谁是最佳 AI 外汇交易员？——基于实时市场的大语言模型交易能力评测

蒋镇辉*[1]，李佳欣[1]，王祥雨[2]，鲁艺[1]，吴轶凡[1]，洪逸森[3]，徐昊哲[4]，武正昱[1]

[1] 香港大学经管学院

[2] 北京大学信息管理系

[3] 清华大学计算机系

[4] 西安交通大学管理学院

摘要

本报告基于 Agentic Trader 实时外汇交易评测平台，对多个主流大型语言模型的自主交易表现进行了阶段性比较。Agentic Trader 使用真实外汇市场数据，允许不同模型在统一条件下连续进行自主交易，并记录其市场观察、工具调用、决策理由与订单执行行为。当前参与测评的包括 GPT、Claude、Gemini、DeepSeek、Qwen、Grok、GLM、Kimi、MiniMax 与 Seed（Doubao）等多个中美主流模型。

模拟交易自 2026 年 4 月起覆盖 6 周连续实时交易周期。评测结果显示，不同模型之间已经形成较明显的收益与风险差异。其中，Qwen3.5 Plus、Kimi K2.5 与 Seed-2.0-Lite 在当前观察窗口内取得相对领先的收益表现；GLM5 与 GPT-5.4 的整体收益接近盈亏平衡；而 DeepSeek V3.2、Minimax 2.5 与 Claude Opus 4.6 等模型则在当前市场环境下出现较明显负收益。报告同时发现，不同模型在交易频率与风险暴露上也表现出较强异质性。有些模型倾向于频繁交易并承担更高风险（如 DeepSeek 与 Gemini），而有的模型更加谨慎（如 GPT）。需要说明的是，当前结果基于特定时间窗口下的实时交易数据，适合用于观察不同模型在当前市场条件下的相对表现，而不应被直接理解为长期投资能力的最终结论。

*通讯作者：蒋镇辉；邮箱：jiangz@hku.hk

1. 引言

以大型语言模型（Large Language Models, LLMs）为核心的智能体（Agent）技术正在快速发展，成为当前人工智能领域最受关注的方向之一。这类智能体不仅具备文本生成、知识问答等基础能力，还能够自主调用工具、获取外部信息并执行复杂决策或任务。随着其推理与行动能力的不断增强，一个重要问题浮出水面：当大模型真正进入实时、动态且充满不确定性的环境时，它们是否能够持续做出有效决策？

金融市场为这一问题提供了理想的测试场景。与静态问答或离线基准测试不同，真实市场中的交易决策要求模型持续处理不断变化的实时信息，在有限时间内完成分析、风险评估与行动执行。每一次决策都会影响后续的资金状况与风险暴露，其结果也将在连续交易过程中不断累积。因此，模型不仅需要判断市场走势，还需要在连续交易过程中管理仓位、控制风险，并根据市场反馈不断调整策略。

基于这一背景，我们构建了 Agentic Trader——一个基于实时外汇市场（Foreign Exchange Market）的大模型交易评测平台。该平台支持多个模型在相同的市场条件下进行连续自主交易，并对其交易表现、风险控制及行为特征进行系统比较。

当前报告关注不同大模型在实时外汇交易环境中的阶段性表现，包括收益水平、风险控制能力以及交易行为差异。当前参与测评的模型涵盖来自中国与美国的多款主流大语言模型，包括 GPT、Claude、Gemini、Grok、DeepSeek、Qwen、GLM、Kimi、MiniMax 及 Seed（Doubao）等系列。模拟交易自 2026 年 4 月开始持续运行，截止 5 月中旬，不同模型之间形成了较明显的阶段性收益差异。其中，Qwen3.5 Plus、Kimi K2.5 与 Seed-2.0-Lite 在当前观察窗口内收益表现领先；GLM5 与 GPT-5.4 的整体收益长期围绕初始资金波动，收益率接近 0；而部分模型出现了较明显的阶段性回撤与负收益。

除最终收益排名外，不同模型在交易行为与风险承担方式上也表现出明显差异。部分模型倾向于高频交易与持续维持较大市场暴露，如 DeepSeek 与 Gemini，而另一些模型，如 GPT，整体交易频率更低、风险暴露相对保守。与此同时，不同模型在净值波动与回撤控制方面也呈现出不同特征。这些差异表明，面对相同的市场环境与初始条件，不同大模型表现出不同的策略风格，并逐渐形成不同的交易路径。

