BenchSaaS

企业应用领域
大模型评测报告

100个领域测试任务/30个模型配置

报告版本 V1.0 | 2026 年 11 月

项目官网：mingdao.com/landing/benchsaas

目录

章节	页码
摘要与主要发现	3
1. 项目背景与评测范围	4
2. 任务来源、设计与任务集构成	6
3. 任务执行平台与评测机制	10
4. 业务领域、业务类型与评测维度排行	13
5. 评测维度细分表现与失败分析	26
6. 执行稳定性、执行效率与成本	39
7. 典型业务任务实例	43
8. 评测结论与适用范围	48
补充图表	50
版权说明与合作伙伴征集	51

摘要与主要发现

BenchSaaS 以企业应用中的具体工作检验模型能否理解目标、取得正确事实、遵守操作条件，并形成可使用的业务结果。本期使用 BenchSaaS 候选题库中的 100 道任务，对 10 个模型的低、中、高三档配置进行评测，每题重复 3 次，共纳入 9,000 条运行记录。

本期按业务领域、业务类型、评测维度三类排行呈现结果。按最终工作目标和必要数据来源划分，CRM（客户关系管理）20 题、库存 52 题、财务 27 题、跨应用 1 题。CRM 包含客户资料、销售订单管理与客户订单跟进。

- **整体完成质量** 合并 30 项配置的运行结果后，任务解决率为 90.25%。Gemini 3.8 Flash 中档以 97.94% 位列总体第一，高档以 97.24% 位列第二。
- **CRM** 6 项配置以 100% 并列第一，分别是 Kimi K3.1 Pro 高档、Qwen 3.8 Max 中档、GPT-6-Luna 中档和高档，以及 DeepSeek V4.1 Flash 低档和高档。
- **库存** GPT-6.1-Terra 低档任务解决率为 98.72%，在库存领域的 30 项配置中排名第一。
- **财务** DeepSeek V4 Pro 低档在票款、账户及收付款相关任务中以 100% 排名第一。
- **重复与投入** Gemini 3.8 Flash 中档有 90 道题在三次重复运行中均通过。各模型提高推理档位后，任务解决率的变化方向不同；作答耗时与模型费用分别在第 6 章呈现。

1. 项目背景与评测范围

1.1 项目背景

企业应用中的模型使用，以完成明确的业务工作为目标。查询客户资料、分析订单进度、核对票款关系或创建业务单据，都要求模型在具体的数据结构、记录状态和业务约束下取得事实、作出判断，并交付可使用的结果。模型的回答质量需要与业务事实和执行结果结合，才能判断任务是否完成。

这类工作对评测提出了三个要求：评测环境应当保留业务对象之间的关联，使模型能够基于上下文理解数据；评测任务应当说明工作目标与边界，使完成条件可以判断；评测结果应当同时反映必要信息是否正确、适用操作条件是否满足，以及写入后是否形成要求的业务结果。通用能力测试提供的知识、推理或生成能力信息，还需要与这些具体业务条件下的执行证据结合，才能支持应用选择。

BenchSaaS 围绕企业业务任务建立评测，将自然语言指令、HAP 业务应用、模型可用工具和结果检查连接起来，通过固定配置和重复运行，观察模型在同一批任务上的完成质量、表现差异与执行投入。

本期聚焦 CRM、库存、财务与跨应用四个业务领域，以七类业务类型组织工作形式，以七个一级维度呈现相关任务表现、重复稳定性与执行投入。其价值在于为业务适配判断与模型能力定位提供具有任务上下文、运行记录和结果依据的比较。

1.2 评测定位与目标

本报告以一次具体业务任务的运行结果为评价单位，同时观察配置、工具和执行环境共同作用下的表现。

评测目标	对应证据
任务完成质量	回答与最终业务记录的检查结果
不同场景下的表现差异	领域、类型和维度分组排行
重复执行的可靠程度	同题三次均通过的比例
执行投入	时长与费用各自可用的运行记录

1. 项目背景与评测范围

1.3 参评配置与样本

评测运行时间为 2026 年 10 月 22 日至 2026 年 11 月 19 日（北京时间）。2026 年 11 月 19 日从 HAP 回读运行状态。10 个模型各取低、中、高三个档位，每档执行同一组 100 道题，每题 3 次，共 30 项配置、9,000 个题次。

模型	低档	中档	高档
Claude Sonnet 5.5	low	medium	high
DeepSeek V4 Pro	low	high	max
DeepSeek V4.1 Flash	low	high	max
GLM-5.3	low	high	max
GPT-6-Luna	low	medium	high
GPT-6.1-Terra	low	medium	high
Gemini 3.8 Flash	low	medium	high
Kimi K3.1 Pro	low	high	max
MiniMax M3.1 Flash	low	high	max
Qwen 3.8 Max	low	medium	xhigh

表中列出实际请求的推理强度参数值。GPT 两款模型使用 reasoning.effort，其余使用 reasoning_effort。DeepSeek 两款模型及 GLM 同时设置 thinking.type=enabled，Qwen 设置 enable_thinking=true。低、中、高仅标识本期配置，不代表不同模型投入相同的推理资源。

1.4 结果阅读方法

任务解决率采用通过次数 ÷（通过次数＋失败次数），无效运行单列。汇总任务解决率按各配置的通过题次总数除以有效题次总数计算，不是各配置解决率的算术平均。

总体、业务领域、业务类型及评测维度排行，均按任务解决率从高到低排列。双栏排行先读左栏，再读右栏。名次依据未四舍五入的比例计算，相同比例并列。稳定性、作答耗时和模型费用分别按各图注明的口径统计。

排行反映这批实际任务的运行结果。不同配置可用的运行数、题目之间的相似性及测试环境都会影响比较。不同批次留存的环境标识与数据副本不同，现有材料尚不能证明全部初始状态等价；本报告据实际记录比较，细微的排名差异不作为稳定能力优势。

2. 任务来源、设计与任务集构成

2.1 任务来源与业务环境

本期任务来自 BenchSaaS 题库。题库基于同一 BenchSaaS 测试组织中的采购与销售、财务管理两个 HAP 应用编制，利用客户、商品、订单、库存、履约、对账、发票与收付款等业务对象及其关联，构造业务人员可能提出的工作需求。

题库索引记录的编制数据副本于 2026 年 10 月 9 日 11:23 采集，包含 40 张业务数据表、18,014 条记录；组织成员与部门依据测试组织资料确定。这份数据副本是题库编制依据，运行过程及成绩以实际留存记录为准。

任务来源包含业务目标、数据关系和操作条件三方面：以查询、统计、判断、追溯、核对、维护及流程受理组织工作目标；以已有记录及关联确定必要业务事实；以前置条件、时间状态、数量单位和允许修改数据的范围限定允许的结果。题库使用业务数据表中的记录、应用结构与组织信息，不涉及知识库、附件或扫描件。

2.2 任务设计原则与出题思路

业务领域说明工作所属的范围；业务类型说明要完成哪种工作；评测维度说明从哪些方面观察完成质量和执行投入。三者分别分类、分别统计。

任务设计以完整、可判断的业务结果为单位。题面说明目标、对象与必要边界，模型自行识别所需数据并选择数据处理路径。难度由条件组合、关联关系、分批履约和异常情形形成，每题按最终业务结果归入一个业务类型。

目标明确、路径开放。题面提供足以确定业务对象和结果边界的上下文，不直接指定数据表、数据项、操作接口名称或固定步骤；相似字段须有业务线索可供区分。

事实有据、口径确定。参考结果依据题库数据与业务关系编制。数量与计量单位共同构成事实，日期、状态、金额与取整口径以具体任务要求为准，缺失信息不允许推测补全。

以业务结果判断完成。查询与分析任务要求必要业务事实和结论正确；创建或修改业务记录的任务同时关注回答与目标单据及明细。需填写的必需值由题面、来源记录或明确配置确定，字段由模型填写或由应用正确带出均可接受。

任务的交付要求

查询、统计、判断与追溯需要交付可核实的事实和依据。维护与办理需要核对目标记录和明细。题面限定的前置条件不成立时，停止办理并说明原因也可以满足任务要求。

2. 任务来源、设计与任务集构成

七类业务类型的定义与本期实际覆盖例子

业务类型	通用含义	本期实际覆盖例子
查询业务记录	检索、定位并交付指定业务对象的必要信息。	客户资料、报价及指定业务记录查询。
统计业务数据	按确定的时间、状态、分组和计算口径形成业务统计。	订单金额、库存与账户等数据统计。
业务判断与决策支持	依据业务事实和约束比较方案、判断条件或提供行动依据。	库存缺口、安全调拨余量与报价判断。
业务追溯与进度还原	沿业务关联查明来源、处理过程、当前进度或历史状态。	订单履约、收退货批次及票款关系追溯。
审核、核对与对账	检查业务事实、记录或金额的一致性，识别并说明差异。	数量、时间、仓库约定、票款和盘点核对。
维护业务数据	在规定对象和影响范围内，更新或补充已有业务信息。	订单交付约定、报价、供应商资料及安全库存维护。
创建业务单据	按要求创建业务单据，填写必要信息及明细；条件不满足或已有记录时，说明原因并停止创建。	采购申请单、盘点单、收付款计划等的创建，以及按条件停止创建。

本期“创建业务单据”覆盖采购申请单、盘点单、收付款计划等的创建，以及条件不满足或已有记录时停止创建，共 19 题，其中 16 题要求创建记录，3 题要求停止创建。本期成绩反映上述任务的完成情况，完整审批、实际出入库和收付款确认等后续环节不在本类任务的评测范围内。

每题保存业务指令、出题说明、答案依据与标准答案，创建或修改业务记录的题目另有预期影响及保护范围。参考资料服务于评估，不进入被测模型的题面；评分不要求采用唯一的数据处理路径。实际评估机制及证据范围见第 3 章。

2. 任务来源、设计与任务集构成

2.3 本期任务集构成

候选库 500 题分为公开调试 100 题、正式隐藏 300 题、储备 100 题。本轮分别取 20、69、11 题。

七类业务类型：候选库与本轮样本

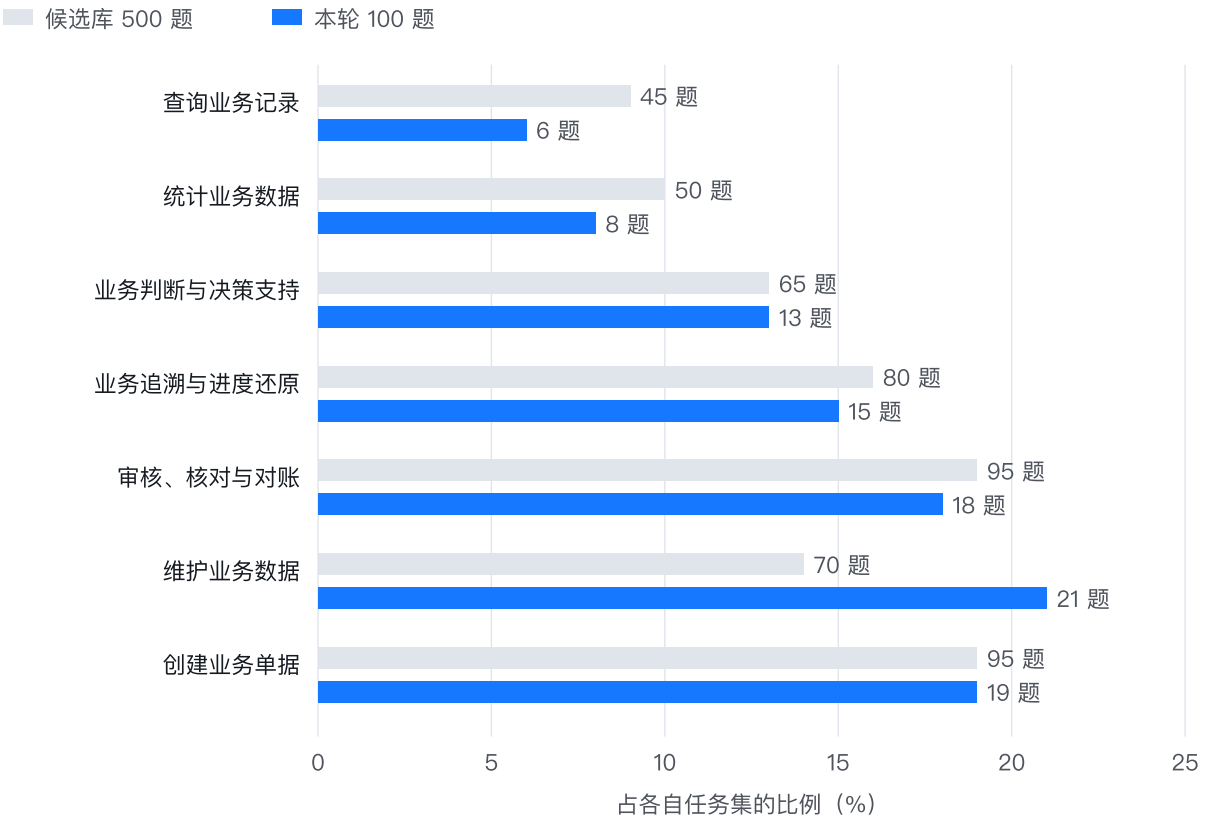


图 1　候选库与本轮七类业务类型分布

2. 任务来源、设计与任务集构成

2.4 样本特点与覆盖边界

每道任务按最终业务目标归入一个主领域，必要数据来源用于判断是否跨应用；辅助业务背景不重复计数。

业务领域	题数	通用含义	本期实际覆盖例子
CRM	20	客户资料、销售活动及客户相关事项的管理。	客户资料、销售订单分析、交付约定及客户订单进度。
库存	52	库存、仓储作业及相关采购业务。	采购、库存缺口、调拨盘点、出入库与退货批次。
财务	27	票据、往来款项、账户和收支相关业务。	对账票款、账户、收付款安排及回款跟进。
跨应用	1	必须取得或处理多个业务应用中的必要信息，才能完成目标的工作。	从财务付款与对账追至采购收货状态和实际时间。

销售订单进度跟进与仓储批次处理按工作目标区分；票款多表追溯仍可属于财务。只有必须交付两个应用中的业务事实，才计入跨应用任务。

本轮 100 题包含只读任务 60 题、涉及数据变更要求的任务 40 题，后者包括创建 16 题、更新 20 题，以及按要求不变更或拒绝办理 4 题。这里按执行要求分类：只读任务包含查询、统计、判断、追溯及核对，并非七类业务类型中的“查询业务记录”6 题；“维护业务数据”21 题也包含 1 道不变更任务，不能等同于实际更新的 20 题。基础、中等、复杂题分别为 8、36、56 题；普通正常业务、复杂正常链路、异常或条件不满足分别为 46、39、15 题。

