放网络信息，但高价值垂直领域在数据异质性和来源权威性方面提出了独特挑战。一名擅长总结政治新闻的通用型智能体，可能无法正确解读结构化财务披露、气象图表或流行病学表格。FutureX-Pro 探究了这一差距，评估智能体能否超越通用网页搜索，有效地将其推理建立在权威的领域专属数据源之上。

FutureX-Pro 借鉴了 FutureX「从设计上杜绝数据污染」的架构，通过聚焦可验证的未来事件，确保评估稳健可靠。该基准测试通过模拟初级分析师的核心职责来评估基础能力：

- 金融：分析实时报告，预测短期财务表现。
- 零售：综合电子商务数据，预测销售排名和趋势。
- 公共卫生：解读官方公告，预测近期疾病指标。
- 自然灾害：处理气象数据，预测灾害影响。

本报告介绍了特定领域数据集的设计与构建，并对领先的智能体大语言模型进行了比较分析。我们旨在清晰评估当前人工智能系统在需要对塑造我们世界的高价值领域进行推理时的表现。

在我们的预测基准测试之外，我们还设计了 **FutureX-Search**，它利用 FutureX 和 FutureX-Pro 的历史数据，将针对已揭晓事件的预测任务转化为纯粹的搜索与检索挑战。本报告其余部分将详细介绍四个垂直领域以及 FutureX-Search 基准测试 [1]。

**加密货币领域。** 有关加密货币领域的评估，请参阅我们的专用基准测试 CryptoBench [6]。

**流程实现。** 有关 FutureX-Pro 构建流程的详情，请参阅 FutureX [18] 中的方法论。

**实时更新。** 新的垂直领域任务即将纳入我们每周举办的 FutureX 竞赛，详见 https://huggingface.co/datasets/futurex-ai/Futurex-Online。

## 2 FutureX-Finance

金融预测代表了复杂自适应系统的巅峰，在这一系统中，智能体必须应对波动性、解读市场情绪，并处理定量报告 [4, 7, 20]。**FutureX-Finance** 评估智能体作为量化分析师开展工作的能力，要求其对全球交易所的股市指标进行精准的数值预测。

[1]本报告仅介绍 FutureX-Pro 和 FutureX-Search。有关详细的构建流程，请参阅 FutureX [18, 第 3 节2]。

## 2.1 数据构建与市场结构

为确保全面评估对全球经济的理解，我们整理了一个数据集，涵盖全球两个最大经济体的主要股票。该数据集追踪由*150* 个不同实体组成的多元化投资组合，这些实体经过筛选，代表各个市场的结构核心。

**市场覆盖率**。

- 美国市场（纳斯达克/纽约证券交易所）：我们从纳斯达克100指数和标准普尔500指数中选取*100*家成分公司。这些公司包括英伟达（*NVDA*）和微软（*MSFT*）等全球科技巨头，以及福泰制药（*VRTX*）等生物科技领军企业。

- 中国*A*股市场（上海证券交易所/深圳证券交易所）：我们选取*50*家顶尖企业，它们代表着中国经济的中坚力量，例如贵州茅台（*600519.SH*）和宁德时代（*300750.SZ*）。

**行业构成分析**。如图1所示，该数据集反映了两国鲜明的经济特征。

- 美国样本的权重主要集中在信息技术和通信服务领域，体现了美国股票市场以创新为驱动的特点。

- 中国样本呈现出更加均衡的分布，在工业、原材料和必需消费品领域占有显著比重，体现了中国经济的制造业与消费基本面。

这种双市场结构考验智能体能否根据不同的监管环境和经济驱动因素调整其预测逻辑。

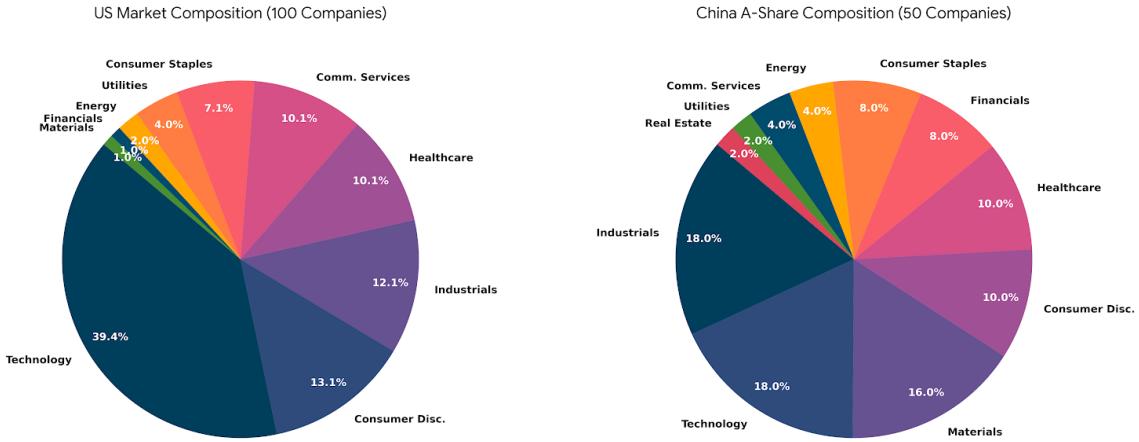


**图1** 150 个跟踪实体的行业分布。美国子集侧重信息技术，而中国子集则突出工业和原材料。

**任务制定**。 我们定义了三种不同的预测任务，用以模拟交易员评估短期市场走势的工作流程。

任务类型*1*：点位预测。 智能体必须预测某个特定术语在单个目标交易日的确切数值。

*“{*特斯拉公司 *(ID: TSLA)}在{2025-12-25}的{*收盘价*}*是多少？”

任务类型*2*：窗口极值预测。智能体必须识别某个指标在未来 $N$ 个交易日窗口内的最大绝对值。这考验了对波动范围进行建模的能力。

“从.....开始的 *{5}* 个交易日内，*{*中国移动 *(ID: 600941.SH)}* 的*{*最高价格*}*是多少？”

任务类型*3*：方向性动量预测。 智能体关注价格变动，例如预测指定时间窗口内的最大涨幅。

“什么　　将在 *{3}* 个交易日内，*{NVIDIA (ID: NVDA)}* 的*{highest Positive Change}*是多少.....？

## 2.2 评估指标

金融市场要求极高的精度；哪怕很小的百分比误差也可能导致巨大的套利损失。因此，我们采用高敏感度误差指标。

设 $\hat{y}$ 为预测值， $y$ 为实际市场值。分数 $S$ 使用陡峭的线性惩罚函数计算：

$$S = \max\left(0, 1 - 20 \times \frac{|\hat{y} - y|}{|y|}\right) \tag{1}$$

解读。“20”的系数设定了严格的容差范围。

- 相对误差为 1%（$0.01 \times 20 = 0.2$）时，得分为 0.8。
- 相对误差为 5%（$0.05 \times 20 = 1.0$）时，得分为 0.0。

任何偏离真实值超过 5% 的预测均得零分。这一严格标准确保高分反映的是真实的市场洞察，而非侥幸得出的近似结果。

## 2.3 结果

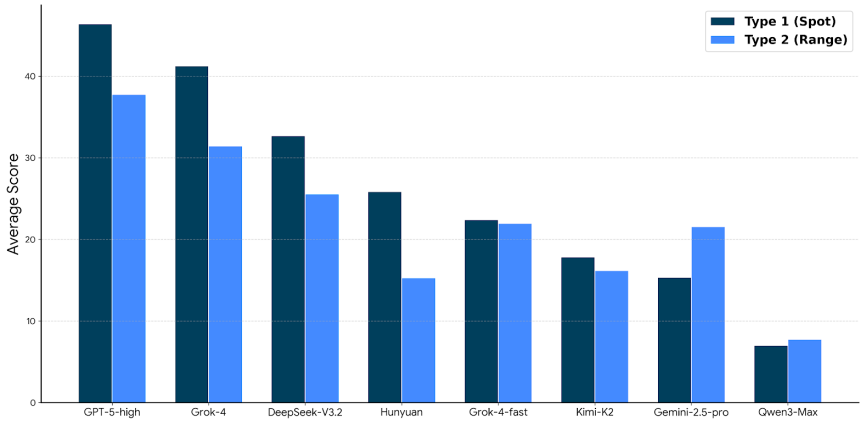


**图 2** FutureX-Finance 在 $24^{\text{th}}$ 10 月至 $28^{\text{th}}$11 月期间的结果。

- 洞察 1：性能分层。评估结果显示出明显的性能群组。GPT-5-high 和 Grok-4 在基准测试中领先，类型1平均得分分别达到 46.37 和 41.25。这表明，这些模型更善于将数值预测范围缩小至 5%误差容限以内。相比之下，通义千问 Qwen3-Max 和 Kimi-K2 等模型得分明显较低，往往无法达到评分函数要求的精度阈值。