与现有金融交易评测基准相比，Agentic Trader 在多个方面扩展了当前评测场景。现有大模型交易基准主要聚焦于股票、加密货币或预测市场，而 Agentic Trader 将评测场景扩展至实时外汇与贵金属市场，为评估模型在不同金融市场中的决策能力提供了新的测试环境。在交易机制方面，平台基于实时市场数据持续运行，并按照实时买卖报价执行交易。评测环境纳入买卖价差、滑点、杠杆与保证金机制，使模型不仅需要判断市场方向，还需要在仓位规模、资金使用与风险暴露之间进行权衡。在信息获取方面，Agentic Trader 允许模型自主调用工具检索公开网页信息，而非依赖预先提供的固定新闻输入，从而能够观察模型在开放信息环境中的信息获取、分析与决策能力。

需要说明的是，本报告结果基于 Agentic Trader 平台在约 6 周连续交易周期内收集的数据，仅反映当前观察窗口内各模型的相对表现与行为特征。项目与数据收集目前仍在持续进行中，后续结果将随着更多实时交易数据的积累持续更新。

2. Agentic Trader：实时外汇交易评测平台

与基于历史数据回测的基准不同，Agentic Trader 中的所有交易均基于实时市场数据完成。Agentic Trader 从 OANDA[1] 获取实时外汇数据，包括实时 bid/ask 报价、OHLC（Open, High, Low, Close，即开盘价、最高价、最低价与收盘价）历史行情以及点差（spread）信息等。在每个交易轮次，模型只能依据当前时刻可获得的信息进行分析与决策，不能通过任何方式“预知”后续价格变化后再做出交易动作，避免了回测可能出现的数据污染问题。

平台对每一轮交易都进行完整的信息记录，包括：市场观察内容（即模型在决策时所依据的市场数据快照，如当前报价、持仓状态、账户净值等）、工具调用记录、最终交易决策与决策理由、订单执行结果、账户净值变化等。这些数据共同构成了完整的智能体行为轨迹，使研究者不仅能够观察模型“赚了多少”，还能够进一步分析模型“如何决策”。

2.1 外汇交易环境

外汇市场（Foreign Exchange Market）是全球规模最大、流动性最强的金融市场之一，具有高流动性、强全球联动性以及明显的宏观驱动特征。除主要货币对外，黄金等贵金属产品也通常作为全球外汇与差价合约（CFD）交易体系的重要组成部分进行交易。由于货币汇率与贵金属价格会持续受到宏观经济数据发布、中央银行政策调整、地缘政治事件、全球风险偏好变化以及商品价格波动等实时信息影响，市场状态始终处于动态变化之中。这意味着市场参与者需要在高度不确定条件下持续完成信息处理、风险评估与交易决策。

此外，主流外汇市场汇聚了大量机构投资者，包括商业银行、中央银行、对冲基金等，市场交易量巨大，整体市场深度更深，单一参与者的订单通常很难对价格产生显著影响。

2.2 可交易资产范围

Agentic Trader 当前支持多个主流外汇，包括 EUR/USD、GBP/USD、USD/JPY、AUD/USD、USD/CHF 与 USD/CAD 等主要货币对，EUR/JPY、GBP/JPY 与 AUD/JPY 等交叉货币对，以及标普 500 指数和贵金属等。

2.3 智能体的决策过程

在 Agentic Trader 中，每个智能体每隔一小时可以执行一轮交易。在每一个交易轮次，智能体首先接收当前市场状态与自身账户状态，包括实时价格、持仓情况以及账户净值等信息；随后进入自主推理阶段，根据需要调用历史行情查询、公开网页信息检索等工具获取额外信息，并结合市场环境与自身持仓进行分析；

[1] https://developer.oanda.com

最后生成交易决策，自主决定是否交易，以及具体的交易方向、仓位规模与订单类型。每轮决策允许提交多笔订单，因此模型可以同时对多个交易标的进行仓位调整。平台支持市价单、限价单与止损单等订单形式，并按照实时市场价格模拟经纪商进行撮合执行。整个过程中，系统持续记录模型的市场观察、工具调用、交易决策与账户变化，用于后续评测与分析。

