BenchSaaS

Enterprise Application

LLM Benchmark Report

100 Domain Tasks / 30 Model Configurations
Report Version V1.0 | November 2026
Website: mingdao.com/landing/benchsaas

Contents

Section	Page
Executive Summary & Key Findings	3
1. Background & Evaluation Scope	4
2. Task Origins, Design & Dataset Composition	6
3. Execution Platform & Evaluation Mechanism	10
4. Rankings by Domain, Task Type & Dimension	13
5. Sub-Dimension Performance & Failure Analysis	26
6. Execution Stability, Efficiency & Cost	39
7. Representative Business Task Case Studies	43
8. Conclusions & Scope of Applicability	48
Supplementary Charts	50
Copyright & Partner Invitation	51

Executive Summary & Key Findings

BenchSaaS evaluates whether models can understand goals, retrieve accurate facts, respect operational constraints, and deliver usable business outcomes in enterprise applications. This release tests 100 tasks from the BenchSaaS pool across Low, Medium, and High configurations of 10 models, with 3 runs per task, totaling 9,000 execution records.

Results are presented across three ranking categories: business domain, business task type, and evaluation dimension. By final goal and required data sources, the benchmark covers 20 CRM (Customer Relationship Management) tasks, 52 Inventory tasks, 27 Finance tasks, and 1 Cross-App task. CRM covers customer profiles, sales orders, and order follow-up.

- Overall Quality: Across all 30 configurations, the overall Task Resolution Rate is 90.25%. Gemini 3.8 Flash (Medium) ranks first overall at 97.94%, followed by its High configuration at 97.24%.

- CRM: Six configurations tie for first at 100%: Kimi K3.1 Pro (High), Qwen 3.8 Max (Medium), GPT-6-Luna (Medium, High), and DeepSeek V4.1 Flash (Low, High).

- Inventory: GPT-6.1-Terra (Low) ranks first among all 30 configurations in Inventory with a 98.72% Task Resolution Rate.
- Finance: DeepSeek V4 Pro (Low) ranks first at 100% on tasks involving invoices, payments, accounts, and settlements.
- Stability & Cost: Gemini 3.8 Flash (Medium) passes all 3 repeat runs on 90 tasks. Increasing reasoning effort shifts Task Resolution Rates in different directions across models; latency and cost are detailed in Section 6.

1. Background & Evaluation Scope

1.1 Project Background

In enterprise applications, LLMs are deployed to complete concrete business work. Looking up customer profiles, analyzing order progress, reconciling invoices and payments, or creating business documents all require retrieving facts, making judgments, and delivering usable results under specific schemas, record states, and business rules. Response quality must be judged together with business facts and execution outcomes.

Such work places three demands on evaluation: the environment must preserve object relationships so models can interpret data in context; tasks must define clear goals and boundaries so completion is verifiable; and grading must jointly check factual accuracy, operational compliance, and post-mutation record states. General capability benchmarks must be paired with concrete business execution evidence to guide deployment decisions.

BenchSaaS centers evaluation on enterprise business tasks, connecting natural-language instructions, HAP business apps, model tools, and outcome verification across fixed configurations and repeated runs to measure completion quality, performance variance, and execution cost.

This release focuses on four domains (CRM, Inventory, Finance, Cross-App), organizes work into seven business task types, and evaluates performance, repeat stability, and execution cost across seven primary dimensions – providing grounded evidence with task context and run logs for model selection.

1.2 Positioning & Objectives

This report evaluates individual business task runs as its basic unit, observing joint performance across model configurations, tools, and the execution environment.

Evaluation Objective	Supporting Evidence
Task Completion Quality	Verification of responses and final business records
Performance Across Scenarios	Rankings by domain, task type, and dimension
Repeat Execution Reliability	Proportion of tasks passing all three repeated runs
Execution Cost & Effort	Valid run logs for response latency and model cost

1. Background & Evaluation Scope

1.3 Evaluated Configurations & Sample

Evaluation runs took place from Oct 22 to Nov 19, 2026 (UTC+8), with final statuses synced from HAP on Nov 19, 2026. Ten models were tested across Low, Medium, and High reasoning tiers on the same 100 tasks (3 runs per task), totaling 30 configurations and 9,000 Planned Runs.

Model	Low	Medium	High
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

Table shows actual requested reasoning effort parameters. Both GPT models use reasoning.effort; others use reasoning_effort. Both DeepSeek models and GLM also set thinking.type=enabled; Qwen sets enable_thinking=true. Low, Medium, and High denote benchmark tiers rather than equal compute across models.