- 洞察 2：任务复杂度的影响。从类型1（点位预测）转向类型2（窗口极值预测）时，性能通常会下降。例如，混元在类型1中的得分为 25.85，在类型2中则降至 15.29。这一趋势表明，在多日

窗口中识别极值，比预测单个点更具挑战性。不过，DeepSeek-V3.2 在这两种任务类型中的表现都相对稳定，这意味着它在处理时间序列数据时具有更好的一致性。

- 洞察 3：指标敏感性与模型局限性。没有模型的平均得分超过 50。这一绝对表现偏低，凸显了该基准测试的难度。模型虽然能够处理财经新闻，但要将这些信息转化为误差低于 5% 的数值预测，仍是一项重大挑战。严格的惩罚函数能有效筛除宽泛或模糊的估算，只奖励高精度的输出。

## 3 FutureX-Retail

在快节奏的电子商务行业中， [12], 准确的需求预测是供应链效率的引擎。**FutureX-Retail** 模拟库存分析师的工作，要求智能体根据可收集到的历史趋势和市场信号预测产品销量。与此前信息可通过公开渠道获取的领域不同，这一领域面临着数据稀缺挑战：历史销售数据往往属于专有信息，或转瞬即逝。因此，智能体必须结合提供的内部快照（HTML）与外部市场情报，对消费者需求进行建模。

### 3.1 数据构建与分类体系

该数据集聚焦于竞争极为激烈的跨境电子商务市场，数据来源是快速发展的全球平台 Temu。

**产品分类体系**。 该数据集涵盖*240* 种不同的产品，并将其划分为*4* 个主要类别和*24* 个子类别的层级分类体系，以确保消费者行为模式的多样性。

- 电子产品： 包括耳机、键盘和智能手表等波动性较高的商品。
- 服装和配饰： 包括女式连衣裙、男士T恤和珠宝等季节性商品。
- 家居生活： 包括家具、厨房用具和照明用品等耐用品。
- 美容和个人护理： 包括护肤品、化妆品和婴儿护理用品等快速消费品（FMCG）。

### 3.2 任务配置与变体

为了诊断智能体在不同信息密度和不确定性建模水平下的能力，我们设计了一个 $2 \times 3$ 实验矩阵，由此形成六种不同的任务配置（例如任务 1a、2b）。示例任务见表 1。

**输入条件（信息密度）**。我们改变提示词中提供的历史背景，以测试静态提取和动态趋势外推能力。

- **A类：仅快照上下文（静态）**。
  智能体被提供 仅提供了7天前（$T_{-7}$）产品页面的HTML源代码 .
  – 挑战：智能体必须仅根据快照中嵌入的元数据（例如“最近售出”标签、评论增长）推断销售速度，而没有历史数值基线。– 输入：截至 $T_{-7}$的HTML快照。

- **B类：稀疏时间序列上下文（动态）**。
  除了来自 $T_{-7}$ 的 HTML快照外，还会明确向智能体提供 14 天前的 exactsalesvolume（$T_{-14}$）。

  – 挑战： 这使智能体能够——也要求智能体——通过计算 $T_{-14}$ 与估算的 $T_{-7}$ 状态之间的销量差值，执行趋势外推，从而预测目标 $T_0$。– 输入： $T_{-14}$ + 时的销售量； $T_{-7}$ 时的 HTML 快照。

**输出定义（预测粒度）**。我们要求智能体以三种复杂度逐渐提高的格式输出预测，以评估其量化不确定性的能力。

• Level 1：确定性点预测（任务 1）智能体必须输出一个确定的整数，代表最可能的销售量。{"predicted_sales_volume": 551}

• Level 2：Top-K 概率预测（任务 2）智能体必须确定最可能的前三种 结果，并为每种结果分配一个概率，概率总和为 1.0。这项任务用于测试智能体对高置信度候选结果的校准能力。{"551": 0.50, "550": 0.48, "552": 0.02}

• Level 3：完整分布预测（任务 3）智能体必须列出所有 合理的结果及其对应概率。这项任务将严格评估智能体对“尾部风险”和需求方差全貌的理解。{"551": 0.50, "550": 0.26, "552": 0.13, ..., "549": 0.02}

**表 1** FutureX-Retail 中 $2\times 3$ 个任务的示例。

| **任务类型** | **1a** | **1b** | **2a** | **2b** | **3a** | **3b** |
|---|---|---|---|---|---|---|
| 核心任务示例 | 你是一个能够预测未来事件的智能体。需要预测的事件是：“预测销售商家 {***} 在美国地区的 Temu 上商品 {***} 于 2025-12-05 14:00 GMT+8 的销售量。” | | | | | |
| 历史背景 | | The 销售截至2025年的销量11月21日, 14:00 GMT+8 was {***}。 | | The 销售额截至2025年的销量11月21日, 14:00 GMT+8 was {***}。 | | The 销售额2025 年的成交量11-21, 14:00 GMT+8 was {***}。 |
| 要求 | | | 列出恰好三个你认为最有可能的销售量。Give以概率分布的形式给出你的预测。列出这些三个销售量，并按概率从高到低排列（从 从最高到最低），以及对应的概率（保留两位小数的数字）。三个概率之和必须为1。我将根据真实结果是否来评估你的预测属于你预测的销售量范围以及你提供的概率分布你提供的内容。 | | 请以概率分布的形式给出你的预测。列出你认为所有可能的销售量可能。请按以下方式给出这些销售量：按概率从高到低排序，并列出各销售量及其概率（保留两位小数）。所有概率之和必须为 1。我将评估你的预测。根据真实结果是否落在您预测的销售量范围内以及您提供的概率分布来评分。你提供的贡献。 | |
| | # 注意：<br>1) 您需要从提供的 HTML 信息中找出该产品的销售量。<br>2) 你可以使用搜索工具收集必要的信息。<br># 输出格式（严格）<br>必须严格返回两行，不得包含其他文本： | | | | | |
| 输出格式 | 1) **预测行**：一个 JSON 对象，包含在 <预测>内标签， 其中包含你最终预测的销售量。 | | 1) **预测行**：一个由 <预测> 标签括起的 JSON 对象，其中包含你最终预测的销售量和概率。 | | | |
| 输出示例 | <预测>{{"预测_销售_量": 551}} </预测-测><br><我的推理>[[The ...]] </myrea-soning> | | <预测>{{"551": 0.50, "550": 0.48, "552": 0.02}} </预测><br><我的推理>[[The ...]] </myrea-soning> | | <预测>{{"551": 0.50, "550": 0.26, "552": 0.13, "553": 0.09, "549": 0.02}} </预测><br><我的推理>[[The ...]] </我的推-理> | |

鉴于历史数据在零售领域的重要性，以下我们主要评估模型在任务1b、2b和3b中的表现。

### 3.3 评估指标

鉴于电子商务数据的粒度各不相同（例如，精确的销售数量与模糊的“10k+”徽章），我们采用混合评分机制。令 $y$ 表示真实值（GT）。

- 如果 $y$ 为 None，则该样本不纳入评估。
- 如果智能体的回答为空或无法解析，则得分为 0。

对于有效回答，评分定义如下：

**基础评分函数。** 我们针对单个预测值 $v$ 与真实值 $y$，根据 $y$的粒度，定义评分函数 $S(v, y)$。

情况 *1*：真实值精确（$y \leq 999$）。当有高精度数据可用时，我们使用带有严格容差阈值 $\epsilon = 0.05$的线性衰减指标。得分根据相对误差线性衰减：

$$S(v, y) = \max\left(0, 1 - \frac{|v - y|}{y \times \epsilon}\right) \tag{2}$$

如果预测偏差超过阈值 $\epsilon$，得分将降至 0。

情况 *2*：粗粒度真实值（$y > 999$）。当平台仅显示模糊徽标（例如“1.5k+”）时，我们会将 $y$映射到一个闭区间范围 $R_y = [L, U]$（例如，“1.5k+” $\rightarrow [1500, 1599]$）。评估将变为二元范围检查：

$$S(v, y) = \mathbb{I}(v \in R_y) = \begin{cases} 1 & \text{if } L \leq v \leq U \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