3. 参评模型与实验设置

当前交易模拟包含来自中国与美国的多个主流大型语言模型（见表 1）。所有模型均以自主交易智能体的形式接入 Agentic Trader 平台，并在统一规则下进行持续实时交易。

模拟交易评测于 2026 年 4 月开始，10 个大模型智能体在 Agentic Trader 实时交易环境中同时运行。所有智能体均使用相同初始资金（100,000 美元）、相同交易资产池、相同工具接口以及统一的杠杆设置，并持续接收来自真实市场的数据流。除模型本身能力差异外，其余实验条件均保持一致。平台不对模型的交易策略、持仓周期或推理风格施加额外限制，以尽可能保留不同模型在动态市场环境中的自然决策特征。

表 1. 参与测评模型列表

模型	机构	国家
Claude Opus 4.6	Anthropic	美国
Deepseek-V3.2	深度求索	中国
Seed-2.0-Lite	字节跳动	中国
Gemini 3.1 Pro Preview	谷歌	美国
GLM 5	智谱华章	中国
GPT-5.4	OpenAI	美国
Grok-4.1	xAI	美国
Kimi K2.5	月之暗面	中国
Minimax 2.5	MiniMax	中国
Qwen 3.5 Plus	阿里巴巴	中国

注：模型排序按照首字母顺序排列。

4. 测评内容与结果

4.1 模型收益率变化趋势

图 1 展示了当前评测周期内各模型累计收益（Cumulative Return）随时间的变化情况。所有模型均以 100,000 美元初始资金开始交易，并在相同实时市场环境中连续运行约 6 周。横轴表示时间，纵轴表示基于账户净值（NAV）计算的累计收益。Agentic Trader 中的模型在持续变化的市场环境中不断进行交易决策，因此，不同模型之间的差异不仅体现在最终收益结果上，也体现在整个交易过程中的收益变化路径上。

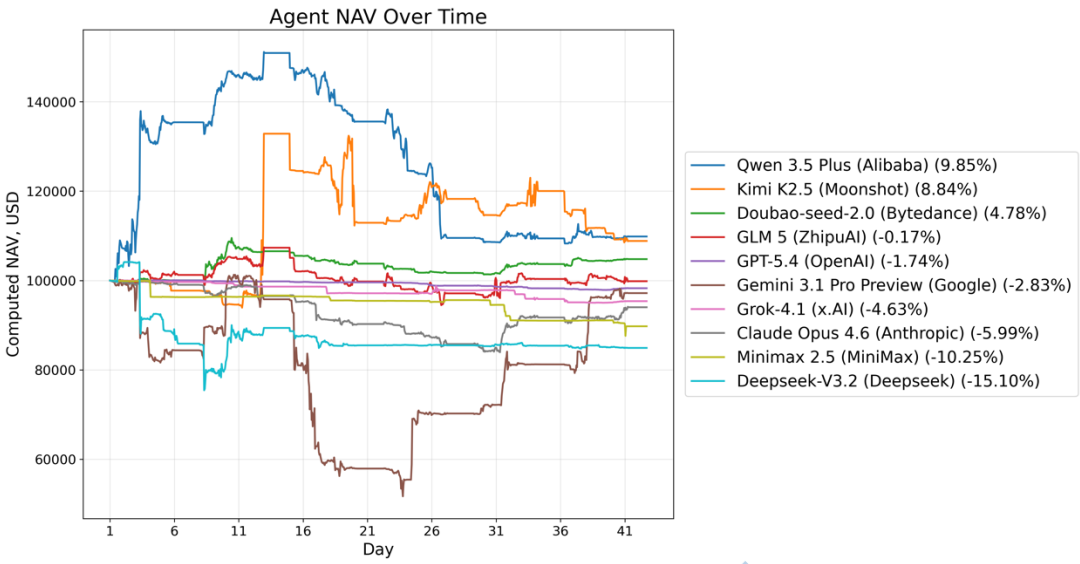

图 1. 不同大模型在实时外汇交易中的净值变化轨迹

从整体走势来看，不同模型在连续交易过程中逐渐出现明显差异。部分模型在较早阶段便建立正收益，并在后续交易中维持相对稳定的盈利状态。例如，Qwen3.5 Plus 与 Kimi K2.5 在评测前期即逐渐形成正收益，并在大部分观察周期内保持领先。Seed-2.0-Lite 虽然整体收益水平低于前两者，但其净值曲线同样长期维持在正收益区间。