索引候选库的复杂题为 175/500，即 35%；本轮为 56/100，即 56%。本轮更偏向复杂题，领域和业务类型题量也不均衡，排行应按这个实际样本构成理解。

3. 任务执行平台与评测机制

3.1 HAP：承载企业数据与流程的应用平台

HAP（Hyper Application Platform）是明道云的企业应用平台。用户可以通过可视化配置建立业务数据、数据视图、工作流和角色权限，将业务对象组织为可持续使用的应用。工作表即业务数据表，用于存放资料与单据，视图面向不同工作场景呈现数据，工作流支持数据处理和流转，统计图表与自定义页面用于结果呈现。

平台通过关联、公式与汇总组织数据关系，并提供标准开放接口，使外部系统或模型工具能够访问应用数据。这为企业任务执行提供了统一的数据与操作入口。

3.2 HAP：企业应用场景的最佳测试床

BenchSaaS 选择 HAP 这样的 APaaS（应用平台即服务）作为企业应用场景的最佳测试床，核心依据是通用性、同构性和完善的日志信息：既能承载真实业务，又能用统一机制执行任务、检查结果并诊断失败。

通用性：企业应用虽分属 CRM、采购、库存、财务等领域，却普遍由业务对象、关联数据、流程和权限构成。APaaS 能以通用能力配置这些场景，覆盖查询、统计、核对、维护和单据创建等任务，使评测从单一应用扩展到多种企业工作。

同构性：不同业务应用建立在同一套数据模型、关联机制和操作接口之上。模型面对的是不同业务语义下的一致操作体系，评测可复用工具与检查方法，减少接口差异对比较的干扰。同构不等于历史数据初始状态完全相同，本期仍保留相关比较边界。

日志信息：平台提供的数据变更、流程执行等日志，与评测程序保留的模型交互、工具调用及运行记录结合，可串联业务指令、操作过程和最终结果，帮助定位理解错误、调用失败或记录缺失，并分析耗时与费用。具体诊断以该次实际留存的日志为准。

HAP 模型评测中心管理任务、模型运行设置与评测结果；自动执行程序调度任务、保留处理过程并衔接结果检查。平台具备的能力不代表本期已全部测量，缺少完整验证证据的项目不单独计分。

3.3 任务如何执行

一次运行从固定题目与评测配置开始。模型收到业务指令后，通过数据操作工具读取应用中的数据结构、查找记录、查看明细和关联、进行统计；任务允许时，也可以新增或更新业务记录。这些工具是模型访问业务数据的操作入口。

模型完成回答后，评测核对必要业务事实，并按任务要求检查最终业务记录或是否调用了禁止的修改操作，形成通过、失败或无效结论。模型自行选择处理路径，评测不预设唯一的操作顺序。

对需要修改数据的任务，按适用流程保存处理结果，并安排恢复测试数据。恢复完成不单独证明任务范围外的数据在执行时一直保持不变；缺少完整验证证据的项目未单独计分。

3. 任务执行平台与评测机制

3.4 如何判断任务完成

任务情形	检查内容	完成要求
查询、统计、判断与追溯	核对回答与参考业务事实	必要对象、数值、状态、关系及依据正确，且没有明确矛盾。
需要创建或修改业务记录	核对回答并检查最终业务记录	目标记录、必需字段与相互关联符合任务要求。
要求不变更的任务	核对回答并检查是否调用禁止的修改操作	说明正确事实与停止原因，并满足任务规定的操作限制。

评测结合程序核对与独立评估模型的判断。程序核对适用的检查要求、操作限制和调用次数上限；评估模型按业务含义比较回答、参考答案及适用的最终结果。这里的“通过”指满足本期检查要求；语义判断仍可能存在误差，本期材料未提供人工复评一致性的量化结果。

评价允许等义表述，不要求模型使用固定的回答格式。缺少必要业务事实、对象或金额错误、状态或关联矛盾，都可能导致失败。主单创建成功不足以代替明细完整，文字中的“已完成”也不足以代替实际结果。

本报告按运行 ID 读取 HAP 中的最终状态。经复核后判为通过的记录中，35 条原判有误，122 条原判标准过严；这 157 条均计入通过数。

3. 任务执行平台与评测机制

3.5 重复执行、次数上限与结果分类

每题每配置重复 3 次。正式配置快照显示交互上限 31 轮、工具调用上限 30 次、运行时限 1,800 秒。

HAP 状态	纳入题次	统计处理
PASS	7,987	计入解决率的分子和分母
FAIL	863	仅计入解决率的分母
INVALID	150	计入无效题次，不计入解决率的分母

合计 9,000 个正式题次，其中 8,850 个取得通过或失败判定。

BenchSaaS 的任务执行平台分工

HAP 模型评测中心

任务与模型运行设置

各批次的过程与结果

评分及模型比较结果

自动执行与评测

模型交互与业务数据处理

检查结果并保留处理过程

恢复测试数据并保存结果

HAP 业务应用

采购与销售：32 张业务数据表

财务管理：8 张业务数据表

提供业务资料与单据

评测中心管理任务与结果；自动执行程序完成模型交互、数据处理和结果检查；业务应用提供测试数据。
表数为报告编制期的结构核查结果，不证明各历史运行的初始状态等价。

图 2　任务管理、自动执行与业务数据的分工

4. 业务领域、业务类型与评测维度排行

4.1 总体排行

名次	模型配置	任务解决率	通过 / 有效	0 — 100%
1	Gemini 3.8 Flash · 中	97.94%	285 / 291	
2	Gemini 3.8 Flash · 高	97.24%	282 / 290	
3	Kimi K3.1 Pro · 高	96.99%	290 / 299	
4	DeepSeek V4 Pro · 中	96.96%	287 / 296	
5	DeepSeek V4 Pro · 低	96.25%	282 / 293	
6	GLM-5.3 · 高	94.54%	277 / 293	
7	Gemini 3.8 Flash · 低	94.50%	275 / 291	
8	Qwen 3.8 Max · 中	93.58%	277 / 296	
9	Claude Sonnet 5.5 · 低	92.62%	276 / 298	
10	Qwen 3.8 Max · 高	92.59%	275 / 297	
10	DeepSeek V4 Pro · 高	92.59%	275 / 297	
12	GPT-6.1-Terra · 中	92.28%	275 / 298	
13	DeepSeek V4.1 Flash · 中	92.23%	273 / 296	
13	GPT-6-Luna · 高	92.23%	273 / 296	
15	Claude Sonnet 5.5 · 高	92.00%	276 / 300	
16	GPT-6.1-Terra · 高	91.95%	274 / 298	
17	Kimi K3.1 Pro · 中	91.25%	271 / 297	
18	GPT-6-Luna · 中	90.97%	272 / 299	
19	DeepSeek V4.1 Flash · 低	90.91%	270 / 297	
20	Qwen 3.8 Max · 低	90.67%	272 / 300	
21	Claude Sonnet 5.5 · 中	90.60%	270 / 298	
22	GLM-5.3 · 中	90.27%	269 / 298	
23	DeepSeek V4.1 Flash · 高	89.83%	265 / 295	
24	GPT-6.1-Terra · 低	88.29%	264 / 299	
25	Kimi K3.1 Pro · 低	88.22%	262 / 297	
26	GLM-5.3 · 低	81.82%	243 / 297	
27	MiniMax M3.1 Flash · 低	79.79%	229 / 287	
28	MiniMax M3.1 Flash · 中	77.42%	216 / 279	
29	MiniMax M3.1 Flash · 高	76.41%	217 / 284	
30	GPT-6-Luna · 低	73.13%	215 / 294	

图 3　总体任务解决率排行

在完整任务集上，Gemini 3.8 Flash 中档以 97.94% 排名第一。为比较相同样本，另取全部 30 项配置均取得三次有效判定的 38 道题。每项配置均有 114 个题次，其中 DeepSeek V4 Pro 低档通过 113 次，解决率为 99.12%。该结果仅用于检查样本差异对排名的影响。

4. 业务领域、业务类型与评测维度排行

4.2 业务领域排行

CRM 共 20 题，汇总通过率 95.31%。6 个配置并列第一，最高有效通过率 100.00%。

名次	模型配置	任务解决率	通过 / 有效	0 — 100%
1	Kimi K3.1 Pro · 高	100.00%	59 / 59	
1	Qwen 3.8 Max · 中	100.00%	58 / 58	
1	GPT-6-Luna · 高	100.00%	58 / 58	
1	GPT-6-Luna · 中	100.00%	59 / 59	
1	DeepSeek V4.1 Flash · 低	100.00%	60 / 60	
1	DeepSeek V4.1 Flash · 高	100.00%	58 / 58	
7	Gemini 3.8 Flash · 低	98.33%	59 / 60	
7	Claude Sonnet 5.5 · 高	98.33%	59 / 60	
7	GLM-5.3 · 中	98.33%	59 / 60	
10	DeepSeek V4 Pro · 中	98.31%	58 / 59	
10	DeepSeek V4 Pro · 低	98.31%	58 / 59	
10	GLM-5.3 · 低	98.31%	58 / 59	
13	Gemini 3.8 Flash · 中	96.67%	58 / 60	
13	Claude Sonnet 5.5 · 低	96.67%	58 / 60	
13	Claude Sonnet 5.5 · 中	96.67%	58 / 60	
13	Kimi K3.1 Pro · 低	96.67%	58 / 60	
17	GLM-5.3 · 高	96.61%	57 / 59	
17	Qwen 3.8 Max · 高	96.61%	57 / 59	
17	DeepSeek V4 Pro · 高	96.61%	57 / 59	
17	DeepSeek V4.1 Flash · 中	96.61%	57 / 59	
17	GPT-6.1-Terra · 高	96.61%	57 / 59	
22	GPT-6.1-Terra · 中	95.00%	57 / 60	
23	Qwen 3.8 Max · 低	93.33%	56 / 60	
24	Gemini 3.8 Flash · 高	92.98%	53 / 57	
25	Kimi K3.1 Pro · 中	91.53%	54 / 59	
26	MiniMax M3.1 Flash · 低	89.29%	50 / 56	
27	GPT-6-Luna · 低	88.14%	52 / 59	
28	GPT-6.1-Terra · 低	86.67%	52 / 60	
29	MiniMax M3.1 Flash · 中	85.71%	48 / 56	
30	MiniMax M3.1 Flash · 高	76.27%	45 / 59	

图 4　CRM：客户与销售订单管理

4. 业务领域、业务类型与评测维度排行

库存领域的模型排行

库存共 52 题，汇总通过率 90.02%。GPT-6.1-Terra · 低 居首，最高有效通过率 98.72%。

名次	模型配置	任务解决率	通过 / 有效	0 — 100%
1	GPT-6.1-Terra · 低	98.72%	154 / 156	
2	Gemini 3.8 Flash · 高	98.08%	153 / 156	
3	Gemini 3.8 Flash · 中	98.04%	150 / 153	
4	GPT-6.1-Terra · 中	97.42%	151 / 155	
5	GPT-6.1-Terra · 高	96.77%	150 / 155	
6	Kimi K3.1 Pro · 高	96.15%	150 / 156	
6	Claude Sonnet 5.5 · 中	96.15%	150 / 156	
8	Gemini 3.8 Flash · 低	95.97%	143 / 149	
9	DeepSeek V4 Pro · 中	95.45%	147 / 154	
10	GLM-5.3 · 高	94.77%	145 / 153	
11	DeepSeek V4 Pro · 低	93.55%	145 / 155	
12	Claude Sonnet 5.5 · 低	93.51%	144 / 154	
13	Claude Sonnet 5.5 · 高	92.95%	145 / 156	
14	Qwen 3.8 Max · 中	92.31%	144 / 156	
15	DeepSeek V4 Pro · 高	91.61%	142 / 155	
16	Kimi K3.1 Pro · 中	90.97%	141 / 155	
17	Kimi K3.1 Pro · 低	90.91%	140 / 154	
18	DeepSeek V4.1 Flash · 低	90.85%	139 / 153	
19	Qwen 3.8 Max · 低	90.38%	141 / 156	
20	Qwen 3.8 Max · 高	88.96%	137 / 154	
20	GPT-6-Luna · 高	88.96%	137 / 154	
22	DeepSeek V4.1 Flash · 中	88.31%	136 / 154	
23	GLM-5.3 · 中	86.45%	134 / 155	
24	DeepSeek V4.1 Flash · 高	86.27%	132 / 153	
25	GPT-6-Luna · 中	84.62%	132 / 156	
26	GLM-5.3 · 低	83.87%	130 / 155	
27	MiniMax M3.1 Flash · 低	78.38%	116 / 148	
28	MiniMax M3.1 Flash · 高	76.71%	112 / 146	
29	MiniMax M3.1 Flash · 中	73.61%	106 / 144	
30	GPT-6-Luna · 低	67.53%	104 / 154	

图 5　库存：采购、库存及仓储作业

4. 业务领域、业务类型与评测维度排行

财务领域的模型排行

财务共 27 题，汇总通过率 86.56%。DeepSeek V4 Pro · 低 居首，最高有效通过率 100.00%。

名次	模型配置	任务解决率	通过 / 有效	0 — 100%
1	DeepSeek V4 Pro · 低	100.00%	76 / 76	
2	DeepSeek V4 Pro · 中	98.75%	79 / 80	
3	Gemini 3.8 Flash · 中	98.67%	74 / 75	
4	Gemini 3.8 Flash · 高	98.65%	73 / 74	
5	Kimi K3.1 Pro · 高	96.30%	78 / 81	
5	Qwen 3.8 Max · 高	96.30%	78 / 81	
5	GPT-6-Luna · 中	96.30%	78 / 81	
8	DeepSeek V4.1 Flash · 中	96.25%	77 / 80	
9	GPT-6-Luna · 高	92.59%	75 / 81	
10	GLM-5.3 · 高	92.31%	72 / 78	
11	DeepSeek V4 Pro · 高	91.25%	73 / 80	
11	Kimi K3.1 Pro · 中	91.25%	73 / 80	
11	GLM-5.3 · 中	91.25%	73 / 80	
14	Qwen 3.8 Max · 中	91.14%	72 / 79	
15	Qwen 3.8 Max · 低	88.89%	72 / 81	
15	DeepSeek V4.1 Flash · 高	88.89%	72 / 81	
17	Gemini 3.8 Flash · 低	88.61%	70 / 79	
18	Claude Sonnet 5.5 · 低	87.65%	71 / 81	
19	Claude Sonnet 5.5 · 高	85.19%	69 / 81	
20	DeepSeek V4.1 Flash · 低	83.95%	68 / 81	
21	GPT-6.1-Terra · 中	80.00%	64 / 80	
22	GPT-6.1-Terra · 高	79.01%	64 / 81	
23	MiniMax M3.1 Flash · 中	77.63%	59 / 76	
24	Kimi K3.1 Pro · 低	76.54%	62 / 81	
25	MiniMax M3.1 Flash · 低	75.00%	60 / 80	
25	MiniMax M3.1 Flash · 高	75.00%	57 / 76	
27	Claude Sonnet 5.5 · 中	74.68%	59 / 79	
28	GPT-6-Luna · 低	71.79%	56 / 78	
29	GPT-6.1-Terra · 低	68.75%	55 / 80	
30	GLM-5.3 · 低	65.00%	52 / 80	