**任务特定聚合。**基于基础函数 $S(v, y)$，我们计算每种任务类型的最终得分：

任务 *1*（确定性预测）。智能体输出一个单一值 $\hat{y}$。得分定义如下：

$$\text{Score}_{\text{task1}} = S(\hat{y}, y) \tag{4}$$

任务 *2* 和 *3*（概率预测）。智能体输出一组预测值 $N$及其对应的概率 $\{p_1, ..., p_N\}$，其中 $\sum p_i = 1$。最终得分为智能体预测分布下的期望得分：

$$\text{Score}_{\text{prob}} = \sum_{i=1}^{N} p_i \times S(v_i, y) \tag{5}$$

该指标会奖励那些对准确值赋予更高置信概率的智能体。

### 3.4 结果

我们使用内部智能体评估框架（配备搜索工具）对 18 个领先的大语言模型进行全面评估，包括 GPT-5.1-High、GPT-5-High、GPT-5-Medium、Claude-Opus-4.1、Claude-Sonnet-4.5、Kimi- K2-0905、Kimi-K2-1104、Kimi-K2-thinking、Grok-4、DeepSeek-V3.1、Gemini-2.5-flash[2], o3-High、GLM-4.6、GLM-4.5、DeepSeek-R1 和 Qwen3-Max。

图 3 展示了领先大语言模型在 FutureX-Retail 上的表现，并凸显了以下三个关键趋势：

[2]由于 API 错误率较高，我们未在此处纳入 Gemini-2.5-pro 的结果。

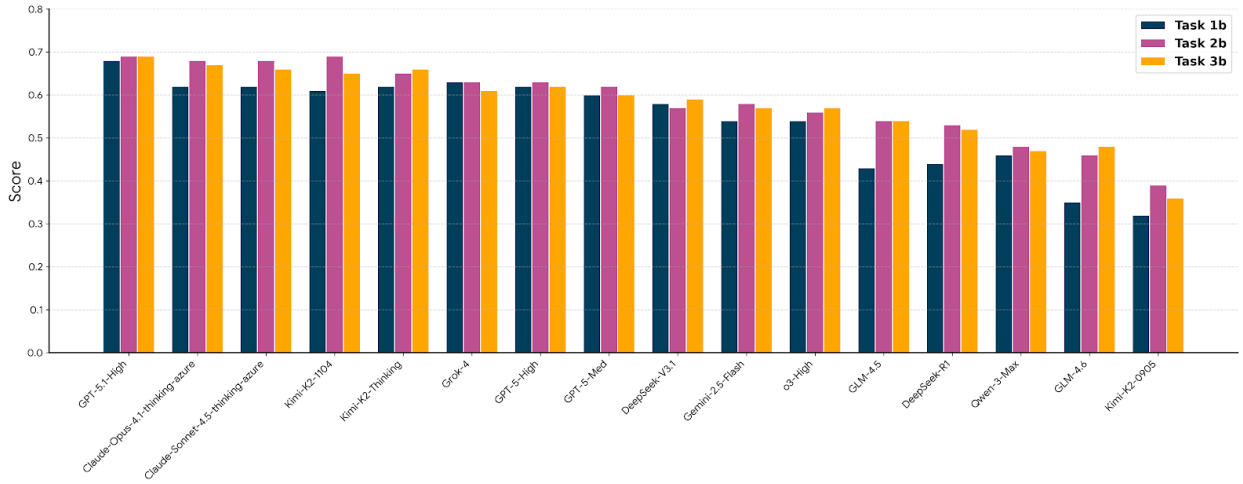


**图3** FutureX-Retail 在 $12^{\text{th}}$ 年11月至 $3^{\text{rd}}$年12月的结果。

- 洞见 1：“概率优势”。大多数模型，包括 Claude-Opus 和 Kimi-K2，在任务2b（前三）和任务3b（完整分布）中的得分均高于任务1b（点预测）。这表明，尽管智能体可能难以准确给出整数销售额，但它们的不确定性校准能力很强。它们能够成功识别真实值的“高概率邻域”，而在库存风险管理中，这往往比脆弱的点估计更有价值。

- 洞察 2：SOTA前沿的稳定性。GPT-5.1-High 和 Grok-4 在所有三种任务类型中都表现出非凡的稳定性（1b、2b 和 3b 的得分分别为 $\approx 0.69$ 和 0.63）。与不同任务之间表现出较大方差的较低级别模型不同，SOTA模型展现出对市场稳健的内部表征，有效地将不确定性分布收敛为高置信度预测。

### 3.4.1 历史背景的影响

为量化历史背景的影响，我们对任务 1a（仅快照）和任务1b（历史增强）的模型性能进行了比较分析。如图4所示，纳入历史数据后，所有模型的性能均大幅提升。这一实证证据验证了将 B类任务设计为评估趋势外推能力的必要基准测试。

### 3.4.2 上下文学习敏感性分析

我们进一步考察少样本示例（示例）的特征如何影响模型行为。实验表明，所提供的示例与模型的输出分布在两个关键维度上存在显著的正相关：预测数量和最大概率。

**1. 对输出粒度（预测数量）的敏感性**。模型与所提供示例的粒度表现出高度一致性。

- 单一来源影响：增加示例中预测值的数量，会直接促使模型输出更大的支撑集。– 当示例包含*5* 个值时，平均输出数量为*5.1* （众数：4.8）。– 当示例增加到*8* 个值时，平均输出数量跃升至*7.3* （众数：6.9），最大数量则从 9.6 增至 14.0。

- 混合来源插值：当提供多个粒度不同的示例（例如，一个示例包含 5 个值，另一个包含 8 个值）时，模型的行为会进行插值，但略微偏向较大的值，从而生成以两个锚点之间为中心的分布。

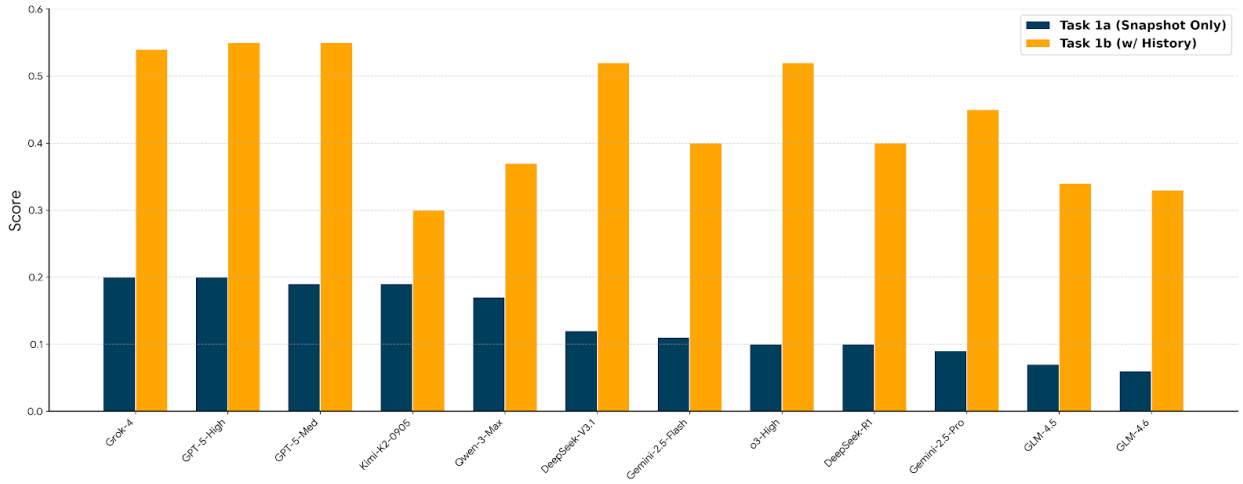


**图4** FutureX-Retail 上任务 1a 与任务1b的比较。

**2. 对置信度校准（最大概率）的敏感性**。示例中展示的置信度会影响模型自身的确定程度，但对平均统计量的影响会逐渐减弱。

- 任务 2（前三项预测）：随着示例的最大概率从 $0.5 \to 0.7 \to 0.9$增加：– 模型输出的<styled id='5'>平均最大概率略有上升（$0.51 \to 0.56 \to 0.58$）。– 然而，观测到的<styled id='10'>最高概率显著提升（$0.76 \to 0.76 \to 0.84$），这表明高质量示例能够激发模型进行高置信度预测的潜力。