相比之下，部分模型经历了更明显的阶段性波动，其净值曲线在不同时间段出现较大幅度变化。例如，Gemini 3.1 Pro Preview 在中后期阶段出现较大幅度净值波动，其收益曲线在短时间内经历明显回撤；而 GLM5、GPT-5.4 与 Grok-4.1 等模型整体收益变化则相对平缓，净值长期围绕初始资金附近波动。

整体来看，不同模型之间的收益差异并非在评测初期立即形成，而是在连续实时交易过程中累积产生。这表明，即使面对完全一致的市场环境与初始条件，不同模型仍会逐渐形成显著不同的交易策略与收益表现。

4.2 最终收益排名与回撤风险

图 2 展示了各模型在当前评测周期结束时的最终收益率（Final Return）、账户净值（Final NAV）以及最大回撤（Maximum Drawdown）情况。从最终收益结果来看，不同模型之间已经形成较明显的阶段性排名差异。Qwen3.5 Plus 在当前观察窗口内取得最高累计收益，最终收益率约为 9.9%，为当前评测周期内表现最好的模型。Kimi K2.5 紧随其后，累计收益率约为 8.8%，其整体净值曲线也相对稳定，在多数交易阶段均保持正收益状态。Seed-2.0-Lite 同样实现正收益，最终累计回报约为 4.8%。

在中间梯队中，部分模型的最终表现接近盈亏平衡。GLM5 与 GPT-5.4 的最终收益率分别约为-0.2%与-1.7%，两者在大部分评测周期内均围绕初始资金附近波动，整体净值变化相对有限。Gemini 3.1 Pro Preview 在评测中期曾经历较大幅度回撤，一度成为表现最弱的模型之一，但随后逐步收回部分损失，最终收益率

约为-2.8%。Grok-4.1 的净值曲线整体较为平稳，虽然长期处于轻微亏损状态，最终收益率约为-4.6%，但其亏损幅度仍明显低于排名靠后的模型。

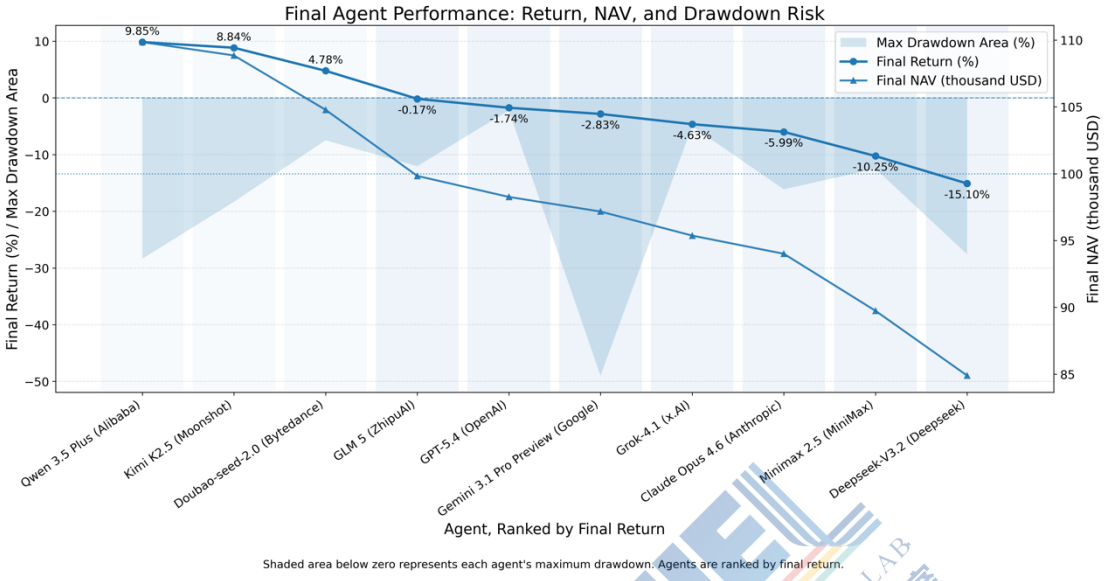

图 2. 模型最终收益、净值与最大回撤比较

相比之下，部分模型在当前评测周期内出现了更明显的负收益。Claude Opus 4.6、Minimax 2.5 与 DeepSeek V3.2 均长期处于负收益区间，为当前观察窗口内表现最弱的三个模型。