图 6　财务：票款、账户及收付款

4. 业务领域、业务类型与评测维度排行

4.3 业务类型排行

任务分组	通过 / 有效	无效	解决率
查询业务记录	444 / 523	17	84.89%
统计业务数据	605 / 682	38	88.71%
业务判断与决策支持	1066 / 1155	15	92.29%
业务追溯与进度还原	1270 / 1332	18	95.35%
审核、核对与对账	1522 / 1600	20	95.12%
维护业务数据	1782 / 1872	18	95.19%
创建业务单据	1298 / 1686	24	76.99%

七类业务按任务最终目标互斥归类。表中按各类任务合并 30 项配置的题次计算解决率；后续各图分别排列该类任务下的 30 项配置。

4. 业务领域、业务类型与评测维度排行

七类业务类型的模型排行

按任务解决率从高到低排列，先读左栏，再读右栏。每行依次为名次、模型档位和解决率。按未舍入的比例确定名次，相同比例并列。

查询业务记录

1	Gemini 3.8 Flash · 中	100.0%	16	Claude Sonnet 5.5 · 高	88.9%
1	Gemini 3.8 Flash · 高	100.0%	16	Qwen 3.8 Max · 低	88.9%
1	DeepSeek V4 Pro · 中	100.0%	18	Gemini 3.8 Flash · 低	83.3%
1	DeepSeek V4 Pro · 低	100.0%	18	DeepSeek V4.1 Flash · 低	83.3%
1	Qwen 3.8 Max · 高	100.0%	20	GLM-5.3 · 高	80.0%
1	GPT-6.1-Terra · 中	100.0%	21	Qwen 3.8 Max · 中	77.8%
1	GPT-6.1-Terra · 高	100.0%	21	DeepSeek V4 Pro · 高	77.8%
1	Kimi K3.1 Pro · 中	100.0%	21	DeepSeek V4.1 Flash · 高	77.8%
1	GPT-6-Luna · 中	100.0%	24	GLM-5.3 · 低	76.5%
1	GPT-6.1-Terra · 低	100.0%	25	GLM-5.3 · 中	70.6%
11	Kimi K3.1 Pro · 高	94.4%	26	Kimi K3.1 Pro · 低	61.1%
11	Claude Sonnet 5.5 · 低	94.4%	26	MiniMax M3.1 Flash · 低	61.1%
11	DeepSeek V4.1 Flash · 中	94.4%	26	GPT-6-Luna · 低	61.1%
11	GPT-6-Luna · 高	94.4%	29	MiniMax M3.1 Flash · 高	55.6%
11	Claude Sonnet 5.5 · 中	94.4%	30	MiniMax M3.1 Flash · 中	38.9%

统计业务数据

1	Gemini 3.8 Flash · 高	100.0%	16	DeepSeek V4 Pro · 高	87.5%
1	Kimi K3.1 Pro · 高	100.0%	16	GPT-6-Luna · 中	87.5%
1	DeepSeek V4 Pro · 中	100.0%	16	GLM-5.3 · 中	87.5%
1	Claude Sonnet 5.5 · 低	100.0%	16	DeepSeek V4.1 Flash · 高	87.5%
1	GPT-6.1-Terra · 中	100.0%	20	DeepSeek V4 Pro · 低	87.0%
1	Claude Sonnet 5.5 · 高	100.0%	21	Qwen 3.8 Max · 中	83.3%
1	GPT-6.1-Terra · 高	100.0%	21	Qwen 3.8 Max · 低	83.3%
1	Kimi K3.1 Pro · 中	100.0%	23	Qwen 3.8 Max · 高	82.6%
1	GPT-6.1-Terra · 低	100.0%	23	GLM-5.3 · 低	82.6%
1	Kimi K3.1 Pro · 低	100.0%	25	MiniMax M3.1 Flash · 高	81.0%
11	Claude Sonnet 5.5 · 中	95.8%	26	DeepSeek V4.1 Flash · 中	78.3%
12	Gemini 3.8 Flash · 中	95.0%	27	Gemini 3.8 Flash · 低	69.6%
13	GLM-5.3 · 高	91.7%	28	MiniMax M3.1 Flash · 低	66.7%
13	GPT-6-Luna · 高	91.7%	29	GPT-6-Luna · 低	52.2%
15	DeepSeek V4.1 Flash · 低	91.3%	30	MiniMax M3.1 Flash · 中	50.0%

图 7　七类业务类型的模型排行（1/4）

4. 业务领域、业务类型与评测维度排行

七类业务类型的模型排行（续）

按任务解决率从高到低排列，先读左栏，再读右栏。每行依次为名次、模型档位和解决率。按未舍入的比例确定名次，相同比例并列。

业务判断与决策支持

1	Gemini 3.8 Flash · 高	100.0%	10	Claude Sonnet 5.5 · 中	94.9%
1	Kimi K3.1 Pro · 高	100.0%	10	GLM-5.3 · 中	94.9%
1	GPT-6.1-Terra · 中	100.0%	18	DeepSeek V4.1 Flash · 高	94.7%
1	GPT-6.1-Terra · 高	100.0%	18	Kimi K3.1 Pro · 低	94.7%
1	GPT-6.1-Terra · 低	100.0%	20	Qwen 3.8 Max · 高	92.3%
6	DeepSeek V4 Pro · 中	97.4%	20	GPT-6-Luna · 高	92.3%
6	DeepSeek V4 Pro · 低	97.4%	20	GLM-5.3 · 低	92.3%
6	Qwen 3.8 Max · 中	97.4%	23	MiniMax M3.1 Flash · 低	91.4%
9	Gemini 3.8 Flash · 低	97.3%	24	Qwen 3.8 Max · 低	89.7%
10	Gemini 3.8 Flash · 中	94.9%	25	DeepSeek V4.1 Flash · 低	89.5%
10	GLM-5.3 · 高	94.9%	26	DeepSeek V4.1 Flash · 中	87.2%
10	Claude Sonnet 5.5 · 低	94.9%	26	Claude Sonnet 5.5 · 高	87.2%
10	DeepSeek V4 Pro · 高	94.9%	26	GPT-6-Luna · 低	87.2%
10	Kimi K3.1 Pro · 中	94.9%	29	MiniMax M3.1 Flash · 中	63.2%
10	GPT-6-Luna · 中	94.9%	30	MiniMax M3.1 Flash · 高	52.9%

业务追溯与进度还原

1	Gemini 3.8 Flash · 中	100.0%	15	GLM-5.3 · 中	97.7%
1	Gemini 3.8 Flash · 高	100.0%	15	Kimi K3.1 Pro · 低	97.7%
1	Kimi K3.1 Pro · 高	100.0%	18	Claude Sonnet 5.5 · 低	95.6%
1	GLM-5.3 · 高	100.0%	18	GPT-6.1-Terra · 中	95.6%
1	Gemini 3.8 Flash · 低	100.0%	18	DeepSeek V4.1 Flash · 中	95.6%
1	GPT-6-Luna · 高	100.0%	18	Claude Sonnet 5.5 · 高	95.6%
1	GPT-6.1-Terra · 高	100.0%	22	Qwen 3.8 Max · 高	95.5%
1	DeepSeek V4.1 Flash · 高	100.0%	23	GPT-6.1-Terra · 低	93.3%
9	DeepSeek V4 Pro · 中	97.8%	24	MiniMax M3.1 Flash · 中	91.1%
9	DeepSeek V4 Pro · 低	97.8%	25	MiniMax M3.1 Flash · 低	90.9%
9	Qwen 3.8 Max · 中	97.8%	26	Kimi K3.1 Pro · 中	88.9%
9	GPT-6-Luna · 中	97.8%	27	Claude Sonnet 5.5 · 中	86.7%
9	DeepSeek V4.1 Flash · 低	97.8%	27	GLM-5.3 · 低	86.7%
9	Qwen 3.8 Max · 低	97.8%	29	GPT-6-Luna · 低	83.7%
15	DeepSeek V4 Pro · 高	97.7%	30	MiniMax M3.1 Flash · 高	81.4%

图 7　七类业务类型的模型排行（2/4）

4. 业务领域、业务类型与评测维度排行

七类业务类型的模型排行（续）

按任务解决率从高到低排列，先读左栏，再读右栏。每行依次为名次、模型档位和解决率。按未舍入的比例确定名次，相同比例并列。

审核、核对与对账

1	Gemini 3.8 Flash · 中	100.0%	16	DeepSeek V4.1 Flash · 中	98.1%
1	Kimi K3.1 Pro · 高	100.0%	17	Claude Sonnet 5.5 · 低	96.3%
1	DeepSeek V4 Pro · 低	100.0%	17	GPT-6-Luna · 高	96.3%
1	Qwen 3.8 Max · 中	100.0%	17	GPT-6-Luna · 中	96.3%
1	GPT-6.1-Terra · 中	100.0%	17	Qwen 3.8 Max · 低	96.3%
1	Claude Sonnet 5.5 · 高	100.0%	21	GLM-5.3 · 高	96.2%
1	GPT-6.1-Terra · 高	100.0%	22	DeepSeek V4 Pro · 高	92.5%
1	Claude Sonnet 5.5 · 中	100.0%	22	DeepSeek V4.1 Flash · 低	92.5%
1	GPT-6.1-Terra · 低	100.0%	24	Kimi K3.1 Pro · 低	90.7%
10	Gemini 3.8 Flash · 高	98.1%	25	GLM-5.3 · 低	88.9%
10	DeepSeek V4 Pro · 中	98.1%	26	DeepSeek V4.1 Flash · 高	86.8%
10	Qwen 3.8 Max · 高	98.1%	26	MiniMax M3.1 Flash · 低	86.8%
10	GLM-5.3 · 中	98.1%	26	GPT-6-Luna · 低	86.8%
14	Gemini 3.8 Flash · 低	98.1%	29	MiniMax M3.1 Flash · 中	80.8%
14	Kimi K3.1 Pro · 中	98.1%	30	MiniMax M3.1 Flash · 高	78.4%

维护业务数据

1	DeepSeek V4 Pro · 低	100.0%	16	Qwen 3.8 Max · 高	95.2%
1	DeepSeek V4 Pro · 高	100.0%	16	GPT-6.1-Terra · 中	95.2%
3	Kimi K3.1 Pro · 高	98.4%	16	Qwen 3.8 Max · 低	95.2%
3	DeepSeek V4.1 Flash · 低	98.4%	16	GLM-5.3 · 中	95.2%
3	Claude Sonnet 5.5 · 中	98.4%	16	GLM-5.3 · 低	95.2%
3	Kimi K3.1 Pro · 低	98.4%	21	Qwen 3.8 Max · 中	95.1%
7	DeepSeek V4 Pro · 中	98.4%	22	MiniMax M3.1 Flash · 中	95.0%
8	GLM-5.3 · 高	98.4%	23	GPT-6-Luna · 高	93.5%
9	Gemini 3.8 Flash · 中	96.8%	24	DeepSeek V4.1 Flash · 高	92.1%
9	Gemini 3.8 Flash · 低	96.8%	25	Gemini 3.8 Flash · 高	91.9%
9	DeepSeek V4.1 Flash · 中	96.8%	25	GPT-6-Luna · 中	91.9%
9	Claude Sonnet 5.5 · 高	96.8%	25	MiniMax M3.1 Flash · 高	91.9%
13	Claude Sonnet 5.5 · 低	96.8%	28	MiniMax M3.1 Flash · 低	88.7%
13	Kimi K3.1 Pro · 中	96.8%	29	GPT-6.1-Terra · 低	87.3%
15	GPT-6.1-Terra · 高	96.7%	30	GPT-6-Luna · 低	84.1%

图 7　七类业务类型的模型排行（3/4）

4. 业务领域、业务类型与评测维度排行

七类业务类型的模型排行（续）

按任务解决率从高到低排列，先读左栏，再读右栏。每行依次为名次、模型档位和解决率。按未舍入的比例确定名次，相同比例并列。

创建业务单据

1	Gemini 3.8 Flash · 中	98.2%	16	Claude Sonnet 5.5 · 高	77.2%
2	Gemini 3.8 Flash · 高	96.4%	17	Claude Sonnet 5.5 · 低	76.8%
2	Gemini 3.8 Flash · 低	96.4%	18	GPT-6-Luna · 中	75.4%
4	DeepSeek V4 Pro · 中	90.9%	18	GLM-5.3 · 中	75.4%
5	GLM-5.3 · 高	89.3%	20	MiniMax M3.1 Flash · 高	72.7%
6	DeepSeek V4 Pro · 低	89.1%	21	Kimi K3.1 Pro · 中	71.4%
6	Qwen 3.8 Max · 中	89.1%	21	MiniMax M3.1 Flash · 中	71.4%
8	Kimi K3.1 Pro · 高	87.7%	23	Claude Sonnet 5.5 · 中	70.2%
9	DeepSeek V4.1 Flash · 中	87.5%	24	GPT-6.1-Terra · 中	67.9%
10	DeepSeek V4 Pro · 高	86.0%	25	Kimi K3.1 Pro · 低	66.7%
11	DeepSeek V4.1 Flash · 高	84.2%	26	GPT-6.1-Terra · 高	61.4%
12	Qwen 3.8 Max · 高	83.9%	27	MiniMax M3.1 Flash · 低	57.9%
13	GPT-6-Luna · 高	80.4%	28	GPT-6.1-Terra · 低	57.1%
14	DeepSeek V4.1 Flash · 低	78.9%	29	GLM-5.3 · 低	50.0%
14	Qwen 3.8 Max · 低	78.9%	30	GPT-6-Luna · 低	41.8%

图 7　七类业务类型的模型排行（4/4）

4. 业务领域、业务类型与评测维度排行

4.4 评测维度相关任务组排行

七个评测维度使用通用定义；本期只依据已经取得的运行证据呈现表现。通用定义包含的内容不等于本期已经全部测量。

评测维度	通用定义	本期统计依据
D1 业务理解	理解业务目标、对象、信息含义及相互关系。	29 道相关任务的整题通过情况。
D2 指令遵循	落实用户要求的条件、范围和交付内容。	26 道相关任务的整题通过情况。
D3 结果正确性与完整性	业务事实、计算结果和实际处理结果正确且完整。	全部 100 道任务的通过情况。
D4 业务约束遵循	遵守业务前提、允许的操作及影响范围。	13 道相关任务
D5 异常处理与事实可靠性	识别差异或缺失信息，依据事实处理、反馈或合理停止。	21 道相关任务；缺失信息与如实反馈仅 1 题。
D6 执行稳定性	相同任务重复执行时，保持一致的完成质量。	各配置中三次均有效的任务，统计三次均通过的比例。
D7 执行效率与成本	完成任务所需的时间、交互、操作与费用。	对有相应记录的有效运行，分别统计时长中位数和平均模型费用。