- 任务 3（完整分布）：观察到类似趋势。将示例概率从 $0.5 \to 0.7$提高后，模型的平均最大概率从 $0.39 \to 0.43$上升。

**3. 模型特定行为**。在这些条件下，不同架构呈现出各自鲜明的“个性特征”：

- Grok-4（“长尾”专家）：它是一个结构性离群者，与所有其他模型相比，始终生成显著更多的预测值（平均值和最大值），这表明它倾向于全面覆盖分布。

- GPT 系列（“平等主义”推理者）：GPT 模型展现出更均衡的校准表现。它们的概率分布往往集中在中心区域，峰值较低，体现出一种更“平等主义”或保守的不确定性量化方式，避免对任何单一结果过度自信。

## 4 FutureX-PublicHealth

在全球卫生挑战 [1, 5], 的背景下，追踪传染病的能力是社会韧性的基石。**FutureX-PublicHealth**评估智能体作为实时生物监测工具的效用。与一般查询不同，该领域的任务要求智能体处理结构化的流行病学报告、解读人口统计分层，并以高时间分辨率追踪病原体动态。

### 4.1 数据构建与来源

为模拟公共卫生分析师的工作流程，该领域严格以高频官方公告为依据。

**数据来源权威性**。数据仅来源于国家公共卫生机构，具体包括美国疾病控制与预防中心（美国疾病控制与预防中心）和中国疾病预防控制中心（中国疾控中心）。这些来源是呼吸道病毒监测的权威标准。

**规模与粒度**。该数据集由*70* 个周期性周度事件模板扩展而成，包含*382* 个具体变量。该领域的一个显著特点是时间频率较高；所有任务均每周更新，因此智能体需要区分最新报告期与历史数据。如图5所示，这些任务涵盖多个维度，包括特定疾病、地理区域和人口群体（例如 0–4 岁）。

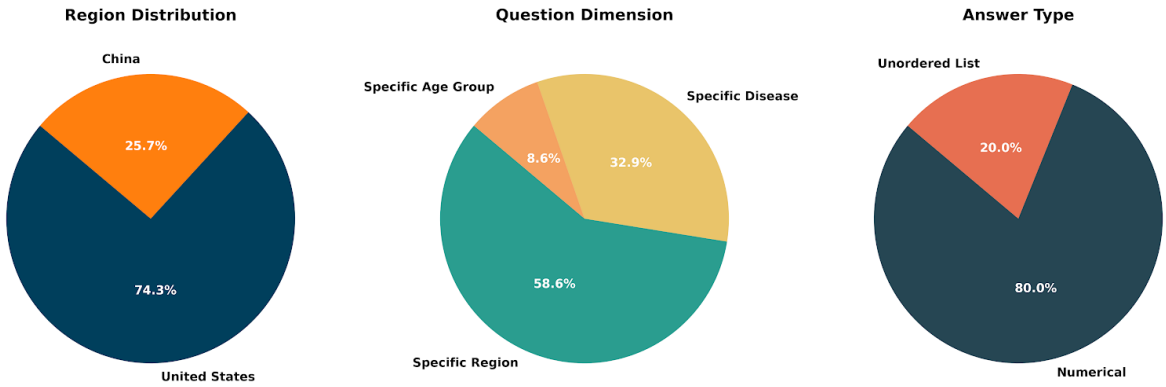


**图5** FutureX-PublicHealth 统计数据。该基准测试聚焦于不同流行病学维度的高频每周更新。

**任务定义**。FutureX-PublicHealth 中的任务要求解析分层数据结构，例如病毒谱系细分或按年龄分层的病原体阳性率。以下展示两个任务模板：

示例*1*：病毒分型（美国疾病控制与预防中心）。

> “根据美国疾病控制与预防中心的流感监测报告（*FluView*）中公共卫生实验室的数据，截至 *{Date}*，*{t}* 的阳性病例数是多少？”

- 变量 $t$ 从 {A (H3)、A (H1N1) pdm09、A（未进行分型）、B（维多利亚谱系）、B（谱系未指定）} 中选取。
- 该任务测试智能体在密集的实验室报告中区分特定病毒亚型（例如 H3 与 H1N1 pdm09）的能力。

示例 *2*：人口统计分析（中国疾控中心）。

> “根据中国疾控中心 *{Date}*的数据，在 *0-4* 岁年龄组中，*{t}*呼吸道样本核酸检测阳性率最高的三种病原体是什么？”

- 变量 $t$ 从{Outpatient influenza-like cases, hospitalized severe acute respiratory infection cases}中选取。
- 这项任务要求交叉比对病原体流行率与人口统计数据，以提取有序或无序的风险列表。

## 4.2 评估指标

鉴于所需的精确度，我们针对答案类型制定了严格的评分规程。

**数值预测评分**。我们采用相对误差评分机制 。设 $y$ 为真实值，$\hat{y}$ 为智能体的预测。我们根据变量的波动性定义严格容差阈值 $\epsilon \in \{0.05, 0.10\}$ 。评分 $S_{\text{num}}$ 计算如下：

$$S_{\text{num}} = \begin{cases} 0 & \text{if } \frac{|\hat{y}-y|}{y} > \epsilon \\ 1 - \frac{|\hat{y}-y|}{y \times \epsilon} & \text{if } \frac{|\hat{y}-y|}{y} \leq \epsilon \end{cases} \tag{6}$$

任何偏差超出阈值 $\epsilon$ 的预测都会获得 0 分，严格惩罚幻觉式精确性。

**无序列表/分类评分。** 对于涉及类别状态或集合的任务（例如，受影响省份列表），我们使用基于集合的重叠度量。令 $G$ 为真实元素集合， $P$ 为预测元素集合：

$$S_{\text{cat}} = \begin{cases} 1.0 & \text{if } P = G \text{ (Exact Match)} \\ 0.5 & \text{if } P \subset G \text{ or } P \cap G \neq \emptyset \text{ (Partial Match)} \\ 0.0 & \text{otherwise (Failure)} \end{cases} \tag{7}$$

这种严格的评估确保高分反映的是真正的领域 grounding，而非概率性猜测。

## 4.3 结果

我们考察了 8 个领先的智能体大语言模型：GPT-5-High、Grok-4、Kimi-K2-thinking、Gemini-2.5-pro 和 Grok-4-fast（通过工具使用进行评估），以及通义千问Qwen3-Max、DeepSeek-V3.2-Exp 和元宝（通过网页浏览进行评估）。11 月 $14^{\text{th}}$ 至 11 月 $28^{\text{th}}$ 期间数据上的表现汇总于 图6。