1.4 How to Read the Results

Task Resolution Rate equals Passes / (Passes + Failures), with invalid runs listed separately. Aggregate Task Resolution Rates are computed as total Valid Passes divided by total Valid Runs across configurations, rather than the simple arithmetic mean of configuration rates.

Overall, domain, task type, and dimension rankings are ordered by Task Resolution Rate from highest to lowest. In two-column charts, read the left column before the right. Ranks are computed from unrounded ratios (ties share the same rank). Stability, latency, and cost follow the specific sample rules noted on each chart.

Rankings reflect empirical results on this task set. Valid run counts, inter-task similarity, and environment conditions affect comparisons. Because environment IDs and data snapshots vary across batches, full initial-state equivalence cannot be proven from retained logs; minor rank differences should not be over-interpreted as durable capability gaps.

2. Task Origins, Design & Dataset Composition

2.1 Task Origins & Business Environment

Tasks are drawn from the BenchSaaS task pool, built upon two HAP applications – Procurement & Sales and Financial Management – within a shared BenchSaaS test organization, using customers, products, orders, inventory, fulfillment, reconciliation, invoices, and payments to reflect realistic operational requests.

The baseline data snapshot recorded in the task pool index was captured on Oct 9, 2026 at 11:23, comprising 40 business tables and 18,014 records, with organization members and departments defined by test org profiles. This snapshot underpins task design, while run logs govern actual scoring.

Task design integrates business goals, data relationships, and operational constraints: organizing goals around query, aggregation, judgment, tracing, verification, maintenance, and workflow intake; grounding facts in linked records; and bounding valid outcomes via preconditions, dates, states, units, and mutation scopes. Only structured table records, app schemas, and org data are used (no knowledge bases, attachments, or scans).

2.2 Task Design Principles

Business domains define functional scope; business task types define the nature of work; evaluation dimensions define the axes of quality and cost observation. Each is classified and tallied independently.

Tasks are designed around complete, verifiable business outcomes. Prompts specify goals, target entities, and boundaries, leaving models to identify required data and processing paths. Difficulty arises from condition combinations, multi-table links, partial fulfillment, and exceptions; each task maps to one task type by final outcome.

Clear goals, open paths: Prompts provide sufficient context to identify target objects and outcome boundaries without prescribing table names, field IDs, API endpoints, or fixed steps; similar fields are distinguishable via business cues.

Evidence-based facts, deterministic rules: Reference answers are derived from snapshot data and business logic. Quantities and units jointly form facts; dates, states, amounts, and rounding follow prompt specifications, and missing facts must never be guessed.

Outcome-driven verification: Read-only tasks require accurate facts and conclusions; mutation tasks verify both the response and target documents/line items. Required field values are defined by the prompt, source records, or app defaults – whether filled by the model or auto-populated by the app.

Task Delivery Requirements

Query, aggregation, judgment, and tracing tasks require verifiable facts and supporting basis. Maintenance and creation tasks verify target records and line items. When prompt preconditions are unmet, halting execution and stating the reason fulfills the task.

2. Task Origins, Design & Dataset Composition

Definitions and Covered Examples of the Seven Business Task Types

Task Type	General Definition	Covered Examples in V1.0
Query Business Records	Retrieve, locate, and deliver essential information for specified business entities.	Customer profiles, quotes, and targeted record lookups.
Aggregate Business Data	Compute business aggregations under specified time, status, grouping, and calculation rules.	Aggregations over order amounts, inventory, and financial accounts.
Business Judgment & Decision Support	Compare options, evaluate conditions, or provide actionable recommendations from business facts and constraints.	Inventory shortfalls, safe transfer margins, and quote evaluations.
Lineage & Progress Tracing	Trace origins, processing history, current progress, or historical states across linked records.	Order fulfillment, receipt/return batches, and invoice–payment lineage.
Audit, Verification & Reconciliation	Check consistency across business facts, records, or amounts, identifying and explaining discrepancies.	Reconciling quantities, dates, warehouse terms, invoice–payments, and cycle counts.
Maintain Business Data	Update or supplement existing business records within prescribed targets and impact boundaries.	Updating order delivery terms, quotes, supplier profiles, and safety stock.
Create Business Documents	Create required business documents with necessary headers and line items; halt and explain if preconditions fail or duplicates exist.	Creating purchase requisitions, count sheets, and payment plans, or halting when conditions fail.