D1—D5 的题目组可重叠，使用整题完成结果。下表按各维度任务组合并 30 项配置的题次计算解决率。D6 衡量同一题目三次均通过的比例，D7 从有计时或费用的运行中分别统计。

任务分组	通过 / 有效	无效	解决率
D1 业务理解	2423 / 2571	39	94.24%
D2 指令遵循	2176 / 2313	27	94.08%
D3 结果正确性与完整性	7987 / 8850	150	90.25%
D4 业务约束遵循	995 / 1151	19	86.45%
D5 异常处理与事实可靠性	1787 / 1868	22	95.66%

4. 业务领域、业务类型与评测维度排行

D1–D5 相关任务组排行

按任务解决率从高到低排列，先读左栏，再读右栏。每行依次为名次、模型档位和解决率。按未舍入的比例确定名次，相同比例并列。

D1 业务理解

1	GPT-6.1-Terra · 高	100.0%	14	GPT-6-Luna · 中	95.4%
2	Gemini 3.8 Flash · 高	98.8%	14	Qwen 3.8 Max · 低	95.4%
2	Kimi K3.1 Pro · 高	98.8%	18	GLM-5.3 · 中	95.3%
2	GLM-5.3 · 高	98.8%	19	Kimi K3.1 Pro · 中	94.3%
2	Kimi K3.1 Pro · 低	98.8%	20	GPT-6-Luna · 高	94.1%
6	Gemini 3.8 Flash · 中	98.8%	21	DeepSeek V4.1 Flash · 高	94.0%
7	DeepSeek V4 Pro · 中	97.7%	22	DeepSeek V4.1 Flash · 中	93.1%
7	GPT-6.1-Terra · 中	97.7%	23	Qwen 3.8 Max · 高	93.0%
9	Claude Sonnet 5.5 · 低	97.7%	24	Claude Sonnet 5.5 · 中	92.0%
10	Claude Sonnet 5.5 · 高	96.6%	25	MiniMax M3.1 Flash · 低	91.8%
10	DeepSeek V4.1 Flash · 低	96.6%	26	Gemini 3.8 Flash · 低	91.6%
10	GPT-6.1-Terra · 低	96.6%	27	GLM-5.3 · 低	87.4%
13	DeepSeek V4 Pro · 高	96.5%	28	MiniMax M3.1 Flash · 中	82.3%
14	DeepSeek V4 Pro · 低	95.4%	29	GPT-6-Luna · 低	81.0%
14	Qwen 3.8 Max · 中	95.4%	30	MiniMax M3.1 Flash · 高	80.5%

D2 指令遵循

1	Gemini 3.8 Flash · 中	100.0%	14	DeepSeek V4.1 Flash · 中	94.9%
2	Kimi K3.1 Pro · 高	98.7%	14	DeepSeek V4.1 Flash · 低	94.9%
3	DeepSeek V4 Pro · 中	97.4%	18	GPT-6-Luna · 高	94.7%
3	Claude Sonnet 5.5 · 低	97.4%	19	GPT-6-Luna · 中	93.5%
3	Kimi K3.1 Pro · 中	97.4%	19	GLM-5.3 · 中	93.5%
6	Gemini 3.8 Flash · 高	97.4%	19	GLM-5.3 · 低	93.5%
6	Kimi K3.1 Pro · 低	97.4%	22	Qwen 3.8 Max · 低	92.3%
8	GPT-6.1-Terra · 高	97.4%	22	DeepSeek V4.1 Flash · 高	92.3%
9	Claude Sonnet 5.5 · 高	96.2%	24	Qwen 3.8 Max · 高	91.0%
9	Claude Sonnet 5.5 · 中	96.2%	24	GPT-6.1-Terra · 低	91.0%
11	GLM-5.3 · 高	96.1%	26	Qwen 3.8 Max · 中	90.9%
11	Gemini 3.8 Flash · 低	96.1%	27	MiniMax M3.1 Flash · 高	90.5%
11	DeepSeek V4 Pro · 高	96.1%	28	MiniMax M3.1 Flash · 中	90.0%
14	DeepSeek V4 Pro · 低	94.9%	29	MiniMax M3.1 Flash · 低	87.0%
14	GPT-6.1-Terra · 中	94.9%	30	GPT-6-Luna · 低	77.9%

图 8　D1–D5 相关任务组的模型排行（1/3）

4. 业务领域、业务类型与评测维度排行

D1–D5 相关任务组排行（续）

按任务解决率从高到低排列，先读左栏，再读右栏。每行依次为名次、模型档位和解决率。按未舍入的比例确定名次，相同比例并列。

D3 结果正确性与完整性

1	Gemini 3.8 Flash · 中	97.9%	16	GPT-6.1-Terra · 高	91.9%
2	Gemini 3.8 Flash · 高	97.2%	17	Kimi K3.1 Pro · 中	91.2%
3	Kimi K3.1 Pro · 高	97.0%	18	GPT-6-Luna · 中	91.0%
4	DeepSeek V4 Pro · 中	97.0%	19	DeepSeek V4.1 Flash · 低	90.9%
5	DeepSeek V4 Pro · 低	96.2%	20	Qwen 3.8 Max · 低	90.7%
6	GLM-5.3 · 高	94.5%	21	Claude Sonnet 5.5 · 中	90.6%
7	Gemini 3.8 Flash · 低	94.5%	22	GLM-5.3 · 中	90.3%
8	Qwen 3.8 Max · 中	93.6%	23	DeepSeek V4.1 Flash · 高	89.8%
9	Claude Sonnet 5.5 · 低	92.6%	24	GPT-6.1-Terra · 低	88.3%
10	Qwen 3.8 Max · 高	92.6%	25	Kimi K3.1 Pro · 低	88.2%
10	DeepSeek V4 Pro · 高	92.6%	26	GLM-5.3 · 低	81.8%
12	GPT-6.1-Terra · 中	92.3%	27	MiniMax M3.1 Flash · 低	79.8%
13	DeepSeek V4.1 Flash · 中	92.2%	28	MiniMax M3.1 Flash · 中	77.4%
13	GPT-6-Luna · 高	92.2%	29	MiniMax M3.1 Flash · 高	76.4%
15	Claude Sonnet 5.5 · 高	92.0%	30	GPT-6-Luna · 低	73.1%

D4 业务约束遵循

1	Claude Sonnet 5.5 · 中	100.0%	16	DeepSeek V4 Pro · 高	87.2%
2	Gemini 3.8 Flash · 高	97.4%	16	GPT-6-Luna · 高	87.2%
3	DeepSeek V4 Pro · 中	97.3%	18	Kimi K3.1 Pro · 中	86.8%
3	Gemini 3.8 Flash · 低	97.3%	18	DeepSeek V4.1 Flash · 低	86.8%
5	Gemini 3.8 Flash · 中	94.9%	20	GPT-6-Luna · 中	82.1%
5	Kimi K3.1 Pro · 高	94.9%	20	Qwen 3.8 Max · 低	82.1%
5	GPT-6.1-Terra · 低	94.9%	20	GLM-5.3 · 中	82.1%
8	DeepSeek V4.1 Flash · 中	94.7%	20	DeepSeek V4.1 Flash · 高	82.1%
9	GLM-5.3 · 高	92.3%	24	Qwen 3.8 Max · 高	81.6%
9	Qwen 3.8 Max · 中	92.3%	24	Kimi K3.1 Pro · 低	81.6%
11	DeepSeek V4 Pro · 低	92.1%	24	GLM-5.3 · 低	81.6%
11	Claude Sonnet 5.5 · 低	92.1%	27	MiniMax M3.1 Flash · 低	74.3%
13	Claude Sonnet 5.5 · 高	89.7%	28	GPT-6-Luna · 低	66.7%
13	GPT-6.1-Terra · 高	89.7%	29	MiniMax M3.1 Flash · 中	61.5%
15	GPT-6.1-Terra · 中	89.5%	30	MiniMax M3.1 Flash · 高	59.5%

图 8　D1–D5 相关任务组的模型排行（2/3）

4. 业务领域、业务类型与评测维度排行

D1–D5 相关任务组排行（续）

按任务解决率从高到低排列，先读左栏，再读右栏。每行依次为名次、模型档位和解决率。按未舍入的比例确定名次，相同比例并列。

D5 异常处理与事实可靠性

1	Gemini 3.8 Flash · 中	100.0%	16	DeepSeek V4.1 Flash · 中	98.4%
1	Kimi K3.1 Pro · 高	100.0%	17	Claude Sonnet 5.5 · 低	96.8%
1	DeepSeek V4 Pro · 低	100.0%	17	GPT–6–Luna · 高	96.8%
1	Qwen 3.8 Max · 中	100.0%	17	GPT–6–Luna · 中	96.8%
1	GPT–6.1–Terra · 中	100.0%	17	Qwen 3.8 Max · 低	96.8%
1	Claude Sonnet 5.5 · 高	100.0%	21	GLM–5.3 · 高	96.8%
1	GPT–6.1–Terra · 高	100.0%	22	DeepSeek V4 Pro · 高	93.5%
1	Claude Sonnet 5.5 · 中	100.0%	22	DeepSeek V4.1 Flash · 低	93.5%
1	GPT–6.1–Terra · 低	100.0%	24	Kimi K3.1 Pro · 低	92.1%
10	Gemini 3.8 Flash · 高	98.4%	25	GLM–5.3 · 低	90.5%
10	DeepSeek V4 Pro · 中	98.4%	26	DeepSeek V4.1 Flash · 高	88.7%
10	Qwen 3.8 Max · 高	98.4%	26	MiniMax M3.1 Flash · 低	88.7%
10	GLM–5.3 · 中	98.4%	28	GPT–6–Luna · 低	87.1%
14	Gemini 3.8 Flash · 低	98.4%	29	MiniMax M3.1 Flash · 高	81.7%
14	Kimi K3.1 Pro · 中	98.4%	30	MiniMax M3.1 Flash · 中	80.3%

图 8　D1–D5 相关任务组的模型排行（3/3）

5. 评测维度细分表现与失败分析

5.1 D1 业务理解

按任务解决率从高到低排列，先读左栏，再读右栏。每行依次为名次、模型档位和解决率。按未舍入的比例确定名次，相同比例并列。

D1.1 业务对象与信息含义

1	Gemini 3.8 Flash · 中	100.0%	15	Kimi K3.1 Pro · 高	96.3%
1	DeepSeek V4 Pro · 低	100.0%	15	DeepSeek V4 Pro · 中	96.3%
1	GLM-5.3 · 高	100.0%	15	Gemini 3.8 Flash · 低	96.3%
1	Qwen 3.8 Max · 中	100.0%	15	Qwen 3.8 Max · 高	96.3%
1	Claude Sonnet 5.5 · 低	100.0%	15	Claude Sonnet 5.5 · 高	96.3%
1	DeepSeek V4 Pro · 高	100.0%	15	GLM-5.3 · 中	96.3%
1	GPT-6.1-Terra · 中	100.0%	15	GLM-5.3 · 低	96.3%
1	GPT-6.1-Terra · 高	100.0%	15	MiniMax M3.1 Flash · 低	96.3%
1	Kimi K3.1 Pro · 中	100.0%	24	DeepSeek V4.1 Flash · 中	92.6%
1	DeepSeek V4.1 Flash · 低	100.0%	25	GPT-6-Luna · 高	88.9%
1	Qwen 3.8 Max · 低	100.0%	25	GPT-6-Luna · 中	88.9%
1	Claude Sonnet 5.5 · 中	100.0%	27	DeepSeek V4.1 Flash · 高	85.2%
1	GPT-6.1-Terra · 低	100.0%	28	MiniMax M3.1 Flash · 高	77.8%
1	Kimi K3.1 Pro · 低	100.0%	29	GPT-6-Luna · 低	70.4%
15	Gemini 3.8 Flash · 高	96.3%	30	MiniMax M3.1 Flash · 中	68.0%

D1.2 时间、状态与数值口径

1	Gemini 3.8 Flash · 高	100.0%	15	DeepSeek V4.1 Flash · 低	93.8%
1	Kimi K3.1 Pro · 高	100.0%	15	GLM-5.3 · 中	93.8%
1	GPT-6.1-Terra · 高	100.0%	15	GPT-6.1-Terra · 低	93.8%
4	DeepSeek V4 Pro · 中	97.9%	19	DeepSeek V4 Pro · 低	91.7%
4	GLM-5.3 · 高	97.9%	19	Qwen 3.8 Max · 中	91.7%
4	Claude Sonnet 5.5 · 低	97.9%	19	DeepSeek V4.1 Flash · 中	91.7%
4	GPT-6-Luna · 中	97.9%	19	Qwen 3.8 Max · 低	91.7%
4	Kimi K3.1 Pro · 低	97.9%	23	Qwen 3.8 Max · 高	89.4%
9	Gemini 3.8 Flash · 中	97.8%	24	MiniMax M3.1 Flash · 低	89.1%
9	DeepSeek V4.1 Flash · 高	97.8%	25	Gemini 3.8 Flash · 低	87.0%
11	GPT-6.1-Terra · 中	95.8%	26	MiniMax M3.1 Flash · 中	85.7%
11	Claude Sonnet 5.5 · 高	95.8%	27	Claude Sonnet 5.5 · 中	85.4%
13	DeepSeek V4 Pro · 高	95.7%	28	GPT-6-Luna · 低	84.8%
13	GPT-6-Luna · 高	95.7%	29	GLM-5.3 · 低	81.2%
15	Kimi K3.1 Pro · 中	93.8%	30	MiniMax M3.1 Flash · 高	79.5%