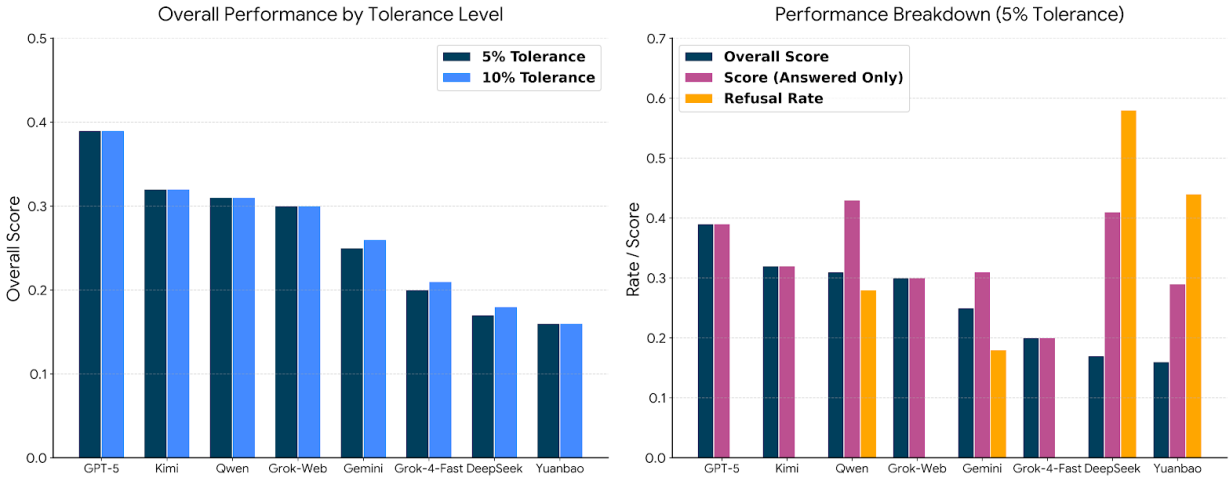


**图6** FutureX-PublicHealth 在 11 月 $14^{\text{th}}$ 至 11 月 $28^{\text{th}}$期间的结果。

根据公共卫生领域的性能指标，我们总结了关于智能体可靠性和精确率-召回率权衡的三个关键观察结果。

- **洞察 1：精确率-召回率权衡。**仅评估已回答的查询时，结果出现了显著逆转。GPT-5-High 和 Kimi-K2-thinking 凭借较高的覆盖率（拒答率为 0%）领先总体排行榜，但在已回答任务的纯准确率方面，Qwen3-Max 和 DeepSeek-V3.2-Exp 显著优于它们。

  – 如果将拒答排除在分母之外，通义千问Qwen3-Max和DeepSeek-V3.2-Exp将升至第$1^{st}$ 和第 $2^{nd}$ 位，超过GPT-5-High。– 这表明，通义千问Qwen3-Max和DeepSeek-V3.2-Exp在高置信度样本上具有更高的领域精确度，但检索阈值过于保守。它们更像是“高精度专家”，而非“通用型监测器”。

- **洞察 2：医疗场景中的安全策略差异。** 对于某些模型，公共卫生领域的拒答率显著高于自然灾害领域，这反映出它们在医疗/健康主题上采取了更严格的安全对齐策略。

  – DeepSeek-V3.2-Exp 的拒答率极高，接近 60%，其次是元宝（44%）。– 相比之下，GPT-5-High、Kimi-K2-thinking 和 Grok-4 的拒答率均为 0%。– 这种分化凸显了策略上的差异：激进型智能体优先考虑实用性（尽力给出估算），而保守型智能体优先考虑安全（在健康相关情境中避免潜在错误信息，即使这会牺牲实用性）。

## 5 FutureX-NaturalDisaster

随着气候波动加剧，预测环境灾害的能力已成为至关重要的安全要务 [14, 17]。**FutureX-NaturalDisaster** 用于评估智能体作为生物监测和环境监测工具的能力。该领域要求智能体综合多模态信号（如气象数据、地质报告），并进行严谨的时空推理，以预测灾害的发展轨迹及其影响。

### 5.1 数据构建与来源

为满足高风险安全场景所要求的可靠性，FutureX-NaturalDisaster 完全基于国家及国际政府机构发布的权威数据。

**来源权威性。**我们汇总了来自官方机构的实时数据，包括美国国家海洋和大气管理局（NOAA）[3]、美国地质调查局（USGS）[4]、中国国家气象中心（NMC）[5],以及英国气象局。

**规模与多样性。**该数据集包含*92* 种不同的事件模板，并实例化为*446* 个具体变量。如图7所示，该基准测试覆盖全球 15 个区域，涵盖多种灾害类型，从地震等高频事件到干旱等缓慢发生的现象。

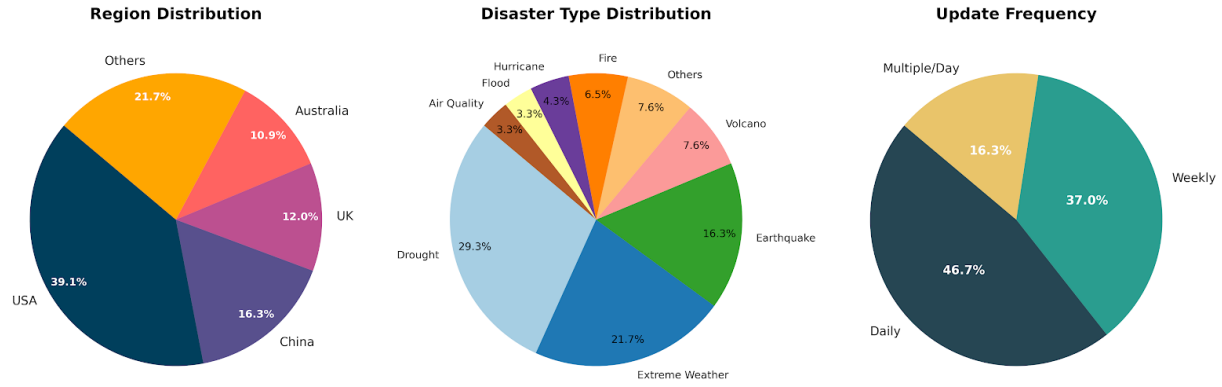


**图 7** FutureX-NaturalDisaster 的统计数据。该数据集涵盖不同区域和灾害类型，更新频率各异。

**任务定义。**该领域的任务模拟风险分析师监控实时数据流的工作流程。通常分为数值预测和状态分类。下面展示两个任务模板：

[3] https://www.drought.gov/[4]
https://www.usgs.gov/programs/VHP/[5]
https://www.nmc.cn/

模板示例 *1*：数值预测（干旱监测）。

> *"{Date}* 根据美国国家海洋和大气管理局的国家综合干旱信息系统，美国将有多大比例被分类为*{t}*？"

- 变量 $t$ 从{异常干燥、中度干旱、严重干旱、极端干旱、特重干旱}中选择。
- 智能体必须解读定量报告，并输出精确的百分比数值。

示例 *2*：状态分类（气旋警报）。

> *"{Date}* 根据雅加达热带气旋警报中心发布的热带气旋展望，预计未来 *{$t_1$}* 小时内热带气旋的潜在发展情况为 *{$t_2$}*？"

- 变量 $t_1$ 从{24、48、72}中选择。
- 变量 $t_2$ 从 {Low/Medium/High} 中选取。
- 该 智能体必须找到特定的区域公告并提取概率预报。

## 5.2 评估指标

我们采用与公共卫生领域（第 4 节）相同的严格评分规程，以确保在各高风险领域间保持一致。

**数值预测**。对于定量指标（例如病例数），我们使用相对误差评分机制函数，并采用严格容差阈值（$\epsilon \in \{0.05, 0.10\}$）。随着误差接近阈值，得分线性下降；一旦超过阈值，得分便立即归零。

**列表生成**。对于需要识别主要病原体（无序列表）的任务，我们采用基于集合的重叠度量：

- *1.0* 分：完全匹配（识别出所有正确元素，且没有幻觉）。
- *0.5* 分：部分匹配（仅识别出部分正确元素，或包含了错误元素）。
- *0.0* 分：未能检索到有效数据。

## 5.3 结果

我们考察了 8 个领先的智能体大语言模型：GPT-5-High、Grok-4、Kimi-K2-thinking、Gemini-2.5-pro 和 Grok-4-fast（使用工具评估），以及通义千问Qwen3-Max、DeepSeek-V3.2-Exp 和元宝（通过网页浏览评估）。 $24^{\text{th}}$ 10 月至 11 月 $28^{\text{th}}$ 期间数据上的表现汇总于 图8。

基于自然灾害领域的定量结果，我们总结出关于模型行为和可靠性的四项关键见解。

- **见解 1：激进检索型智能体占据主导地位**。GPT-5 和 Grok-web 凭借极低的拒答率（$\leq$ 1%）确立了明显优势。关键在于，这些事件均有客观的政府数据（例如美国国家海洋和大气管理局、USGS）作为依据，因此这些模型的成功证明，答案信息可以在网上获取。它们的高性能体现了更强的信息检索（IR）召回能力，使其能够找到其他模型无法找到的特定数据点。