In this release, Create Business Documents covers creating purchase requisitions, inventory count sheets, and payment/collection plans, as well as halting creation when preconditions fail or records already exist (19 tasks total: 16 creation, 3 conditional halt). Scores reflect these intake steps; downstream approval, physical stock movement, and payment confirmation are outside scope.

Each task stores the prompt, design notes, ground–truth evidence, and reference answer; mutation tasks also specify expected record changes and protected scope. Reference notes guide grading and are hidden from evaluated models; any valid data–processing path is accepted. See Section 3 for the grading mechanism.

2. Task Origins, Design & Dataset Composition

2.3 Benchmark Dataset Composition

The 500-task candidate pool comprises 100 public dev tasks, 300 formal hidden tasks, and 100 reserve tasks. This release samples 20, 69, and 11 tasks from each subset respectively.

Seven Business Task Types: Candidate Pool vs. V1.0 Sample

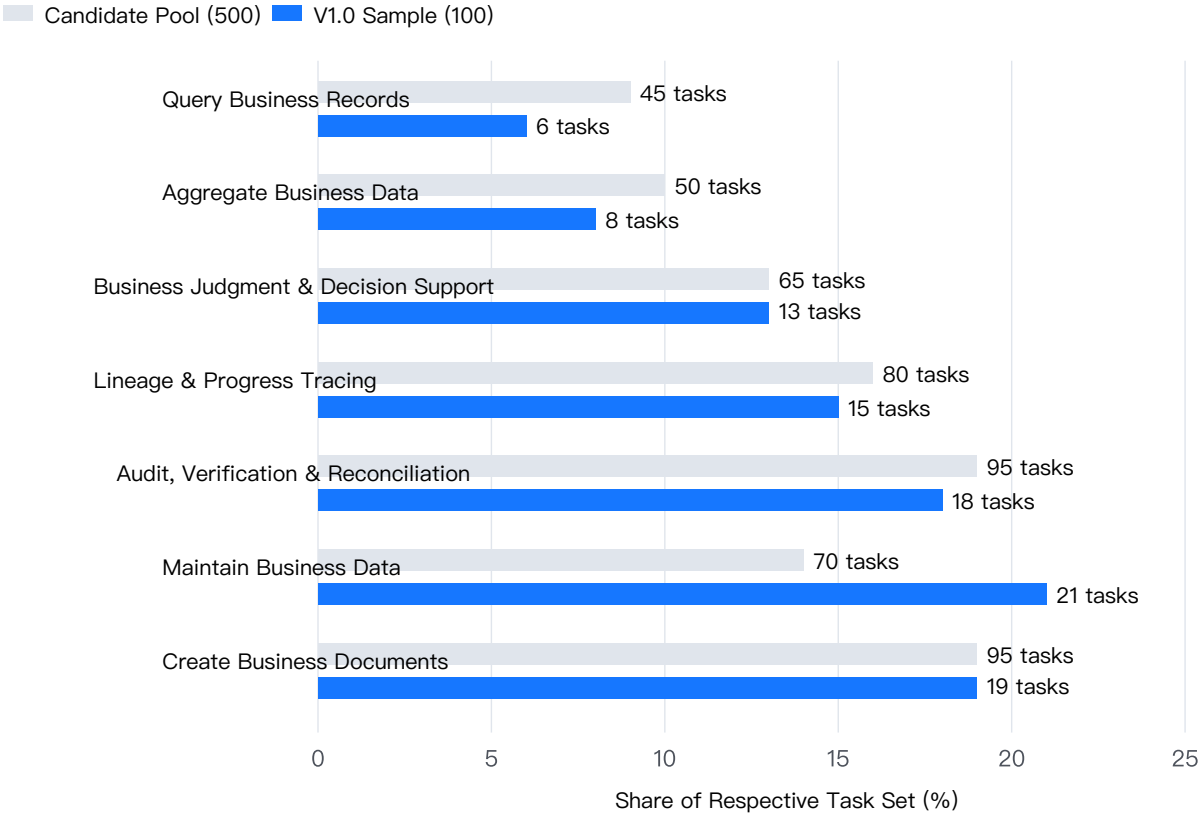


Figure 1 Distribution of the Seven Business Task Types in Candidate Pool vs. V1.0 Sample

2. Task Origins, Design & Dataset Composition

2.4 Sample Characteristics & Coverage Boundaries

Each task is assigned to one primary domain by its final business goal, while required data sources determine whether it is Cross-App; auxiliary context is not double-counted.