图 9　D1 相关细分任务组排行（整题结果）（1/2）

5. 评测维度细分表现与失败分析

5.1 D1 业务理解（续）

按任务解决率从高到低排列，先读左栏，再读右栏。每行依次为名次、模型档位和解决率。按未舍入的比例确定名次，相同比例并列。

D1.3 业务关联理解

1	Gemini 3.8 Flash · 中	100.0%	15	GLM-5.3 · 中	97.7%
1	Gemini 3.8 Flash · 高	100.0%	15	Kimi K3.1 Pro · 低	97.7%
1	Kimi K3.1 Pro · 高	100.0%	18	Claude Sonnet 5.5 · 低	95.6%
1	GLM-5.3 · 高	100.0%	18	GPT-6.1-Terra · 中	95.6%
1	Gemini 3.8 Flash · 低	100.0%	18	DeepSeek V4.1 Flash · 中	95.6%
1	GPT-6-Luna · 高	100.0%	18	Claude Sonnet 5.5 · 高	95.6%
1	GPT-6.1-Terra · 高	100.0%	22	Qwen 3.8 Max · 高	95.5%
1	DeepSeek V4.1 Flash · 高	100.0%	23	GPT-6.1-Terra · 低	93.3%
9	DeepSeek V4 Pro · 中	97.8%	24	MiniMax M3.1 Flash · 中	91.1%
9	DeepSeek V4 Pro · 低	97.8%	25	MiniMax M3.1 Flash · 低	90.9%
9	Qwen 3.8 Max · 中	97.8%	26	Kimi K3.1 Pro · 中	88.9%
9	GPT-6-Luna · 中	97.8%	27	Claude Sonnet 5.5 · 中	86.7%
9	DeepSeek V4.1 Flash · 低	97.8%	27	GLM-5.3 · 低	86.7%
9	Qwen 3.8 Max · 低	97.8%	29	GPT-6-Luna · 低	83.7%
15	DeepSeek V4 Pro · 高	97.7%	30	MiniMax M3.1 Flash · 高	81.4%

图 9　D1 相关细分任务组排行（整题结果）（2/2）

5. 评测维度细分表现与失败分析

5.2 D2 指令遵循

按任务解决率从高到低排列，先读左栏，再读右栏。每行依次为名次、模型档位和解决率。按未舍入的比例确定名次，相同比例并列。

D2.1 条件与范围遵循

1	Gemini 3.8 Flash · 中	100.0%	11	Kimi K3.1 Pro · 中	95.2%
1	DeepSeek V4 Pro · 低	100.0%	11	GLM-5.3 · 中	95.2%
3	Kimi K3.1 Pro · 高	97.6%	18	Gemini 3.8 Flash · 高	95.1%
3	DeepSeek V4 Pro · 中	97.6%	18	GLM-5.3 · 高	95.1%
3	DeepSeek V4.1 Flash · 中	97.6%	20	GPT-6.1-Terra · 高	95.0%
3	DeepSeek V4.1 Flash · 低	97.6%	21	Qwen 3.8 Max · 低	92.9%
3	Claude Sonnet 5.5 · 中	97.6%	22	Qwen 3.8 Max · 中	92.7%
3	Kimi K3.1 Pro · 低	97.6%	23	GPT-6.1-Terra · 低	90.5%
9	DeepSeek V4 Pro · 高	97.6%	24	GPT-6-Luna · 高	90.2%
9	GLM-5.3 · 低	97.6%	25	DeepSeek V4.1 Flash · 高	88.1%
11	Gemini 3.8 Flash · 低	95.2%	25	MiniMax M3.1 Flash · 高	88.1%
11	Claude Sonnet 5.5 · 低	95.2%	27	GPT-6-Luna · 中	87.8%
11	Qwen 3.8 Max · 高	95.2%	28	MiniMax M3.1 Flash · 中	86.5%
11	GPT-6.1-Terra · 中	95.2%	29	MiniMax M3.1 Flash · 低	81.0%
11	Claude Sonnet 5.5 · 高	95.2%	30	GPT-6-Luna · 低	73.8%

D2.2 必要信息交付

1	Gemini 3.8 Flash · 中	100.0%	14	Qwen 3.8 Max · 中	96.3%
1	Gemini 3.8 Flash · 高	100.0%	14	DeepSeek V4.1 Flash · 中	96.3%
1	Kimi K3.1 Pro · 高	100.0%	14	Claude Sonnet 5.5 · 高	96.3%
1	GLM-5.3 · 高	100.0%	14	DeepSeek V4.1 Flash · 低	96.3%
1	Gemini 3.8 Flash · 低	100.0%	14	MiniMax M3.1 Flash · 低	96.3%
1	Claude Sonnet 5.5 · 低	100.0%	14	MiniMax M3.1 Flash · 高	96.3%
1	DeepSeek V4 Pro · 高	100.0%	22	GLM-5.3 · 中	96.2%
1	GPT-6-Luna · 高	100.0%	22	Kimi K3.1 Pro · 低	96.2%
1	GPT-6.1-Terra · 高	100.0%	24	Qwen 3.8 Max · 高	92.6%
1	Kimi K3.1 Pro · 中	100.0%	24	GPT-6.1-Terra · 中	92.6%
1	GPT-6-Luna · 中	100.0%	24	Claude Sonnet 5.5 · 中	92.6%
1	Qwen 3.8 Max · 低	100.0%	24	GLM-5.3 · 低	92.6%
1	DeepSeek V4.1 Flash · 高	100.0%	24	MiniMax M3.1 Flash · 中	92.6%
14	DeepSeek V4 Pro · 中	96.3%	29	GPT-6.1-Terra · 低	88.9%
14	DeepSeek V4 Pro · 低	96.3%	30	GPT-6-Luna · 低	80.8%

图 10　D2 相关细分任务组排行（整题结果）（1/2）

5. 评测维度细分表现与失败分析

5.2 D2 指令遵循（续）

按任务解决率从高到低排列，先读左栏，再读右栏。每行依次为名次、模型档位和解决率。按未舍入的比例确定名次，相同比例并列。

D2.3 交付要求落实

1	Gemini 3.8 Flash · 中	100.0%	16	MiniMax M3.1 Flash · 中	93.3%
1	Gemini 3.8 Flash · 高	100.0%	17	GLM-5.3 · 高	90.5%
1	Kimi K3.1 Pro · 高	100.0%	17	Claude Sonnet 5.5 · 中	90.5%
1	DeepSeek V4 Pro · 中	100.0%	19	GLM-5.3 · 低	90.0%
1	Claude Sonnet 5.5 · 低	100.0%	19	MiniMax M3.1 Flash · 低	90.0%
1	GPT-6.1-Terra · 中	100.0%	21	Qwen 3.8 Max · 中	85.7%
1	GPT-6-Luna · 高	100.0%	21	DeepSeek V4 Pro · 高	85.7%
1	GPT-6.1-Terra · 高	100.0%	21	DeepSeek V4.1 Flash · 中	85.7%
1	Kimi K3.1 Pro · 中	100.0%	21	DeepSeek V4.1 Flash · 低	85.7%
1	GPT-6-Luna · 中	100.0%	21	GLM-5.3 · 中	85.7%
1	GPT-6.1-Terra · 低	100.0%	26	MiniMax M3.1 Flash · 高	82.4%
1	Kimi K3.1 Pro · 低	100.0%	27	DeepSeek V4 Pro · 低	81.0%
13	Claude Sonnet 5.5 · 高	95.2%	27	Qwen 3.8 Max · 低	81.0%
13	DeepSeek V4.1 Flash · 高	95.2%	29	GPT-6-Luna · 低	80.0%
15	Gemini 3.8 Flash · 低	95.0%	30	Qwen 3.8 Max · 高	76.2%

图 10　D2 相关细分任务组排行（整题结果）（2/2）

5. 评测维度细分表现与失败分析

5.3 D3 结果正确性与完整性

按任务解决率从高到低排列，先读左栏，再读右栏。每行依次为名次、模型档位和解决率。按未舍入的比例确定名次，相同比例并列。

D3.1 查询与计算结果

1	GPT-6.1-Terra · 高	100.0%	16	Qwen 3.8 Max · 高	94.4%
2	Kimi K3.1 Pro · 高	99.4%	16	Claude Sonnet 5.5 · 中	94.4%
3	Gemini 3.8 Flash · 高	99.4%	18	GLM-5.3 · 中	93.3%
4	GPT-6.1-Terra · 中	98.9%	19	Gemini 3.8 Flash · 低	93.1%
5	GPT-6.1-Terra · 低	98.3%	20	Qwen 3.8 Max · 低	92.8%
6	DeepSeek V4 Pro · 中	98.3%	21	DeepSeek V4 Pro · 高	92.1%
7	Gemini 3.8 Flash · 中	98.2%	22	DeepSeek V4.1 Flash · 中	92.1%
8	DeepSeek V4 Pro · 低	97.1%	22	DeepSeek V4.1 Flash · 低	92.1%
9	Claude Sonnet 5.5 · 低	96.1%	24	Kimi K3.1 Pro · 低	91.5%
10	GPT-6-Luna · 中	95.6%	25	DeepSeek V4.1 Flash · 高	90.9%
11	Kimi K3.1 Pro · 中	95.5%	26	GLM-5.3 · 低	87.1%
12	GPT-6-Luna · 高	95.5%	27	MiniMax M3.1 Flash · 低	83.9%
13	Claude Sonnet 5.5 · 高	95.0%	28	GPT-6-Luna · 低	79.0%
14	GLM-5.3 · 高	94.9%	29	MiniMax M3.1 Flash · 中	73.0%
15	Qwen 3.8 Max · 中	94.4%	30	MiniMax M3.1 Flash · 高	71.9%

D3.2 信息更新结果

1	DeepSeek V4 Pro · 低	100.0%	16	Qwen 3.8 Max · 高	95.2%
1	DeepSeek V4 Pro · 高	100.0%	16	GPT-6.1-Terra · 中	95.2%
3	Kimi K3.1 Pro · 高	98.4%	16	Qwen 3.8 Max · 低	95.2%
3	DeepSeek V4.1 Flash · 低	98.4%	16	GLM-5.3 · 中	95.2%
3	Claude Sonnet 5.5 · 中	98.4%	16	GLM-5.3 · 低	95.2%
3	Kimi K3.1 Pro · 低	98.4%	21	Qwen 3.8 Max · 中	95.1%
7	DeepSeek V4 Pro · 中	98.4%	22	MiniMax M3.1 Flash · 中	95.0%
8	GLM-5.3 · 高	98.4%	23	GPT-6-Luna · 高	93.5%
9	Gemini 3.8 Flash · 中	96.8%	24	DeepSeek V4.1 Flash · 高	92.1%
9	Gemini 3.8 Flash · 低	96.8%	25	Gemini 3.8 Flash · 高	91.9%
9	DeepSeek V4.1 Flash · 中	96.8%	25	GPT-6-Luna · 中	91.9%
9	Claude Sonnet 5.5 · 高	96.8%	25	MiniMax M3.1 Flash · 高	91.9%
13	Claude Sonnet 5.5 · 低	96.8%	28	MiniMax M3.1 Flash · 低	88.7%
13	Kimi K3.1 Pro · 中	96.8%	29	GPT-6.1-Terra · 低	87.3%
15	GPT-6.1-Terra · 高	96.7%	30	GPT-6-Luna · 低	84.1%

图 11　D3 相关细分任务组排行（整题结果）(1/2)

5. 评测维度细分表现与失败分析

5.3 D3 结果正确性与完整性（续）

按任务解决率从高到低排列，先读左栏，再读右栏。每行依次为名次、模型档位和解决率。按未舍入的比例确定名次，相同比例并列。

D3.3 单据创建结果完整性

1	Gemini 3.8 Flash · 中	97.9%	16	Claude Sonnet 5.5 · 高	72.9%
2	Gemini 3.8 Flash · 高	95.7%	17	Claude Sonnet 5.5 · 低	72.3%
2	Gemini 3.8 Flash · 低	95.7%	18	GPT-6-Luna · 中	70.8%
4	DeepSeek V4 Pro · 中	89.1%	18	GLM-5.3 · 中	70.8%
5	DeepSeek V4 Pro · 低	87.2%	20	MiniMax M3.1 Flash · 中	70.2%
5	GLM-5.3 · 高	87.2%	21	MiniMax M3.1 Flash · 高	67.4%
7	Qwen 3.8 Max · 中	87.0%	22	Kimi K3.1 Pro · 中	66.0%
8	Kimi K3.1 Pro · 高	85.4%	23	Claude Sonnet 5.5 · 中	64.6%
9	DeepSeek V4.1 Flash · 中	85.1%	24	GPT-6.1-Terra · 中	62.5%
10	DeepSeek V4 Pro · 高	83.3%	25	Kimi K3.1 Pro · 低	60.4%
11	DeepSeek V4.1 Flash · 高	81.2%	26	GPT-6.1-Terra · 高	54.2%
12	Qwen 3.8 Max · 高	80.9%	27	MiniMax M3.1 Flash · 低	50.0%
13	GPT-6-Luna · 高	76.6%	28	GPT-6.1-Terra · 低	48.9%
14	DeepSeek V4.1 Flash · 低	75.0%	29	GLM-5.3 · 低	40.4%
14	Qwen 3.8 Max · 低	75.0%	30	GPT-6-Luna · 低	32.6%

图 11　D3 相关细分任务组排行（整题结果）(2/2)

5. 评测维度细分表现与失败分析

5.4 D4 业务约束遵循

按任务解决率从高到低排列，先读左栏，再读右栏。每行依次为名次、模型档位和解决率。按未舍入的比例确定名次，相同比例并列。

D4.1 业务前置条件

1	Gemini 3.8 Flash · 高	100.0%	15	GPT-6.1-Terra · 高	86.7%
1	Gemini 3.8 Flash · 低	100.0%	17	GPT-6.1-Terra · 中	86.2%
1	Claude Sonnet 5.5 · 中	100.0%	17	DeepSeek V4.1 Flash · 低	86.2%
4	DeepSeek V4.1 Flash · 中	96.7%	19	GPT-6-Luna · 高	83.3%
5	DeepSeek V4 Pro · 中	96.4%	20	Kimi K3.1 Pro · 中	82.8%
6	Gemini 3.8 Flash · 中	93.3%	21	GPT-6-Luna · 中	76.7%
6	Kimi K3.1 Pro · 高	93.3%	21	Qwen 3.8 Max · 低	76.7%
6	GPT-6.1-Terra · 低	93.3%	21	GLM-5.3 · 中	76.7%
9	GLM-5.3 · 高	90.0%	24	Qwen 3.8 Max · 高	75.9%
9	Qwen 3.8 Max · 中	90.0%	24	Kimi K3.1 Pro · 低	75.9%
9	DeepSeek V4 Pro · 高	90.0%	24	GLM-5.3 · 低	75.9%
9	DeepSeek V4.1 Flash · 高	90.0%	27	MiniMax M3.1 Flash · 低	74.1%
13	DeepSeek V4 Pro · 低	89.7%	28	MiniMax M3.1 Flash · 高	60.7%
13	Claude Sonnet 5.5 · 低	89.7%	29	MiniMax M3.1 Flash · 中	60.0%
15	Claude Sonnet 5.5 · 高	86.7%	29	GPT-6-Luna · 低	60.0%