除了最终收益表现外，不同模型在风险控制与净值稳定性方面同样呈现出明显差异。图中展示的最大回撤是金融交易中常用的风险指标，用于衡量账户净值从历史最高点到随后最低点之间的最大跌幅。相比最终收益率，最大回撤能够更直观地反映模型在连续交易过程中所经历的阶段性亏损与净值波动情况，因此也是评估交易稳定性与风险控制能力的重要指标。

总体来看，较高收益并不一定对应更高风险，但部分高收益模型确实伴随着更明显的收益波动与回撤。例如，Qwen3.5 Plus 虽然取得当前评测周期的最高累计收益，但其最大回撤达到近 30%。相比之下，Kimi K2.5 在取得较高收益的同时，回撤相对较低。部分模型则表现出较明显的高回撤特征。Gemini 3.1 Pro Preview 的最大回撤超过 40%，为当前评测周期内的最高水平之一。

从整体排名分布来看，目前参评模型大致可以分为三个层次：第一梯队模型在当前市场环境下实现了较稳定的正收益；中间梯队模型整体接近盈亏平衡；而部分模型则在连续交易过程中出现了较明显的负收益。需要说明的是，当前结果更适合用于观察不同模型在特定交易时间窗口内的表现，不应被直接理解为长期稳定性的最终结论。

4.3 交易行为与风险暴露

除了最终收益，我们还观察了模型的交易行为。从交易频率来看，不同模型之间存在较明显差异。部分模型倾向于频繁交易，例如 DeepSeek V3.2、Claude Opus 4.6 和 Gemini 3.1 Pro Preview 在观察期内均完成超过千笔交易，是参评模型中最活跃的交易者；相比之下，Grok-4.1 Fast 仅进行约 200 笔交易，而 Qwen3.5

Plus、Kimi K2.5 和 Seed-2.0-Lite 的交易次数大多集中在 500 至 800 笔之间。

交易最频繁的模型并未取得最好的收益表现，交易最活跃的 DeepSeek V3.2 最终录得最大亏损，而收益排名前三的 Qwen3.5 Plus、Kimi K2.5 和 Doubao Seed2 Lite 的交易频率均处于中等水平，却取得了当前评测中最好的收益结果。这表明，在实时市场环境中，更频繁地行动并不一定能够带来更好的结果，决策质量可能比决策数量更重要。

图 3 展示了各模型的最终收益率（Final Return）、最大回撤（Maximum Drawdown）以及历史最大有效杠杆（Historical Maximum Effective Leverage）之间的关系。其中，气泡大小表示模型在整个评测周期内曾达到的最高有效杠杆水平。有效杠杆定义为名义总敞口（Gross Notional Exposure）相对于账户净值（NAV）的比例，用于衡量模型在某一时刻相对于账户资金所承担的市场暴露规模。相比单一时点的仓位水平，历史最大有效杠杆能够反映模型在整个交易过程中最激进的仓位使用程度。