Domain	Tasks	General Definition	Covered Examples in V1.0
CRM	20	Managing customer profiles, sales activities, and customer-related affairs.	Customer profiles, sales order analysis, delivery terms, and customer order progress.
Inventory	52	Inventory, warehousing operations, and related procurement workflows.	Procurement, stock shortfalls, transfers, cycle counts, inbound/outbound, and return batches.
Finance	27	Invoices, receivables/payables, financial accounts, and revenue/expense operations.	Invoice reconciliation, accounts, payment/collection plans, and collection follow-up.
Cross-App	1	Work requiring retrieval or processing of essential information across multiple business applications.	Tracing from financial payments and reconciliation to purchase receipt status and actual timestamps.

Sales order follow-up and warehouse batch handling are separated by primary goal; multi-table invoice-payment tracing within Finance remains Finance. Only tasks requiring facts delivered from both applications count as Cross-App.

The 100 tasks comprise 60 Read-Only Tasks and 40 Data-Mutation Tasks (16 creation, 20 update, and 4 required no-change/rejection tasks). Classified by execution requirements: Read-Only Tasks cover query, aggregation, judgment, tracing, and verification, rather than just the 6 "Query Business Records" tasks among the seven task types; the 21 "Maintain Business Data" tasks also include 1 non-mutation task and cannot be equated to the 20 actual update tasks. Basic, medium, and complex tasks number 8, 36, and 56; routine normal workflows, complex normal chains, and exceptions or unmet conditions number 46, 39, and 15.

Complex tasks account for 175/500 (35%) in the candidate pool and 56/100 (56%) in this round. Because this round leans toward complex tasks with uneven domain and task-type counts, rankings should be interpreted against this sample composition.

3. Execution Platform & Evaluation Mechanism

3.1 HAP: An Application Platform for Enterprise Data and Workflows

HAP (Hyper Application Platform) is Mingdao Cloud's enterprise application platform. Through visual configuration, users build business data, views, workflows, and role permissions to organize business objects into sustainable applications. Worksheets store master data and documents as business tables, views present data for specific scenarios, workflows automate data processing and routing, and charts and custom pages deliver visual reporting.

The platform structures data relationships via relations, formulas, and rollups, and provides standard open APIs for external systems or model tools to access application data, offering a unified data and operation entry point for enterprise task execution.

3.2 HAP: An Ideal Testbed for Enterprise Application Scenarios

BenchSaaS selects an APaaS (Application Platform as a Service) like HAP as the testbed for enterprise scenarios based on three core attributes—universality, isomorphism, and comprehensive logging—enabling realistic business hosting alongside unified task execution, result verification, and failure diagnosis.

Universality: Across CRM, procurement, Inventory, and Finance, enterprise apps share common building blocks of business objects, relational data, workflows, and permissions. APaaS configures these scenarios with general–purpose capabilities covering lookup, aggregation, verification, maintenance, and document creation, extending evaluation across diverse enterprise workflows.

Isomorphism: Distinct business applications share a common data model, relational mechanism, and operational interface. Models interact with a consistent operational system across varied business semantics, allowing the benchmark to reuse tools and verification methods while minimizing interface noise. Isomorphism does not imply identical initial historical states across all runs; such comparison boundaries are noted in this report.

Comprehensive Logging: Platform logs of data mutations and workflow executions, combined with model interactions, tool calls, and run records captured by the evaluation harness, link business instructions, execution traces, and final outcomes to pinpoint comprehension errors, call failures, or missing records while analyzing latency and cost. Specific diagnoses rely on the logs retained for each run.

The HAP Model Evaluation Center manages tasks, model run configurations, and evaluation results; the automated harness schedules tasks, records execution traces, and connects to result verification. Platform capabilities do not imply full measurement of every aspect in this round, and items lacking complete verification evidence are not scored separately.

3.3 How Tasks Are Executed

Each run begins with a fixed task and evaluation configuration. After receiving a business instruction, the model uses data operation tools to inspect application schemas, search records, view line items and relations, and aggregate data; when permitted, it can also create or update business records. These tools serve as the model's operational gateway to business data.

Once the model completes its response, the benchmark verifies required business facts and inspects final business records or checks for prohibited mutation calls according to task requirements, yielding a pass, fail, or invalid verdict. Models choose their own execution paths without a prescribed operation sequence.

For data–mutation tasks, execution results are saved per applicable procedures and test data restoration is scheduled. Completion of restoration alone does not prove that out–of–scope data remained untouched during execution; items without complete verification evidence are not scored separately.