D4.3 重复办理与处理边界

1	Kimi K3.1 Pro · 高	100.0%	1	Kimi K3.1 Pro · 低	100.0%
1	GLM-5.3 · 高	100.0%	1	GLM-5.3 · 低	100.0%
1	Qwen 3.8 Max · 中	100.0%	18	Gemini 3.8 Flash · 高	93.3%
1	Claude Sonnet 5.5 · 低	100.0%	18	DeepSeek V4 Pro · 中	93.3%
1	Qwen 3.8 Max · 高	100.0%	18	DeepSeek V4 Pro · 低	93.3%
1	GPT-6.1-Terra · 中	100.0%	18	DeepSeek V4.1 Flash · 低	93.3%
1	GPT-6-Luna · 高	100.0%	18	GPT-6-Luna · 低	93.3%
1	Claude Sonnet 5.5 · 高	100.0%	23	DeepSeek V4.1 Flash · 中	92.9%
1	GPT-6.1-Terra · 高	100.0%	24	Gemini 3.8 Flash · 低	92.3%
1	Kimi K3.1 Pro · 中	100.0%	25	Gemini 3.8 Flash · 中	86.7%
1	GPT-6-Luna · 中	100.0%	26	MiniMax M3.1 Flash · 低	85.7%
1	Qwen 3.8 Max · 低	100.0%	27	DeepSeek V4 Pro · 高	80.0%
1	Claude Sonnet 5.5 · 中	100.0%	28	DeepSeek V4.1 Flash · 高	73.3%
1	GLM-5.3 · 中	100.0%	29	MiniMax M3.1 Flash · 中	66.7%
1	GPT-6.1-Terra · 低	100.0%	29	MiniMax M3.1 Flash · 高	66.7%

图 12　D4 相关细分任务组排行（整题结果）

5. 评测维度细分表现与失败分析

5.5 D5 异常处理与事实可靠性

按任务解决率从高到低排列，先读左栏，再读右栏。每行依次为名次、模型档位和解决率。按未舍入的比例确定名次，相同比例并列。

D5.1 差异与异常识别

1	Gemini 3.8 Flash · 中	100.0%	16	DeepSeek V4.1 Flash · 中	98.1%
1	Kimi K3.1 Pro · 高	100.0%	17	Claude Sonnet 5.5 · 低	96.3%
1	DeepSeek V4 Pro · 低	100.0%	17	GPT-6-Luna · 高	96.3%
1	Qwen 3.8 Max · 中	100.0%	17	GPT-6-Luna · 中	96.3%
1	GPT-6.1-Terra · 中	100.0%	17	Qwen 3.8 Max · 低	96.3%
1	Claude Sonnet 5.5 · 高	100.0%	21	GLM-5.3 · 高	96.2%
1	GPT-6.1-Terra · 高	100.0%	22	DeepSeek V4 Pro · 高	92.5%
1	Claude Sonnet 5.5 · 中	100.0%	22	DeepSeek V4.1 Flash · 低	92.5%
1	GPT-6.1-Terra · 低	100.0%	24	Kimi K3.1 Pro · 低	90.7%
10	Gemini 3.8 Flash · 高	98.1%	25	GLM-5.3 · 低	88.9%
10	DeepSeek V4 Pro · 中	98.1%	26	DeepSeek V4.1 Flash · 高	86.8%
10	Qwen 3.8 Max · 高	98.1%	26	MiniMax M3.1 Flash · 低	86.8%
10	GLM-5.3 · 中	98.1%	26	GPT-6-Luna · 低	86.8%
14	Gemini 3.8 Flash · 低	98.1%	29	MiniMax M3.1 Flash · 中	80.8%
14	Kimi K3.1 Pro · 中	98.1%	30	MiniMax M3.1 Flash · 高	78.4%

D5.2 缺失信息与如实反馈

1	Gemini 3.8 Flash · 中	100.0%	1	GPT-6.1-Terra · 高	100.0%
1	Gemini 3.8 Flash · 高	100.0%	1	Kimi K3.1 Pro · 中	100.0%
1	Kimi K3.1 Pro · 高	100.0%	1	GPT-6-Luna · 中	100.0%
1	DeepSeek V4 Pro · 中	100.0%	1	DeepSeek V4.1 Flash · 低	100.0%
1	DeepSeek V4 Pro · 低	100.0%	1	Qwen 3.8 Max · 低	100.0%
1	GLM-5.3 · 高	100.0%	1	Claude Sonnet 5.5 · 中	100.0%
1	Gemini 3.8 Flash · 低	100.0%	1	GLM-5.3 · 中	100.0%
1	Qwen 3.8 Max · 中	100.0%	1	DeepSeek V4.1 Flash · 高	100.0%
1	Claude Sonnet 5.5 · 低	100.0%	1	GPT-6.1-Terra · 低	100.0%
1	Qwen 3.8 Max · 高	100.0%	1	Kimi K3.1 Pro · 低	100.0%
1	DeepSeek V4 Pro · 高	100.0%	1	GLM-5.3 · 低	100.0%
1	GPT-6.1-Terra · 中	100.0%	1	MiniMax M3.1 Flash · 低	100.0%
1	DeepSeek V4.1 Flash · 中	100.0%	1	MiniMax M3.1 Flash · 中	100.0%
1	GPT-6-Luna · 高	100.0%	1	MiniMax M3.1 Flash · 高	100.0%
1	Claude Sonnet 5.5 · 高	100.0%	1	GPT-6-Luna · 低	100.0%

图 13　D5 相关细分任务组排行（整题结果）（1/2）

5. 评测维度细分表现与失败分析

5.5 D5 异常处理与事实可靠性（续）

按任务解决率从高到低排列，先读左栏，再读右栏。每行依次为名次、模型档位和解决率。按未舍入的比例确定名次，相同比例并列。

D5.3 合理停止与原因说明

1	Gemini 3.8 Flash · 中	100.0%	1	GPT-6.1-Terra · 高	100.0%
1	Gemini 3.8 Flash · 高	100.0%	1	Kimi K3.1 Pro · 中	100.0%
1	Kimi K3.1 Pro · 高	100.0%	1	GPT-6-Luna · 中	100.0%
1	DeepSeek V4 Pro · 中	100.0%	1	DeepSeek V4.1 Flash · 低	100.0%
1	DeepSeek V4 Pro · 低	100.0%	1	Qwen 3.8 Max · 低	100.0%
1	GLM-5.3 · 高	100.0%	1	Claude Sonnet 5.5 · 中	100.0%
1	Gemini 3.8 Flash · 低	100.0%	1	GLM-5.3 · 中	100.0%
1	Qwen 3.8 Max · 中	100.0%	1	DeepSeek V4.1 Flash · 高	100.0%
1	Claude Sonnet 5.5 · 低	100.0%	1	GPT-6.1-Terra · 低	100.0%
1	Qwen 3.8 Max · 高	100.0%	1	Kimi K3.1 Pro · 低	100.0%
1	DeepSeek V4 Pro · 高	100.0%	1	GLM-5.3 · 低	100.0%
1	GPT-6.1-Terra · 中	100.0%	1	MiniMax M3.1 Flash · 低	100.0%
1	DeepSeek V4.1 Flash · 中	100.0%	1	MiniMax M3.1 Flash · 高	100.0%
1	GPT-6-Luna · 高	100.0%	29	GPT-6-Luna · 低	88.9%
1	Claude Sonnet 5.5 · 高	100.0%	30	MiniMax M3.1 Flash · 中	77.8%

图 13　D5 相关细分任务组排行（整题结果）（2/2）

5. 评测维度细分表现与失败分析

细分任务组的结果如何解读

前述细分排行按各配置在相应任务组中的整题解决率计算。各组可以重叠；名次来自整题结果，不能据此判定具体能力缺陷或模型失败的内部原因。D3 一级组与总体榜相同，D1.3、D3.2、D5.1 分别与追溯、维护、核对业务类型使用同一任务组；这些对应结果不构成额外独立证据。

小样本的解释范围

缺失信息与如实反馈只含 1 题，合理停止与原因说明含 3 题。三个档位或多次执行增加了运行次数，但没有增加题目的种类。这些结果只能用于解释本期缺失信息、前提不满足或已有安排的情形。

从任务组定位到业务要求

D3 使用全部 100 题。查询与计算要求必要数值、单位和关系齐全；信息维护需核对原有内容与更新后的记录；新增单据则需核对主单和明细。一次整题失败可能涉及其中一个或多个环节，需结合回答检查、最终记录检查及任务要求核查。

5. 评测维度细分表现与失败分析

5.6 失败运行的主要记录类型

> HAP 最终判为失败的 863 个题次中，593 次至少有一项必需检查项未通过，占 68.71%。

记录类型	Runner 错误码	次数	占比
必需检查项未通过	MODEL_REQUIRED_ASSERTION_FAILED	593	68.71%
未满足业务条件或操作限制	MODEL_BUSINESS_RULE_REJECTED	161	18.66%
工具参数无效	MODEL_ARGUMENT_INVALID	51	5.91%
超出工具调用上限	MODEL_TOOL_LIMIT_EXCEEDED	45	5.21%
查询任务触发写入限制	MODEL_QUERY_WRITE_FORBIDDEN	5	0.58%
模型服务拒绝请求	MODEL_PROVIDER_REQUEST_REJECTED	5	0.58%
工具调用格式错误	MODEL_TOOL_CALL_MALFORMED	3	0.35%

图 14　863 次失败运行的主要记录类型

这 593 次包括回答未满足要求 428 次、最终记录未满足要求 165 次。分类以最终尝试的必需检查项结论为依据，每个失败题次只计入一类。

“未满足业务条件或操作限制”指工具请求因业务前置条件或任务允许的对象、操作、字段范围等限制而被拒绝。该分类描述评测运行中的限制，不能据此认定触发了 HAP 的“业务规则”功能。

分类用于说明未满足要求的环节，不能单独确定模型内部原因；被拦截的操作也不等于已造成业务变更。实际案例见第 7 章。

> 本版新增（5.6.1–5.6.2）：依托 ESAB 组织 HAP BenchSaaS 应用 863 次失败运行原始记录（failed_assertions_json、tool_call_log 与 model_turn_log），第 37–38 页逐类拆解全部 7 种失败记录类型在回答检查、数据库快照落地、工具范围拦截、参数校验及调用预算上的具体成因。

5. 评测维度细分表现与失败分析

5.6.1 必需检查项未通过（593 次）HAP 原始记录拆解

> “必需检查项”是每道任务收尾前按顺序执行的两道验收关卡：第一关 ANSWER_MATCH（100 题全部必检）核对最终回答的事实、计算口径、依据与变更回执；第二关 RESULT_CONSISTENCY（仅建单/修改任务触发）回读 HAP 冻结数据库快照核对落库结果。经对 HAP 原始记录逐条拆解，165 次因落库记录未通过的运行 100% 已先通过第一关回答检查——即文字回答“听起来已办妥”，回读系统库表却暴露字段留空、状态空置或言账不符。

一、第一关：回答未满足要求（ANSWER_MATCH · 428 次，占 72.18% | 只读 245 + 变更 183）

失败细分类型	原始记录具体表现与业务问题说明	次数 / 占比
只读 · 筛选、口径或数值计算出错 READ · Filter / Rule / Calc	查漏或多查目标单据（如应为空集却错列多笔、15 笔漏查为 5—9 笔）；违反“不同计量单位不得跨单位相加”等业务口径；或把“待确认”款项误算入“已确认”导致净流量与余款失准。	156 (36.4%)
变更 · 写库回执漏报必交代要素 WRITE · Receipt Omitted	写库后回复过于简略（如仅回“已登记/已追加”），漏报任务要求交代的核验项：未说明沿用原经办归属、未写出追加的新备注原文，或未确认其余明细未作修改。	97 (22.7%)
只读 · 自相矛盾、漏列依据或单位/编号错 READ · Contradiction / Unit / ID	明细表算对但开头总结/总计与下文自相矛盾（38 次）；漏报单据状态、极值、底层凭证明细编号或跟进建议（31 次）；写错计量单位或以系统 UUID 替代业务编号（20 次）。	89 (20.8%)
变更 · 操作未完成或选错范围 WRITE · Scope / Incomplete	批量维护漏改部分符合条件记录或选错对象（23 次）；答复自述新建单据状态仍为空、未设为“待确认/待入库”（20 次）；未登记指定申请人/部门、排序违规或空答复（43 次）。	86 (20.1%)

二、第二关：最终业务记录未满足要求（RESULT_CONSISTENCY · 165 次，占 27.82% | 建单 147 + 维护 18）

失败细分类型	原始记录具体表现与业务问题说明	次数 / 占比
必需字段或表间关联落库留空 NULL Required Field / Link	主单虽已建成，但 HAP 明细或表头必填项实存为空（常误以为关联商品会自动带出文本与库存）。高频缺口：可用库存（17 次，见案例 7.4）、明细商品编号/名称（13 次）、主单经办人/部门/业务员（11 次）、退货入库单位成本（9 次）及物理关联（7 次）。	58 (35.2%)
字段值、编码类型或修改范围错 Wrong Code / Value / Scope	盘点/调拨明细将库存编号 KC-* 误写入商品编号 SP-*（22 次）；备注漏句号、地址多加括注或全角数字（8 次）；库存数量写错或调入/调出仓写反（7 次）；沿用错部门或关联无效 ID（6 次）；批量改漏行或重复建单（4 次）。	47 (28.5%)
核心记录已写对但声明与快照矛盾 Claim vs. DB Mismatch	落库核心字段已达标，但回复附加陈述与冻结快照冲突或无法核实（“言账不符”）——如把含税小计当税额、把待入库数量提前从待收余量扣减、误报对账单原应付金额等。	33 (20.0%)
单据状态空置 [] 或误填实际时间 Empty Status [] / Timestamp	回执声称已设为“待确认 / 待入库 / 待出库”，但 API 写库漏传选项 Key 导致 HAP 状态字段实存为空数组 []（后端不补界面默认值）；另含误设为已确认或在待入库单提前填入实际入库时间。	27 (16.4%)

数据口径说明：以上拆解基于 HAP BenchSaaS 应用 V1.0 正式评测批次 593 条 MODEL_REQUIRED_ASSERTION_FAILED 原始运行记录逐条互斥归因（428 + 165 = 593；案例细节已作脱敏处理，未披露隐藏题编号与测试实体敏感信息）。

5. 评测维度细分表现与失败分析

5.6.2 其余六类失败运行（270 次）HAP 原始记录拆解

> 除必需检查项未通过（593 次）外，其余 270 次（占 863 次失败的 31.29%）主要由工具权限范围拦截、工具参数 Schema 校验失败、30 次调用上限耗尽及只读写入护栏等导致。以下依据 HAP 运行记录、工具调用日志（tool_call_log）及轮次日志（model_turn_log）给出每类的精确构成与实证原因。