- **见解 2：拒答分化**。我们观察到大语言模型行为存在鲜明反差。顶尖模型几乎会尝试回答每个查询，而包括 DeepSeek（46%）、通义千问（28%）和 Gemini（24%）在内的一组模型则表现出较高的拒答率。这种分化表明，智能体处理"未知"问题的方式存在根本差异：激进型智能体会努力寻找一个合理答案，而其他模型在未能立即检索到高置信度证据时，则默认选择不作答。

- **洞见 3：检索诱发的拒答与安全对齐**。虽然拒答通常被归因于安全对齐，但我们的错误分析表明，搜索不足是一个主要原因。

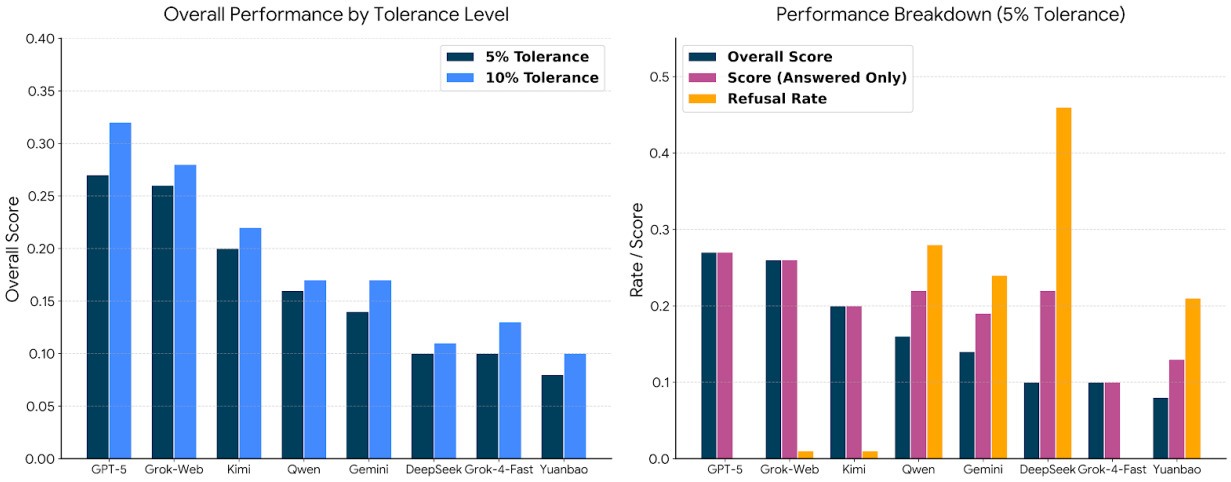


**图8** FutureX-NaturalDisaster 在 $24^{\text{th}}$ 年10月至 $28^{\text{th}}$年11月期间的结果。

– 由于真实数据公开存在（GPT-5 的成功已证明这一点），因此，声称“数据不可用”（通义千问/元宝中常见的说法）表明存在检索假阴性。– 因此，这些模型排名较低，部分原因在于它们无法有效检索深网政府数据库，而不只是因为安全策略过于保守。当搜索组件未能获取准确报告时，推理组件就只能拒答。

- **洞察4：拒答模式的分类体系。** 对拒绝响应进行定性分析，可区分出两种失败模式：

  – 基于能力的拒绝（搜索失败）：通义千问和元宝等模型经常以数据不可用为由（例如，“不存在权威数据”），这客观反映出搜索流程未能找到现有报告。– 基于策略的拒绝（不确定性规避）：DeepSeek 和 Gemini 等模型倾向于明确否定（例如，“无法预测”），这表明当检索到的上下文未达到内部置信度阈值时，某种机制会抑制输出。

## 6 FutureX-Search：从预测到检索

尽管 FutureX 的核心侧重于预测，但任何预测智能体的可靠性，根本上都取决于其准确检索历史数据的能力。**FutureX-Search** 作为一项对照实验，将“过去”的预测任务（其真实值如今已确定）转化为严格的信息检索基准测试。该领域评估智能体能否在答案已存在的情况下有效探索搜索空间。

### 6.1 数据集构建：时间回溯改造

我们将 FutureX [18] 档案中的高质量预测问题转换为搜索查询。构建流程分为三个阶段：校准、转换和混淆。

#### 6.1.1 阶段 1：区分性校准

为确保基准测试能够有效区分模型能力，我们根据 23 个基线模型在原始预测任务上的表现，采用了严格的过滤机制。

- **排除琐事题：** 由 $\geq 19$ 模型正确作答的问题将被移除（过于简单）。

- **排除噪声：** 由 $\leq 3$ 模型正确作答的问题将被移除（过于困难／有歧义）。

这会形成一个“恰到好处”的数据集：其中的信息可以检索到，但需要付出不小的搜索努力。

### 6.1.2 阶段 2：时间转移（StandardSearch）

将筛选后的问题从语法上由将来时转换为过去时，从而构建 FutureX-Search 数据集（$N = 100$）。

> 原始（FutureX）：“在2025年9月4日，中心平价汇率将会是多少……？”↓标准搜索（FutureX-Search）：“在2025年9月4日，中心平价汇率是多少……？”

### 6.1.3 阶段 3：复杂度注入（ComplexSearch）

为评估多跳推理和实体消歧，我们进一步处理一部分问题（$N = 39$），引入“描述性实体掩码”。这一过程包括：

1. 种子提取：在查询中识别一个枢纽实体（例如，“日元”）。2. Wiki-Mask 生成：基于维基百科知识生成描述性谜题（例如，“有一种货币……使用这种货币的国家……”）。3. 融合：将掩码与查询结合，创建 FutureX-Search。

这使我们能够构造类似于 BrowseComp 中的搜索查询（Level 3） [15], ，以评估大语言模型的搜索和推理能力。如图 9所示，这会形成更复杂的搜索任务，所有大语言模型的表现都大幅下降。完整的复杂度注入流程详见 Qiu 等人 [13]。

## 6.2 任务定义

FutureX-Search 设有三个不同的查询复杂度层级，要求智能体弥合隐式描述与显式数据检索之间的差距。

层级 *1*：直接检索。智能体会收到一个关于历史事件的清晰、明确的查询 .

> 查询：“2025年9月4日，人民币对100日元的中心平价汇率是多少？”

*Level 2*：显式实体消解（类型*1*）。 智能体必须先识别被掩盖的实体，然后检索数据。提示词明确询问“这种货币是什么？”

> 查询：“[货币描述]……这种货币是什么？ 2025年9月4日，人民币对100单位这种货币的汇率是多少？”

*Level 3*：隐式多跳检索（类型*2*）。 智能体必须在内部解析被掩盖的实体以执行搜索，而不会被明确提示去识别该实体。这测试了智能体处理模糊指代的能力。

> 查询：“[货币描述]……2025年9月4日，人民币对100单位这种货币的汇率是多少？”