3. Execution Platform & Evaluation Mechanism

3.4 How Task Completion Is Determined

Task Scenario	Verification Scope	Completion Criteria
Lookup, Aggregation, Judgment & Tracing	Check response against reference facts	Required objects, values, states, relations, and rationales are accurate without contradiction.
Record Creation or Modification Required	Check response and inspect final records	Target records, required fields, and mutual relations meet task requirements.
Non-Mutation Required Tasks	Check response and verify no prohibited mutations	State accurate facts and stopping reasons while obeying operational constraints.

Evaluation combines programmatic checks with an independent judge model. Programmatic checks verify applicable inspection rules, operational constraints, and call limits, while the judge model compares responses, reference answers, and applicable final states by business meaning. Here "Pass" means meeting this round's inspection criteria; semantic judgment may still carry residual error, and quantitative inter-rater agreement from human review is not provided in this release.

Evaluation accepts semantically equivalent phrasing without requiring fixed response templates. Missing essential facts, wrong entities or amounts, or contradictory states and relations can all result in failure. Creating a header record cannot substitute for complete line items, nor can stating "completed" in text substitute for actual database outcomes.

This report reads final states from HAP by run ID. Among records adjudicated as passes after review, 35 overturned erroneous initial verdicts and 122 relaxed overly strict initial rubrics; all 157 are counted as Valid Passes.

3. Execution Platform & Evaluation Mechanism

3.5 Repeated Runs, Execution Limits, and Outcome Classification

Each task–configuration pair is run 3 times. Official configuration snapshots set limits of 31 interaction turns, 30 tool calls, and a 1,800–second timeout.

HAP Status	Included Runs	Statistical Treatment
PASS	7,987	Counted in both numerator and denominator of Resolution Rate
FAIL	863	Counted only in denominator of Resolution Rate
INVALID	150	Counted as invalid runs; excluded from Resolution Rate denominator

Out of 9,000 Planned Runs in total, 8,850 Valid Runs received a pass or fail verdict.

Division of Roles in the BenchSaaS Execution Platform

HAP Model Evaluation Center	Automated Execution & Grading	HAP Business Applications
Task & model run configurations	Model interaction & business data processing	Procurement & Sales: 32 business tables
Execution traces & results per batch	Result verification & trace retention	Financial Management: 8 business tables
Scoring & model comparison results	Test data restoration & result saving	Provision of business master data & documents

The Evaluation Center manages tasks and results; the automated harness handles model interaction, data processing, and grading; business apps supply test data. Table counts reflect structural audits during report compilation and do not prove identical initial states across historical runs.

Figure 2 Division of Roles Across Task Management, Automated Execution, and Business Data

4. Rankings by Domain, Task Type & Dimension

4.1 Overall Leaderboard

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

Figure 3 Overall Task Resolution Rate Leaderboard

Across the full 100-task set, Gemini 3.8 Flash (Medium) ranks first at 97.94%. To compare on an identical sample, we also examine the 38 tasks where all 30 configurations obtained three Valid Runs (114 runs each): DeepSeek V4 Pro (Low) passes 113 runs (99.12% Resolution Rate). This subset serves solely to check the sensitivity of rankings to sample differences.

4. Rankings by Domain, Task Type & Dimension

4.2 Business Domain Leaderboards

CRM comprises 20 tasks with a 95.31% pooled resolution rate. Six configurations tie for first at 100.00%.