失败类别与阶段	原始记录拆分构成与实证细节	次数 / 占比
未满足业务条件或操作限制 MODEL_BUSINESS_RULE_REJECTED 161 次（18.66%）· TOOL_EXECUTION	160/161 连续两次被拦后终止，均未造成真实写入（135 条留存细粒度 tool_call_log）： • 75 次（46.6%）create_record 越权提交只读/非允许字段（55 条日志 100% 为 FIELD_OUT_OF_SCOPE）； • 74 次（46.0%）查表结构（58）/ 列表（15）/ 详情（1）传错表标识（70 条日志中 66 次误传短别名如 spxx、ddmx 替代 24 位工作表 ID，4 次跨出许可表范围）； • 12 次（7.5%）get_app_info（7）/ get_app_worksheets_list（5）越界探查未授权应用元数据。	**161 (18.66%)**
工具参数无效 MODEL_ARGUMENT_INVALID 51 次（5.91%）· ARGUMENT_VALIDATION	51/51（100%）连续两次未通过前置 Schema 校验（33 条留存细粒度 tool_call_log）： • 41 次（80.4%）get_record_list 参数错误（31 条日志含分页游标 _runner_read.cursor 冲突 19 次、筛选组 filter.children 误传 JSON 对象而非数组 6 次、漏传 appId 等必填项 6 次，另 10 次未留明细日志）； • 9 次（17.6%）create_record 字段值 JSON 类型不符（如把关联 ID 错包为 {"item":...} 对象）； • 1 次（2.0%）get_record_relations 漏传必填参数 appId。	**51 (5.91%)**
超出工具调用上限 MODEL_TOOL_LIMIT_EXCEEDED 45 次（5.21%）· LOOP_CONTROL	43 次达单题 30 次上限（2 次在 21/29 次停），其中 39/45（86.7%）全程零工具报错（工具可用但未收敛）： • 19 次（42.2%）只读高频翻页或反复试探筛选（get_record_list ⩾ 20 次）； • 16 次（35.6%）只读多表探查与逐条拉取详情发散（14 次跨表扫描 + 2 次逐行详情）； • 8 次（17.8%）变更任务前置查数耗尽额度、尚未进入写库； • 2 次（4.4%）已执行写库后持续反复回查直至超限。	**45 (5.21%)**
查询任务触发写入限制 MODEL_QUERY_WRITE_FORBIDDEN 5 次（0.58%）· TOOL_SELECTION	100% 发生在只读决策支持任务，写入前拦截（executed = false，零数据变更）： • 3 次（60.0%）在只读采购建议任务中擅自调用 create_record 试图新建采购订单； • 2 次（40.0%）在只读客户对账跟进判断中擅自调用 update_record 等写操作。	**5 (0.58%)**
模型服务拒绝请求 MODEL_PROVIDER_REQUEST_REJECTED 5 次（0.58%）· 评分阶段拒收→补评	审计回溯澄清：候选模型均已完成作答（answer_available = true），原发于评分调用被服务拒收；经人工补评（22 次改判 PASS），这 5 次确认为实质未通过并保留原代码： • 3 次（60.0%）面对分段分页数据提前放弃统计排名； • 2 次（40.0%）空集筛选口径判错，误列 5—10 笔无关单据。	**5 (0.58%)**
工具调用格式错误 MODEL_TOOL_CALL_MALFORMED 3 次（0.35%）· MODEL_RESPONSE	3/3 发生在变更任务已完成 5—15 次正常工具调用后，紧接有效工具返回生成破损载荷： • 2 次（66.7%）多步建单（退货/盘点）中途输出畸形工具调用； • 1 次（33.3%）报价更新第 4 轮正常 update_record 后的第 5 轮输出格式非法载荷。	**3 (0.35%)**

数据口径说明：270 = 161 + 51 + 45 + 5 + 5 + 3，与 593 合计恰好等于 V1.0 报告 863 次 FAIL 运行记录；其中部分较早批次运行未完整落盘 tool_call_log 明细载荷（161 中留存 135 条、51 中留存 33 条），已在各行括号中标明实测样本覆盖度。

6. 执行稳定性、执行效率与成本

6.1 执行稳定性

模型 · 档位	三次均通过的题数	三次均有效的题数	稳定通过率	未纳入题数
Gemini 3.8 Flash · 中	90	94	95.74%	6
Kimi K3.1 Pro · 高	92	99	92.93%	1
DeepSeek V4 Pro · 中	90	97	92.78%	3
DeepSeek V4 Pro · 低	87	94	92.55%	6
Gemini 3.8 Flash · 高	85	93	91.40%	7
Gemini 3.8 Flash · 低	84	93	90.32%	7
GLM-5.3 · 高	85	95	89.47%	5
GPT-6.1-Terra · 高	86	98	87.76%	2
Qwen 3.8 Max · 中	85	97	87.63%	3
GPT-6.1-Terra · 中	85	98	86.73%	2
GPT-6-Luna · 高	82	96	85.42%	4
GPT-6.1-Terra · 低	83	99	83.84%	1
Qwen 3.8 Max · 高	81	97	83.51%	3
Kimi K3.1 Pro · 中	81	97	83.51%	3
Claude Sonnet 5.5 · 中	82	99	82.83%	1
Claude Sonnet 5.5 · 低	81	98	82.65%	2
GPT-6-Luna · 中	81	99	81.82%	1
DeepSeek V4 Pro · 高	79	97	81.44%	3
Qwen 3.8 Max · 低	81	100	81.00%	0
GLM-5.3 · 中	79	98	80.61%	2
DeepSeek V4.1 Flash · 中	77	96	80.21%	4
Claude Sonnet 5.5 · 高	80	100	80.00%	0
DeepSeek V4.1 Flash · 高	74	95	77.89%	5
DeepSeek V4.1 Flash · 低	75	97	77.32%	3
Kimi K3.1 Pro · 低	73	97	75.26%	3
GLM-5.3 · 低	62	97	63.92%	3
MiniMax M3.1 Flash · 低	57	92	61.96%	8
GPT-6-Luna · 低	54	94	57.45%	6
MiniMax M3.1 Flash · 中	48	88	54.55%	12
MiniMax M3.1 Flash · 高	45	87	51.72%	13

图 15　D6 执行稳定性：同题三次均通过比例

各配置分别以三次均有效的题目为分母；其任务集合可能不同。

6. 执行稳定性、执行效率与成本

6.2 完成质量与投入一起观察

每行对应一个模型。每个指标由上至下为低、中、高档，依次以蓝、绿、橙三色标出。解决率从 70% 起绘；时长与费用从零起绘。下页列出每个点对应的数值。

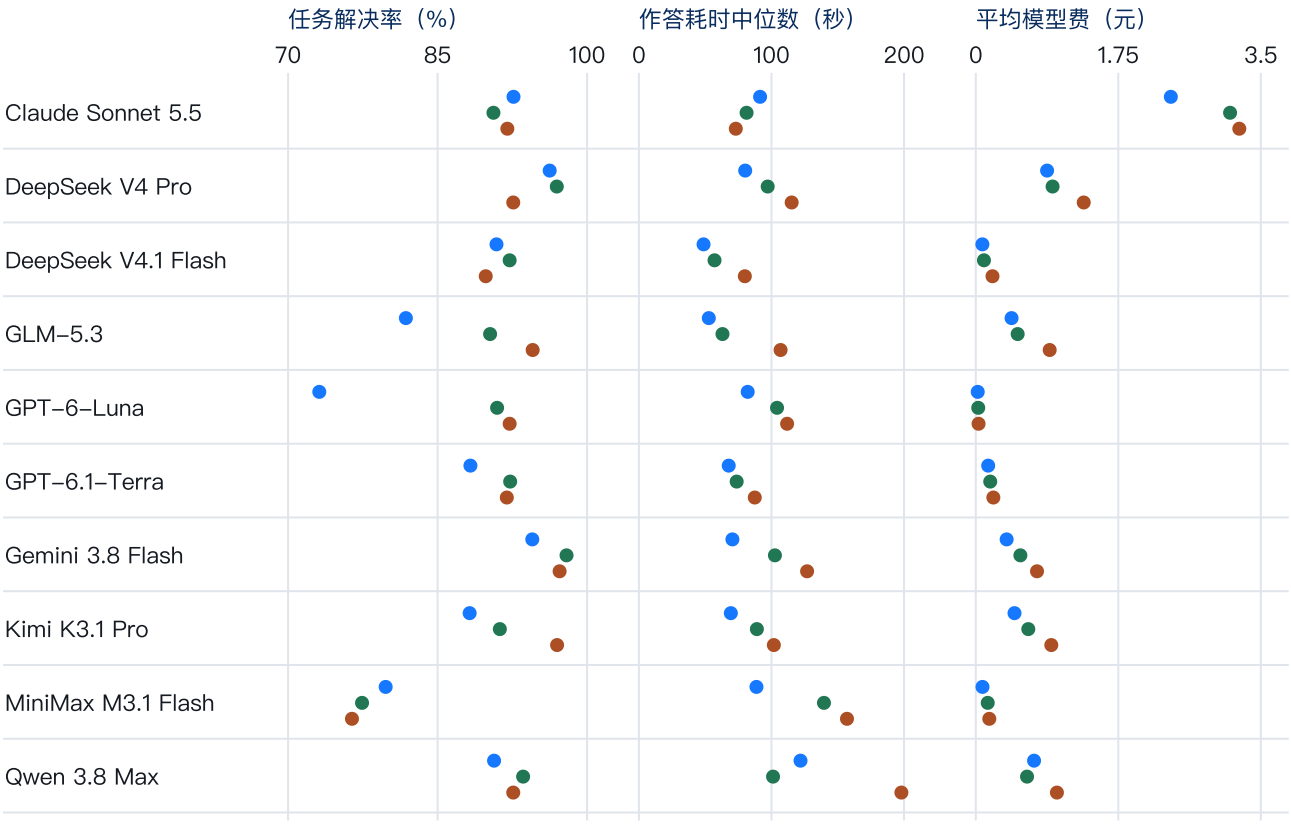


图 16　30 项配置的完成质量、模型费用与时长对照

三个指标分别计量。选择配置时，先对齐工作所需的完成质量，再查看相应的时间与费用。

6. 执行稳定性、执行效率与成本

完成质量与执行投入 · 数值对照

每行对应一个模型，每列对应其推理档位。格内从左到右为任务解决率（%）、作答耗时中位数（秒）、平均模型费用（元）；同一行可直接比较三个档位的完成质量与投入。作答耗时涵盖被测模型生成答案及期间的工具操作，起止定义见第 6.3 节。

模型	低档解决率 / 秒 / 元	中档解决率 / 秒 / 元	高档解决率 / 秒 / 元
Claude Sonnet 5.5	92.62 / 91.42 / 2.3949	90.60 / 81.22 / 3.1224	92.00 / 73.09 / 3.2358
DeepSeek V4 Pro	96.25 / 80.18 / 0.8756	96.96 / 97.14 / 0.9430	92.59 / 115.23 / 1.3254
DeepSeek V4.1 Flash	90.91 / 48.75 / 0.0809	92.23 / 57.03 / 0.1003	89.83 / 79.91 / 0.2052
GLM-5.3	81.82 / 52.73 / 0.4393	90.27 / 63.04 / 0.5141	94.54 / 106.88 / 0.9074
GPT-6-Luna	73.13 / 82.06 / 0.0241	90.97 / 104.19 / 0.0302	92.23 / 111.78 / 0.0341
GPT-6.1-Terra	88.29 / 67.75 / 0.1521	92.28 / 73.75 / 0.1775	91.95 / 87.40 / 0.2157
Gemini 3.8 Flash	94.50 / 70.52 / 0.3789	97.94 / 102.60 / 0.5476	97.24 / 126.87 / 0.7519
Kimi K3.1 Pro	88.22 / 69.35 / 0.4754	91.25 / 88.99 / 0.6447	96.99 / 101.78 / 0.9273
MiniMax M3.1 Flash	79.79 / 88.67 / 0.0825	77.42 / 139.59 / 0.1465	76.41 / 156.97 / 0.1660
Qwen 3.8 Max	90.67 / 121.91 / 0.7155	93.58 / 101.19 / 0.6293	92.59 / 198.01 / 0.9964

以 Gemini 3.8 Flash 为例，中档的任务解决率为 97.94%，作答耗时中位数 102.60 秒，平均费用 0.5476 元。补充图 22 分别按时间与费用排序。

6. 执行稳定性、执行效率与成本

6.3 计量口径与缺失

作答耗时从 Runner 记录的被测模型首次生成调用开始，计至最后一次生成调用结束，包含期间的多轮处理、工具查询、数据写入及等待时间。计时不包含调用开始前的排队和准备，以及回答完成后的评分、结果核验和业务环境恢复。仅统计有最终回答且各轮起止时间戳齐全的有效运行。

模型费用按官方计价标准计算，采用 Runner 留存的人民币费用记录，仅对有费用记录的运行计算平均值。

指标	有记录的题次	缺少该指标的题次	统计结果
作答耗时	8584	266	中位数 86.55 秒
模型费用	8818	32	平均 0.7111 元

有效运行已记录模型费用合计 6,270.12 元。费用与时长分别使用各自的样本，缺失记录排除。该金额汇总被测模型的已记录费用，不含平台和评估模型费用。

7. 典型业务任务实例

以下五例分别说明资料查询、财务统计、信息维护、创建不完整和跨应用追溯。配图为依据同一次实际记录整理的业务过程示意，按业务阶段合并展示，不是原始日志截图；最后一步是评测核对。案例用于说明任务如何执行与判定，不用于比较不同模型或估计失败占比。

图中呈现一次运行；页末统计同题 30 个配置、每配置 3 次，共 90 个题次。

7.1 查询客户开票资料

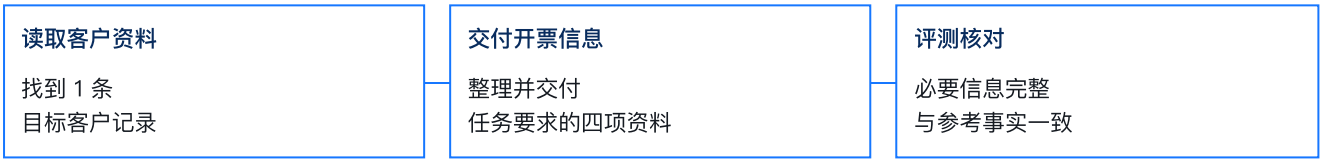
执行模型： DeepSeek V4 Pro（中档）

业务需求：找到指定客户的开票名称、税号、注册地址和注册电话，供开票使用。

客户开票资料：所需信息已完整交付

业务需求：取得指定客户的开票名称、税号、注册地址和注册电话。

本次模型：DeepSeek V4 Pro（中档）



按业务阶段合并实际记录；最后一步为评测核对，前面的整理与交付属于模型处理。

核查内容	任务要求	本次结果
开票名称与税号	对应指定客户	已交付，核对一致
注册地址与注册电话	使用开票所需资料	已交付，核对一致
信息完整性	四项必要资料齐全	满足要求