## 6.3 评估指标

由于这些是有既定真实值的基于事实的检索任务，我们采用完全匹配 *(EM)* 和*F1* 分数[6] 来评估检索到的数值和实体的准确性。对于单个问题，我们

[6]因为有时模型会拒绝回答问题。

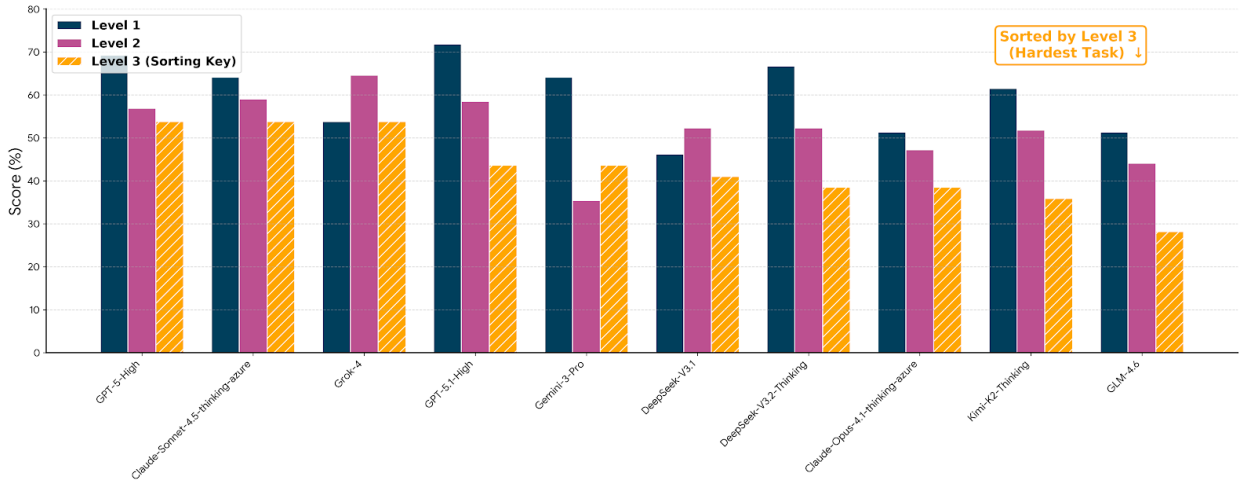


**图 9** FutureX-Search 上的结果。

使用 0/1 损失。对于 Type 1 问题（Level 2），第一问答对得 0.4 分（答错得 0 分），第二问答对得 0.6 分（答错得 0 分）。

## 6.4 结果

基于我们的内部智能体评估框架，我们使用搜索工具测试了不同的大语言模型，包括 GPT-5.1-High、GPT-5-High、Claude-Opus-4.1、Claude-Sonnet-4.5、Grok-4、Gemini-3-pro、DeepSeek-V3.1、DeepSeek-V3.2、Kimi-K2-thinking 和 GLM-4.6。结果如图 9所示。

我们发现：

- **见解 1：任务复杂度验证**。大多数模型的性能急剧下降，验证了 Level 3（隐式多跳）任务的难度。

  – 擅长直接检索（Level 1）的模型，在需要实体消歧时往往表现不佳。例如，GPT-5.1-High 在 Level 1 上取得了 71.8% 的领先成绩，但在 Level 3 上显著降至 43.6%。– 这 ∼28 个百分点的降幅证实，知道一个事实（检索）与推断该查找哪个事实（推理）并不相同。Level 3 任务成功地将“朴素搜索者”与“推理智能体”区分开来。

- **洞察 2：Grok-4 的“不变性”**。在 Grok-4 中观察到一个显著的异常现象。不同于呈下降趋势的其他模型（$L1 > L2 > L3$），Grok-4 保持完全稳定，在 Level 1 和 Level 3 中均获得 53.8% 的得分。

  – 尽管与 GPT-5.1 相比，其直接检索能力（Level 1）表现平平，但其处理复杂隐式查询（Level 3）的能力却处于顶尖水平。– 这表明，Grok-4 的架构优先考虑深层语义对齐，而非原始知识索引，使其能够独特地抵御查询混淆。

- **洞察 3：最先进水平的上限**。尽管当前旗舰模型具备强大的能力，Level 3 的性能仍在约 54% 处触及“玻璃天花板”（Grok-4、Claude-Sonnet-4.5 和 GPT-5-High 均处于这一水平）。

  – 没有任何模型能达到简单检索任务中所见的 70%+ 准确率。– 这表明，隐式实体消歧——即无需明确指令，仅凭模糊描述识别搜索目标的能力——仍是一项尚未解决的挑战，也是下一代搜索智能体面临的关键前沿。

## 7 相关工作

本节回顾了有关大语言模型智能体评测的研究，包括通用智能体能力、动态未来预测和特定领域垂直评测。我们重点介绍 FutureX-Pro 如何建立在以往工作 FutureX [18], 的方法论基础之上，同时应对高风险领域对深度专家级推理的迫切需求。

通用智能体基准测试。 第一波智能体基准测试着重评估开放领域环境中的广泛基础能力，例如工具使用、规划和网页导航。AgentBench [10] 和 GAIA [11] 等基准测试评估智能体作为通用助手的能力，即解决对人类而言概念上简单、但对模型而言需要复杂多步推理的任务。同样，WebArena [19] 和 Mind2Web [3] 在高保真模拟网页环境中评估智能体，测试其执行长时程交互任务的能力。尽管这些基准测试能有效衡量智能体的基础技能，但它们主要针对范围预先设定的静态问题进行评测，无法评估综合异构数据流以开展前瞻性分析的能力。正是这一缺口促使动态评估框架得以发展。

动态评测与未来预测。 为解决静态评测的局限——尤其是数据污染和缺乏时间推理——近期研究转向了动态基准与实时基准。FutureX [18] 提出了“设计上不可能受到污染”的架构，建立了一套流程，用于评估智能体预测可验证的未来事件的能力。FutureX 利用大量多样化的开放网络问题（涵盖体育、娱乐和一般政治），展示了智能体可以成为有效的通才型预测者。ForecastBench [9] 和 LiveBench [16],等其他研究也致力于解决如何在非静态数据上评估模型这一挑战。然而，这些基准测试（包括最初的 FutureX）大多重广度 而轻深度。它们评估智能体适应新信息的通用适应能力，却往往缺少测试专家级推理所需的针对性。例如，预测体育比赛结果（FutureX 中的常见任务）依赖于汇总公众情绪和统计数据，而预测流行病曲线则需要解读高度专业、特定领域的指标——通才型基准测试无法单独考察这种能力。

垂直领域评测：从广度到深度。从通用型助手向专业智能体的转变，需要以垂直领域专业知识为基础的基准测试。现有的特定领域基准，例如金融领域的 FinQA [2] 和 FinSearchComp [7] ，或医学领域的 MedQA [8], 侧重于静态知识检索，以及基于历史文档的推理。尽管这些基准测试能够有效检验领域知识，但无法评估智能体在现实场景中的预测效用。FutureX-Pro 弥合了 FutureX 的动态评测流程与垂直领域严格要求之间的差距。FutureX 汇总广泛可获取的公共信息，而 FutureX-Pro 则聚焦于高价值垂直领域——金融、零售、公共卫生和自然灾害。这些领域在来源权威性和数据解读方面提出了独特挑战。例如，在 FutureX-Pro 的金融领域中，智能体不仅必须阅读新闻，还必须专门解读企业披露，以预测市场走势；这与 FutureX 中的通用新闻综合形成对比。通过将重点从一般性广度转向高风险深度，FutureX-Pro 评估智能体能否超越通用网页搜索，在专业领域中发挥可靠的初级分析师作用。

## 8 结论与未来展望

本报告介绍了 **FutureX-Pro**，这是一个专门用于评估金融、零售、公共卫生和自然灾害等高价值垂直领域中智能体能力的框架。FutureX-Pro 不再局限于通用的开放领域任务或推测性预测基准，而是明确将评估重点转向那些决定经济与社会稳定的行业所提出的严格要求。

我们的研究凸显了“对话流畅度”与“领域能力”之间的根本区别：

- **信息源权威性与依据：** 不同于可能依赖群体智慧或广泛新闻聚合的通用基准测试，FutureX-Pro要求智能体将其推理建立在权威且异质的数据源之上——从财务披露到气象图表。我们在公共卫生和自然灾害领域的研究结果表明，性能瓶颈往往源于“搜索不足”——无法有效地从深网政府数据库中检索数据——而非简单的推理失败。

- **超越合理性的精确性：** 在高价值领域，若缺乏定量严谨性，仅仅给出看似合理的答案是不够的。金融领域的严格评估标准（要求相对误差低于 5%）揭示了一个显著差距：智能体虽然能够处理定性新闻，却难以将其转化为高风险环境所要求的精确数值预测。

- **不确定性的价值：** 我们对零售领域的分析挑战了“预测”等于“确定性”这一观念。结果表明，与被迫给出脆弱的确定性估计相比，智能体通过识别结果的高概率邻域，展现出更大的实用价值，即发挥“概率优势”。

若要推动智能体系统从通用助手发展为高价值领域中能力出众的合作伙伴，未来研究可着眼于三个关键前沿：

- **从表层检索到深度领域扎根：** 智能体的可靠性取决于其信息供应链。未来的工作必须超越通用网页搜索，纳入能够检索权威结构化数据源的“推理优先”策略。如 FutureX-Search 所示，在复杂的垂直领域中，隐式实体消歧以及推断要搜索什么的能力，是取得成功的先决条件。

- **概率推理标准化：** 认识到高价值领域本质上具有随机性，评估框架应不断完善，以奖励**校准而非虚假自信**。未来的智能体应能够生成概率预测（例如置信区间、前k项分布）。这一转变将优先考量智能体对风险和波动性的建模能力；与生成一个可能是幻觉数字的单一数值相比，这种能力更有价值。