Rank	Model Configuration	Resolution Rate	Pass / Valid	0 — 100%
1	Kimi K3.1 Pro · High	100.00%	59 / 59	
1	Qwen 3.8 Max · Med	100.00%	58 / 58	
1	GPT-6-Luna · High	100.00%	58 / 58	
1	GPT-6-Luna · Med	100.00%	59 / 59	
1	DeepSeek V4.1 Flash · Low	100.00%	60 / 60	
1	DeepSeek V4.1 Flash · High	100.00%	58 / 58	
7	Gemini 3.8 Flash · Low	98.33%	59 / 60	
7	Claude Sonnet 5.5 · High	98.33%	59 / 60	
7	GLM-5.3 · Med	98.33%	59 / 60	
10	DeepSeek V4 Pro · Med	98.31%	58 / 59	
10	DeepSeek V4 Pro · Low	98.31%	58 / 59	
10	GLM-5.3 · Low	98.31%	58 / 59	
13	Gemini 3.8 Flash · Med	96.67%	58 / 60	
13	Claude Sonnet 5.5 · Low	96.67%	58 / 60	
13	Claude Sonnet 5.5 · Med	96.67%	58 / 60	
13	Kimi K3.1 Pro · Low	96.67%	58 / 60	
17	GLM-5.3 · High	96.61%	57 / 59	
17	Qwen 3.8 Max · High	96.61%	57 / 59	
17	DeepSeek V4 Pro · High	96.61%	57 / 59	
17	DeepSeek V4.1 Flash · Med	96.61%	57 / 59	
17	GPT-6.1-Terra · High	96.61%	57 / 59	
22	GPT-6.1-Terra · Med	95.00%	57 / 60	
23	Qwen 3.8 Max · Low	93.33%	56 / 60	
24	Gemini 3.8 Flash · High	92.98%	53 / 57	
25	Kimi K3.1 Pro · Med	91.53%	54 / 59	
26	MiniMax M3.1 Flash · Low	89.29%	50 / 56	
27	GPT-6-Luna · Low	88.14%	52 / 59	
28	GPT-6.1-Terra · Low	86.67%	52 / 60	
29	MiniMax M3.1 Flash · Med	85.71%	48 / 56	
30	MiniMax M3.1 Flash · High	76.27%	45 / 59	

Figure 4 CRM: Customer and Sales Order Management

4. Rankings by Domain, Task Type & Dimension

Model Leaderboard in the Inventory Domain

Inventory comprises 52 tasks with a 90.02% pooled resolution rate. GPT-6.1-Terra (Low) leads at 98.72%.

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

Figure 5 Inventory: Procurement, Stock, and Warehouse Operations

4. Rankings by Domain, Task Type & Dimension

Model Leaderboard in the Finance Domain

Finance comprises 27 tasks with an 86.56% pooled resolution rate. DeepSeek V4 Pro (Low) leads at 100.00%.

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

Figure 6 Finance: Invoices, Accounts, and Receipts & Payments

4. Rankings by Domain, Task Type & Dimension

4.3 Business Task Type Leaderboards

Task Group	Pass / Valid	Invalid	Resolution Rate
Query Business Records	444 / 523	17	84.89%
Aggregate Business Data	605 / 682	38	88.71%
Business Judgment & Decision Support	1066 / 1155	15	92.29%
Lineage & Progress Tracing	1270 / 1332	18	95.35%
Audit, Verification & Reconciliation	1522 / 1600	20	95.12%
Maintain Business Data	1782 / 1872	18	95.19%
Create Business Documents	1298 / 1686	24	76.99%

The seven business task types classify tasks into mutually exclusive groups by final objective. The table pools runs across all 30 configurations to compute category resolution rates; subsequent charts rank all 30 configurations within each task type.

4. Rankings by Domain, Task Type & Dimension

Model Leaderboards Across the Seven Business Task Types

Ranked by Task Resolution Rate from highest to lowest (read left column first, then right column). Each row shows rank, model tier, and resolution rate. Ranks are determined by unrounded ratios, with identical ratios tied.

Query Business Records

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

Aggregate Business Data

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

Figure 7 Model Leaderboards Across the Seven Business Task Types (1/4)

4. Rankings by Domain, Task Type & Dimension

Model Leaderboards Across the Seven Business Task Types (Cont.)

Ranked by Task Resolution Rate from highest to lowest (read left column first, then right column). Each row shows rank, model tier, and resolution rate. Ranks are determined by unrounded ratios, with identical ratios tied.

Business Judgment & Decision Support

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

Lineage & Progress Tracing

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

Figure 7 Model Leaderboards Across the Seven Business Task Types (2/4)

4. Rankings by Domain, Task Type & Dimension

Model Leaderboards Across the Seven Business Task Types (Cont.)

Ranked by Task Resolution Rate from highest to lowest (read left column first, then right column). Each row shows rank, model tier, and resolution rate. Ranks are determined by unrounded ratios, with identical ratios tied.

Audit, Verification & Reconciliation

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

Maintain Business Data

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

Figure 7 Model Leaderboards Across the Seven Business Task Types (3/4)

4. Rankings by Domain, Task Type & Dimension

Model Leaderboards Across the Seven Business Task Types (Cont.)

Ranked by Task Resolution Rate from highest to lowest (read left column first, then right column). Each row shows rank, model tier, and resolution rate. Ranks are determined by unrounded ratios, with identical ratios tied.