本次评测：通过 所需资料完整，且与参考业务事实一致。

依据该次实际运行与检查结果整理的业务过程示意。

图 17 业务调用过程与结果核对：客户开票资料

评测结论：通过。所需资料完整且与参考事实一致，满足本次开票资料查询要求。

该题在 30 项模型配置下的汇总结果：88 次通过、2 次失败、0 次无效，有效通过率 97.78%。

BATCH-0134 run_v3_a378a0e6ab77bcec16536869

7. 典型业务任务实例

7.2 计算月度账户净流量

执行模型： DeepSeek V4 Pro（中档）

业务需求：统计指定月份各资金账户的实际收款、实际付款，并计算“收款减付款”的净流量。

账户净流量：按业务口径交付收支结果

业务需求：按指定月份的实际收付款，计算各账户的收款、付款和净流量。

本次模型：DeepSeek V4 Pro（中档）

查看数据结构	读取收付款	读取账户资料	计算并交付	评测核对
收款、付款 及账户资料	收款 17 条 付款 18 条	取得 10 条 账户记录	按账户汇总 交付收支与净额	账户、口径 及金额符合要求

按业务阶段合并实际记录；最后一步为评测核对，前面的整理与交付属于模型处理。

核查内容	任务要求	本次结果
统计期间与款项状态	按实际日期和已确认状态	回答口径通过检查
账户收款与付款	按账户交付相应金额	已交付，核对一致
净流量	按收款减付款计算	已交付，核对一致

本次评测：通过 统计口径、账户及金额事实满足任务要求。

依据该次实际运行与检查结果整理的业务过程示意。

图 18 业务调用过程与结果核对：月度账户净流量

评测结论：通过。各账户的必要统计结果满足任务要求。评测同时关注计算口径与结果，不能只看是否给出了一个净额。

该题在 30 项模型配置下的汇总结果：78 次通过、9 次失败、3 次无效，有效通过率 89.66%。

BATCH-0134 run_v3_327ebd4d4ca44b0f771d9e6b

7. 典型业务任务实例

7.3 更新客户订单的交付约定

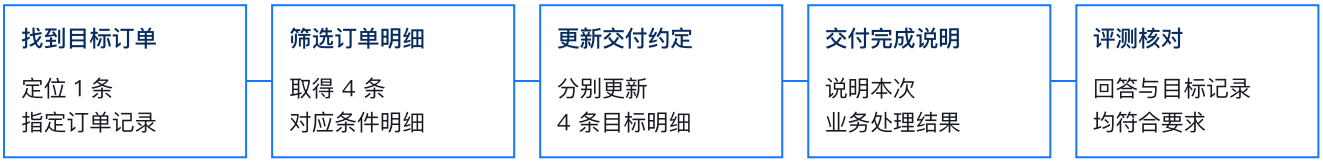
执行模型： Qwen 3.8 Max（低档）

业务需求：在指定订单中，给满足价格条件的明细追加交付要求，并保留任务要求的原有内容。

订单交付约定：目标明细已按要求更新

业务需求：给指定订单中符合价格条件的明细追加交付约定，保留要求的原有内容。

本次模型：Qwen 3.8 Max（低档）



按业务阶段合并实际记录；最后一步为评测核对，前面的整理与交付属于模型处理。

核查内容	任务要求	本次结果
订单与明细范围	限定指定订单和价格条件	取得四条对应明细
目标明细内容	按要求完成交付约定更新	最终记录检查通过
完成说明	与实际办理结果一致	回答检查通过

本次评测：通过 完成说明和目标记录的最终结果均符合要求。

依据该次实际运行与检查结果整理的业务过程示意。

图 19 业务调用过程与结果核对：维护客户订单交付约定

评测结论：通过。完成说明和目标记录的最终结果均符合本题要求。

该题在 30 项模型配置下的汇总结果：83 次通过、7 次失败、0 次无效，有效通过率 92.22%。

BATCH-0145 run_v3_a71faa7d68dc2bf670272bf7

7. 典型业务任务实例

7.4 创建草稿采购申请

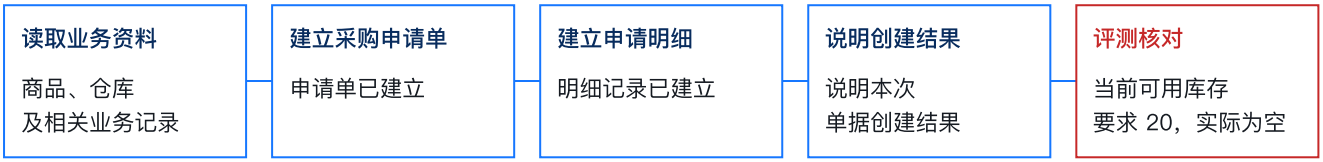
执行模型： Qwen 3.8 Max（中档）

业务需求：按商品现行采购条件创建草稿采购申请，形成完整的采购申请单及明细，不下订单、不入库。

采购申请：单据已建立，明细信息不完整

业务需求：创建草稿采购申请，申请单及明细须完整，不下订单、不入库。

本次模型：Qwen 3.8 Max（中档）



按业务阶段合并实际记录；最后一步为评测核对，前面的整理与交付属于模型处理。

核查内容	任务要求	本次结果
采购申请单	形成符合要求的申请单	已建立，符合要求
申请明细	建立并关联申请明细	记录已建立
明细：当前可用库存	按本题要求记录库存值 20	实际为空，未满足要求

本次评测：未通过　　本题要求的库存值缺失，最终记录检查失败。

依据该次实际运行与检查结果整理的业务过程示意。本题要求的值依据题库库存记录。

图 20　采购申请从业务处理到结果检查的对照

评测结论：未通过。回答检查通过，最终记录检查失败：申请单及明细已建立，但本题要求的库存值为空。创建调用成功与整题通过因此得到不同结论。

该题在 30 项模型配置下的汇总结果：39 次通过、49 次失败、2 次无效，有效通过率 44.32%。

BATCH-0156　run_v3_41c132a9fdcb93040ce4cd26

7. 典型业务任务实例

7.5 从付款追到采购收货

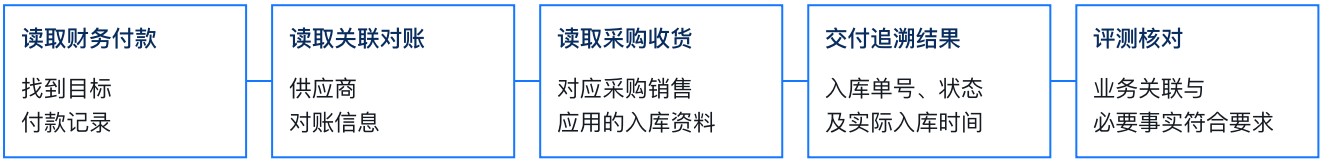
执行模型： DeepSeek V4 Pro（中档）

业务需求：根据指定付款记录，找到对应采购收货，并说明入库单号、入库状态和实际入库时间。

付款追溯收货：跨应用信息已完整交付

业务需求：由指定付款记录找到对应采购收货，说明入库单号、状态和实际时间。

本次模型：DeepSeek V4 Pro（中档）



按业务阶段合并实际记录；最后一步为评测核对，前面的整理与交付属于模型处理。

核查内容	任务要求	本次结果
收货记录及业务关联	沿付款和对账关联定位	已交付对应入库单号
入库状态	取得对应收货状态	已交付，核对一致
实际入库时间	使用入库记录中的实际时间	已交付，核对一致

本次评测：通过 任务要求的跨应用关联与收货信息均满足要求。

依据该次实际运行与检查结果整理的业务过程示意。

图 21 业务调用过程与结果核对：从付款追溯采购收货

评测结论：通过。对应入库单号、状态及实际时间满足本题检查要求；本例只说明这条追溯链的执行过程。

该题在 30 项模型配置下的汇总结果：89 次通过、0 次失败、1 次无效，有效通过率 100.00%。

BATCH-0134 run_v3_1a7f453d18b6f7de6c60226b

8. 评测结论与适用范围

8.1 主要结论

Gemini 3.8 Flash 中档按任务解决率排名第一（285/291，97.94%）；以每项配置的 300 个计划题次计，Kimi K3.1 Pro 高档通过数最多（290/300，96.67%）。库存共 52 题，汇总通过率 90.02%。GPT-6.1-Terra · 低 居首，最高有效通过率 98.72%。

在全部 30 项配置均取得三次有效判定的 38 道共同题目上，每项配置均有 114 个题次。DeepSeek V4 Pro 低档通过 113 次，解决率为 99.12%。这项对照用于检查样本差异对排名的影响。

回答、操作返回和最终业务结果需要分别判断。第 7.4 节的采购申请已建立，回答也通过检查，但明细中的当前可用库存应为 20，实际为空，因此整题未通过。运行记录可帮助区分需要补充回答、修正业务条件，还是补齐目标记录。

重复表现与投入需要共同阅读。执行稳定性、时长和模型费用提供不同信息，成本较低不自动代表任务完成更可靠；不同模型的推理档位也不对应相同投入。具体配置应结合第 4 至第 6 章的实测结果与自身业务条件判断。

8.2 结果的适用方式

本报告可用于相近企业业务任务下的配置初步比较、执行问题定位，以及完成质量与投入的联合观察。实际使用时，先对齐工作目标与任务条件，再阅读对应领域、业务类型和细分组结果。

业务选择问题	对应证据
哪种配置更适合特定工作?	先看相应业务类型和领域，再核对具体题目
未完成的是哪个业务环节?	结合失败类型、回答与最终记录检查，核对调用之后形成的实际结果
能否稳定重复办理?	检查 D6 的三次均通过题数与分母
完成质量需要多少投入?	同时核对任务解决率、时长中位数及平均模型费用

8. 评测结论与适用范围

8.3 覆盖与比较边界

任务组结果不能转化为独立能力分数。有效样本差异、初始环境等价性未完全证实，以及评估模型的判断误差，均限制了对小幅领先和失败原因的解释。缺少完整验证证据的项目未单独计分。

各领域的样本数量不同。9,000 个题次来自同一组 100 道任务；同题重复运行用于观察结果波动。

8.4 数据来源

任务来自 BenchSaaS 题库。通过、失败和无效均取自 2026 年 11 月 19 日回读的 HAP 评测运行状态；作答耗时、模型费用与案例过程取自相应运行 ID 的 Runner 记录。

补充图表

作答耗时与平均模型费用的分别排序

名次	模型配置	作答耗时中位数（秒）	模型配置	平均模型费（元）
1	DeepSeek V4.1 Flash · 低	48.75	GPT-6-Luna · 低	0.0241
2	GLM-5.3 · 低	52.73	GPT-6-Luna · 中	0.0302
3	DeepSeek V4.1 Flash · 中	57.03	GPT-6-Luna · 高	0.0341
4	GLM-5.3 · 中	63.04	DeepSeek V4.1 Flash · 低	0.0809
5	GPT-6.1-Terra · 低	67.75	MiniMax M3.1 Flash · 低	0.0825
6	Kimi K3.1 Pro · 低	69.35	DeepSeek V4.1 Flash · 中	0.1003
7	Gemini 3.8 Flash · 低	70.52	MiniMax M3.1 Flash · 中	0.1465
8	Claude Sonnet 5.5 · 高	73.09	GPT-6.1-Terra · 低	0.1521
9	GPT-6.1-Terra · 中	73.75	MiniMax M3.1 Flash · 高	0.1660
10	DeepSeek V4.1 Flash · 高	79.91	GPT-6.1-Terra · 中	0.1775
11	DeepSeek V4 Pro · 低	80.18	DeepSeek V4.1 Flash · 高	0.2052
12	Claude Sonnet 5.5 · 中	81.22	GPT-6.1-Terra · 高	0.2157
13	GPT-6-Luna · 低	82.06	Gemini 3.8 Flash · 低	0.3789
14	GPT-6.1-Terra · 高	87.40	GLM-5.3 · 低	0.4393
15	MiniMax M3.1 Flash · 低	88.67	Kimi K3.1 Pro · 低	0.4754
16	Kimi K3.1 Pro · 中	88.99	GLM-5.3 · 中	0.5141
17	Claude Sonnet 5.5 · 低	91.42	Gemini 3.8 Flash · 中	0.5476
18	DeepSeek V4 Pro · 中	97.14	Qwen 3.8 Max · 中	0.6293
19	Qwen 3.8 Max · 中	101.19	Kimi K3.1 Pro · 中	0.6447
20	Kimi K3.1 Pro · 高	101.78	Qwen 3.8 Max · 低	0.7155
21	Gemini 3.8 Flash · 中	102.60	Gemini 3.8 Flash · 高	0.7519
22	GPT-6-Luna · 中	104.19	DeepSeek V4 Pro · 低	0.8756
23	GLM-5.3 · 高	106.88	GLM-5.3 · 高	0.9074
24	GPT-6-Luna · 高	111.78	Kimi K3.1 Pro · 高	0.9273
25	DeepSeek V4 Pro · 高	115.23	DeepSeek V4 Pro · 中	0.9430
26	Qwen 3.8 Max · 低	121.91	Qwen 3.8 Max · 高	0.9964
27	Gemini 3.8 Flash · 高	126.87	DeepSeek V4 Pro · 高	1.3254
28	MiniMax M3.1 Flash · 中	139.59	Claude Sonnet 5.5 · 低	2.3949
29	MiniMax M3.1 Flash · 高	156.97	Claude Sonnet 5.5 · 中	3.1224
30	Qwen 3.8 Max · 高	198.01	Claude Sonnet 5.5 · 高	3.2358

图 22　作答耗时与平均模型费用的分别排序

明道云
mingdao.com

版权说明与合作伙伴征集

作者与版权

作者：明道云AI技术组

© 2026 明道云。保留所有权利。

转载本报告或引用其中的文字、图表及评测结果，必须注明出处：“明道云AI技术组《企业应用领域大模型评测报告》（BenchSaaS，V1.0）”。

合作伙伴征集

1. 模型服务商合作

欢迎模型服务商提供 Token 资源，支持 BenchSaaS 持续开展企业应用领域的模型评测，扩大参评模型与任务覆盖。

2. 模型开发商合作

欢迎模型开发商参与合作，获取可公开部分的完整题库和诊断报告，了解模型在企业应用任务中的表现与具体问题，为模型能力改进提供参考。

项目官网

访问项目官网查看完整交互式评测榜单、获取最新版报告并提交合作意向：

mingdao.com/landing/benchsaas