- **区分能力与安全：** 为了提升在公共卫生等敏感行业中的效用，对齐策略必须区分避免伤害（基于策略的拒绝）和未能找到数据（基于能力的拒绝）。真正适用于高价值领域的智能体，必须既有足够的魄力去检索客观的安全数据，又足够谨慎以避免传播错误信息。

总而言之，FutureX-Pro 为下一代大语言模型提供了一项严格的压力测试。它表明，要精通高价值领域，光靠扩展参数规模还不够；还需要精准整合专业化搜索、定量推理和深度领域扎根。

## 9 项贡献

**项目负责人：** Jiashuo Liu、Siyuan Chen、Zaiyuan Wang、Zhiyuan Zeng

**核心贡献者：**（按领域划分）

- FutureX-Finance：Jiacheng Guo[†]、Liang Hu、Lingyue Yin、Suozhi Huang[†]、Wenxin Hao、Yang Yang、Zerui Cheng、Zixin Yao[†]
- FutureX-Retail：Lingyue Yin、Wenxin Hao
- FutureX-PublicHealth：Haoxin Liu[⋆]、Jiayi Cheng、Yuzhen Li、Zezhong Ma
- FutureX-NaturalDisaster：Haoxin Liu[⋆]
- FutureX-Search：Bingjie Wang、Bingsen Qiu、Xiao Liu、Zeyang Zhang、Zijian Liu
- 基础设施：Jinpeng Wang、Mingren Yin、Tianci He、Yali Liao、Yixiao Tian、Zhenwei Zhu

**贡献者：** Anqi Dai、Ge Zhang、Jingkai Liu、Kaiyuan Zhang、Wenlong Wu、Xiang Gao、Xinjie Chen、Zhixin Yao、ZhoufutuWen

**顾问：** B. Aditya Prakash[⋆]、Jose Blanchet[‡]、Mengdi Wang[†]、Nian Si[γ]、Wenhao Huan g

[†] 普林斯顿大学[⋆] 佐治亚理工学院

[‡] 斯坦福大学[γ] 香港科技大学

作者按名字的字母顺序排列。除非另有说明，所有作者均来自字节跳动 Seed 团队。

## 参考文献

[1] Tit Albreht. 新冠肺炎疫情引发的全球卫生挑战。《可持续发展》，15(9):7633，2023。[2] Zhiyu Chen、Wenhu Chen、Charese Smiley、Sameena Shah、Iana Borova、Dylan Langdon、Reema Moussa、Matt Beane、Ting-Hao Huang、Bryan R Routledge 等。FinQA：一个针对金融数据进行数值推理的数据集。载于*2021 年 第 3711 届自然语言处理经验方法会议*，2021 年，页码 3697– 。[3] Xiang Deng、Yu Gu、Boyuan Zheng、Shijie Chen、Sam Stevens、Boshi Wang、Huan Sun 和 Yu Su。Mind2Web：迈向通用型网络智能体。《神经信息处理系统进展》，36，28114，2023，28091：– 。[4] Haoyu Dong、Pengkun Zhang、Yan Gao、Xuanyu Dong、Yilin Cheng、Mingzhe Lu、Adina Yakefu 和 Shuxin Zheng。Finch：对以电子表格为中心的企业工作流中的金融与会计进行基准测试。*arXiv* 预印本 *arXiv:2512.13168*，2025。[5] Laurie Garrett。全球卫生面临的挑战。载于《全球卫生》，页码 525–548。Routledge，2017。[6] Jiacheng Guo、Suozhi Huang、Zixin Yao、Yifan Zhang、Yifu Lu、Jiashuo Liu、Zihao Li、Yanyan Deng、Qixin Xiao、Jia Tian 等。CryptoBench：用于对加密货币领域的大语言模型智能体进行专家级评估的动态基准测试。*arXiv* 预印本 *arXiv:2512.00417*，2025。[7] Liang Hu、Jianpeng Jiao、Jiashuo Liu、Yanle Ren、Zhoufutu Wen、Kaiyuan Zhang、Xuanliang Zhang、Xiang Gao、Tianci He、Fei Hu 等。FinSearchComp：迈向对金融搜索与推理进行真实、专家级评估。*arXiv* 预印本 *arXiv:2509.13160*，2025。[8] Di Jin、Eileen Pan、Nassim Oufattole、Wei-Hung Weng、Hanyi Fang 和 Peter Szolovits。这位患者患有什么疾病？一个来自医学考试的大规模开放领域问答数据集。*arXiv* 预印本 *arXiv:2009.13081*，2020。[9] Ezra Karger、Houtan Bastani、Chen Yueh-Han、Zachary Jacobs、Danny Halawi、Fred Zhang 和 Philip Tetlock。ForecastBench：人工智能预测能力的动态基准测试。载于第十三届国际学习表征会议。[10] Xiao Liu、Hao Yu、Hanchen Zhang、Yifan Xu、Xuanyu Lei、Hanyu Lai、Yu Gu、Hangliang Ding、Kaiwen Men、Kejuan Yang 等。AgentBench：评估作为智能体的大语言模型。载于国际学习表征会议，2024。[11] Grégoire Mialon、Clémentine Fourrier、Thomas Wolf、Yann LeCun 和 Thomas Scialom。GAIA：通用人工智能助手基准测试。载于第十二届国际学习表征会议，2023。[12] Rui Min、Zile Qiao、Ze Xu、Jiawen Zhai、Wenyu Gao、Xuanzhong Chen、Haozhen Sun、Zhen Zhang、Xinyu Wang、Hong Zhou 等。Ecombench：迈向对电子商务领域基础智能体的整体评估。*arXiv* 预印本 *arXiv:2512.08868*，2025。[13] Bingsen Qiu、Zijian Liu、Xiao Liu、Bingjie Wang、Feier Zhang、Yixuan Qin、Chunyan Li、Haoshen Yang 和 Zeren Gao。Bmgq：一种从半结构化数据自底向上生成复杂多跳推理问题的方法。*arXiv* 预印本 *arXiv:2510.24151*，2025。[14] Hannah Ritchie、Pablo Rosado 和 Max Roser。自然灾害。*Our World in Data*，2022。[15] Jason Wei、Zhiqing Sun、Spencer Papay、Scott McKinney、Jeffrey Han、Isa Fulford、Hyung Won Chung、Alex Tachard Passos、William Fedus 和 Amelia Glaese。BrowseComp：一个简单但颇具挑战性的浏览智能体基准测试。*arXiv* 预印本 *arXiv:2504.12516*，2025。[16] Colin White、Samuel Dooley、Manley Roberts、Arka Pal、Ben Feuer、Siddhartha Jain、Ravid Shwartz-Ziv、Neel Jain、Khalid Saifullah、Siddartha Naidu 等。LiveBench：一项颇具挑战性的无污染大语言模型基准测试。*arXiv* 预印本 *arXiv:2406.19314*，4，2024。[17] Manzhu Yu、Chaowei Yang 和 Yun Li。自然灾害管理中的大数据：综述。《地球科学》，8(5)：165，2018。[18] Zhiyuan Zeng、Jiashuo Liu、Siyuan Chen、Tianci He、Yali Liao、Yixiao Tian、Jinpeng Wang、Zaiyuan Wang、Yang Yang、Lingyue Yin 等。FutureX：用于大语言模型智能体未来预测的高级实时基准测试。*arXiv* 预印本 *arXiv:2508.11987*，2025。

[19] Shuyan Zhou、Frank F Xu、Hao Zhu、Xuhui Zhou、Robert Lo、Abishek Sridhar、Xianyi Cheng、Tianyue Ou、Yonatan Bisk、Daniel Fried 等。《Webarena：用于构建自主智能体的真实网络环境》。发表于第十二届国际学习表征会议，*ICLR 2024*，2024。[20] Fengbin Zhu、Xiang Yao Ng、Ziyang Liu、Chang Liu、Xianwei Zeng、Chao Wang、Tianhui Tan、Xuan Yao、Pengyang Shao、Min Xu 等。《Findeepresearch：在严格的金融分析中评估深度研究智能体》。*arXiv* 预印本 *arXiv:2510.13936*，2025。