Create Business Documents

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

Figure 7 Model Leaderboards Across the Seven Business Task Types (4/4)

4. Rankings by Domain, Task Type & Dimension

4.4 Evaluation Dimension Task Group Leaderboards

The seven evaluation dimensions follow general definitions, while this report presents performance strictly based on empirical run evidence collected in this round. Scope covered by a general definition does not imply full measurement in this release.

Dimension	General Definition	Statistical Basis in This Release
D1 Business Understanding	Understand business goals, objects, field semantics, and relationships.	Whole-task pass rates across 29 mapped tasks.
D2 Instruction Following	Fulfill user-specified conditions, scope, and deliverables.	Whole-task pass rates across 26 mapped tasks.
D3 Result Accuracy & Completeness	Ensure business facts, calculations, and execution outcomes are accurate and complete.	Whole-task pass rates across all 100 tasks.
D4 Constraint Compliance	Adhere to business preconditions, permitted actions, and impact scope.	13 mapped tasks
D5 Exception Handling & Fact Reliability	Detect discrepancies or missing info, and process, report, or halt grounded in facts.	21 mapped tasks; only 1 task on missing info & truthful reporting.
D6 Execution Stability	Maintain consistent completion quality across repeated runs of the same task.	Share of 3/3 passes among tasks with three Valid Runs per configuration.
D7 Execution Efficiency & Cost	Time, interaction turns, tool operations, and monetary cost required to complete tasks.	Median duration and mean model cost across Valid Runs with timing or cost logs.

Task groups for D1–D5 can overlap and are evaluated using whole-task pass outcomes. The table below pools runs across all 30 configurations for each dimension's task group. D6 measures the 3/3 pass rate across repeated runs, while D7 summarizes Valid Runs with timing or cost records.

Task Group	Pass / Valid	Invalid	Resolution Rate
D1 Business Understanding	2423 / 2571	39	94.24%
D2 Instruction Following	2176 / 2313	27	94.08%
D3 Result Accuracy & Completeness	7987 / 8850	150	90.25%
D4 Constraint Compliance	995 / 1151	19	86.45%
D5 Exception Handling & Fact Reliability	1787 / 1868	22	95.66%

4. Rankings by Domain, Task Type & Dimension

D1–D5 Mapped Task Group Leaderboards

Ranked by Task Resolution Rate from highest to lowest (read left column first, then right column). Each row shows rank, model tier, and resolution rate. Ranks are determined by unrounded ratios, with identical ratios tied.

D1 Business Understanding

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

D2 Instruction Following

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

Figure 8 Model Leaderboards for D1–D5 Mapped Task Groups (1/3)

4. Rankings by Domain, Task Type & Dimension

D1–D5 Mapped Task Group Leaderboards (Cont.)

Ranked by Task Resolution Rate from highest to lowest (read left column first, then right column). Each row shows rank, model tier, and resolution rate. Ranks are determined by unrounded ratios, with identical ratios tied.

D3 Result Accuracy & Completeness

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

D4 Constraint Compliance

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

Figure 8 Model Leaderboards for D1–D5 Mapped Task Groups (2/3)

4. Rankings by Domain, Task Type & Dimension

D1–D5 Mapped Task Group Leaderboards (Cont.)

Ranked by Task Resolution Rate from highest to lowest (read left column first, then right column). Each row shows rank, model tier, and resolution rate. Ranks are determined by unrounded ratios, with identical ratios tied.

D5 Exception Handling & Fact Reliability

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

Figure 8 Model Leaderboards for D1–D5 Mapped Task Groups (3/3)

5. Sub-Dimension Performance & Failure Analysis

5.1 D1 Business Understanding

Ranked by Task Resolution Rate from highest to lowest (read left column first, then right column). Each row shows rank, model tier, and resolution rate. Ranks are determined by unrounded ratios, with identical ratios tied.

D1.1 Business Objects & Information Semantics

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

D1.2 Time, Status & Numerical Definitions

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

Figure 9 D1 Sub-Group Leaderboards (Whole-Task Outcomes) (1/2)

5. Sub–Dimension Performance & Failure Analysis

5.1 D1 Business Understanding (Cont.)

Ranked by Task Resolution Rate from highest to lowest (read left column first, then right column). Each row shows rank, model tier, and resolution rate. Ranks are determined by unrounded ratios, with identical ratios tied.

D1.3 Business Relationship Understanding

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

Figure 9 D1 Sub–Group Leaderboards (Whole–Task Outcomes) (2/2)

5. Sub-Dimension Performance & Failure Analysis

5.2 D2 Instruction Following

Ranked by Task Resolution Rate from highest to lowest (read left column first, then right column). Each row shows rank, model tier, and resolution rate. Ranks are determined by unrounded ratios, with identical ratios tied.