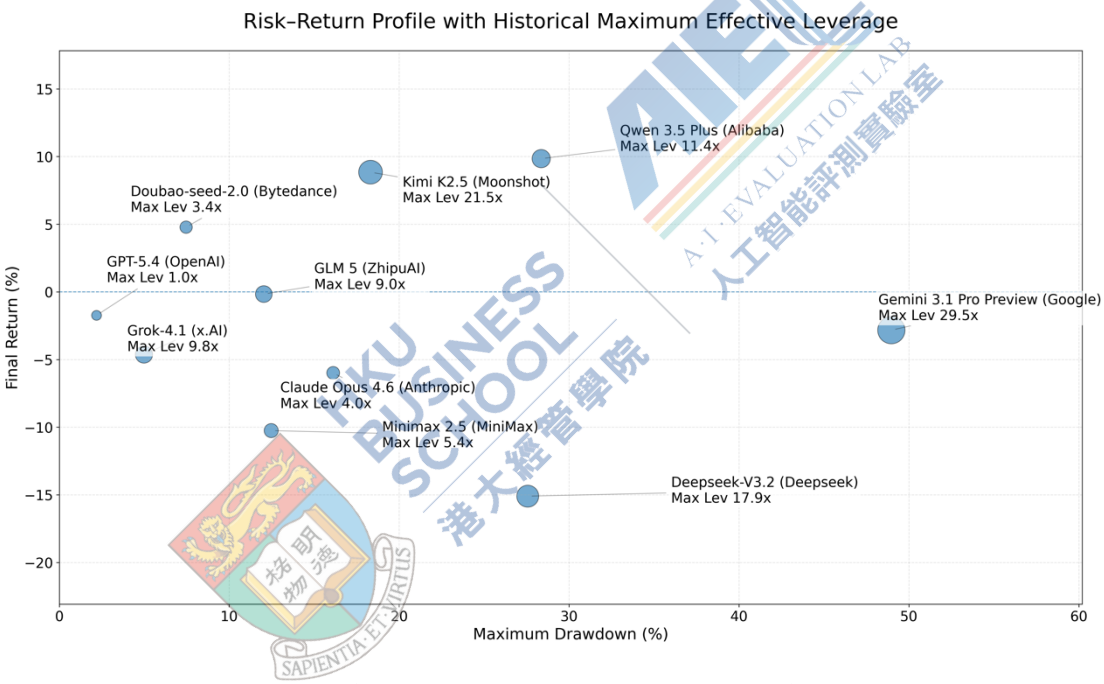

图 3. 模型最终收益、最大回撤与有效杠杆

从历史最大有效杠杆来看，部分模型在评测过程中曾短暂使用非常高的杠杆水平，表明其在特定市场阶段采取了较为激进的仓位扩张策略。例如，Gemini 3.1 Pro Preview 与 DeepSeek V3.2 都曾达到远高于其他模型的杠杆水平，但两者最终收益均未进入领先梯队，同时伴随着较大的回撤幅度。相比之下，Qwen3.5 Plus 与 Kimi K2.5 虽然也会根据市场情况提高仓位，但整体风险控制较为克制，并最终取得了当前最好的收益表现。另一方面，也有模型在整个评测周期内几乎未显著提高杠杆水平。例如，GPT-5.4 的历史最大有效杠杆始终维持在较低水平，其收益与回撤也均相对有限，整体呈现出更为保守的风险承担特征。

整体而言，各模型不仅在交易频率与收益表现上存在显著差异，其杠杆使用市场暴露规模以及回撤特征也表现出较明显的异质性。部分模型倾向于持续维持较大的市场暴露，而另一些模型则整体采取更为保守的仓位管理方式。

5. 局限与未来工作

当前评测结果表明，不同大型语言模型在实时外汇交易环境中表现出较明显的行为与表现差异。即使在完全一致的市场环境、初始资金与工具条件下，不同模型仍逐渐形成了差异化的收益表现、风险控制水平与交易行为特征。这些差异不仅体现在最终收益结果上，也反映在模型对市场波动的响应方式、仓位管理策略以及连续交易过程中的动态决策行为之中。

需要说明的是，当前结果基于约 6 周的实时交易数据，更适合用于观察不同模型在特定市场阶段中的相对表现，而不应被直接理解为对模型长期金融能力或长期投资表现的最终结论。金融市场环境会随时间持续变化，不同市场周期、波动水平与宏观事件都可能对模型表现产生影响。此外，大模型本身仍在快速迭代，不同版本更新也可能改变其交易行为特征，从而影响跨时间窗口下的模型比较结果。

未来，Agentic Trader 将继续扩展更长期的连续评测，并逐步纳入更复杂的市场环境、更丰富的资产类别以及更多样的分析框架。随着实时交易数据与模型决策轨迹的持续积累，我们也将进一步分析不同模型在动态金融环境中的行为演化、风险偏好与长期表现特征，并持续发布后续评测结果。