D2.1 Condition & Scope Adherence

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

D2.2 Required Information Delivery

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

Figure 10 D2 Sub-group Rankings (Task-Level Results) (1/2)

5. Sub-Dimension Performance & Failure Analysis

5.2 D2 Instruction Following (Cont.)

Ranked by Task Resolution Rate from highest to lowest (read left column first, then right column). Each row shows rank, model tier, and resolution rate. Ranks are determined by unrounded ratios, with identical ratios tied.

D2.3 Delivery Requirement Fulfillment

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

Figure 10 D2 Sub-group Rankings (Task-Level Results) (2/2)

5. Sub-Dimension Performance & Failure Analysis

5.3 D3 Result Accuracy & Completeness

Ranked by Task Resolution Rate from highest to lowest (read left column first, then right column). Each row shows rank, model tier, and resolution rate. Ranks are determined by unrounded ratios, with identical ratios tied.

D3.1 Query & Calculation Results

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

D3.2 Information Update Results

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

Figure 11 D3 Sub-group Rankings (Task-Level Results) (1/2)

5. Sub–Dimension Performance & Failure Analysis

5.3 D3 Result Accuracy & Completeness (Cont.)

Ranked by Task Resolution Rate from highest to lowest (read left column first, then right column). Each row shows rank, model tier, and resolution rate. Ranks are determined by unrounded ratios, with identical ratios tied.

D3.3 Document Creation Completeness

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

Figure 11 D3 Sub–group Rankings (Task–Level Results) (2/2)

5. Sub-Dimension Performance & Failure Analysis

5.4 D4 Business Constraint Adherence

Ranked by Task Resolution Rate from highest to lowest (read left column first, then right column). Each row shows rank, model tier, and resolution rate. Ranks are determined by unrounded ratios, with identical ratios tied.

D4.1 Business Preconditions

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

D4.3 Duplicate Processing & Handling Boundaries

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

Figure 12 D4 Sub-group Rankings (Task-Level Results)

5. Sub-Dimension Performance & Failure Analysis

5.5 D5 Exception Handling & Factual Reliability

Ranked by Task Resolution Rate from highest to lowest (read left column first, then right column). Each row shows rank, model tier, and resolution rate. Ranks are determined by unrounded ratios, with identical ratios tied.

D5.1 Discrepancy & Anomaly Detection

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

D5.2 Missing Information & Truthful Reporting

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

Figure 13 D5 Sub-group Rankings (Task-Level Results) (1/2)

5. Sub-Dimension Performance & Failure Analysis

5.5 D5 Exception Handling & Factual Reliability (Cont.)

Ranked by Task Resolution Rate from highest to lowest (read left column first, then right column). Each row shows rank, model tier, and resolution rate. Ranks are determined by unrounded ratios, with identical ratios tied.

D5.3 Proper Halting & Reason Explanation

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

Figure 13 D5 Sub-group Rankings (Task-Level Results) (2/2)

5. Sub-Dimension Performance & Failure Analysis

How to Interpret Sub-group Results

The sub-group rankings above are computed from each configuration's task-level resolution rate within the corresponding task group. Groups may overlap; rankings reflect whole-task outcomes and cannot isolate specific capability deficits or internal failure causes. The top-level D3 group matches the overall leaderboard, while D1.3, D3.2, and D5.1 share the exact same task groups as Lineage & Progress Tracing, Maintain Business Data, and Audit, Verification & Reconciliation respectively, providing no separate independent evidence.

Interpretive Scope of Small Samples

Missing Information & Truthful Reporting contains only 1 task, and Proper Halting & Reason Explanation contains 3 tasks. Running across three tiers or multiple trials increases run counts without increasing task variety. These results only explain scenarios involving missing data, unmet preconditions, or existing arrangements in this release.

Mapping Task Groups to Business Requirements

D3 covers all 100 tasks. Query and calculation tasks require complete values, units, and relationships; data maintenance requires verifying both prior content and updated records; document creation requires verifying both header and line-item records. A single task failure may involve one or more stages and must be diagnosed alongside response checks, final record checks, and task requirements.

5. Sub-Dimension Performance & Failure Analysis

5.6 Primary Categories of Failed Runs

> **Among the 863 runs judged as failed in HAP, 593 failed at least one required check, accounting for 68.71%.**

Record Category	Runner Error Code	Count	Share
Required Check Failed	MODEL_REQUIRED_ASSERTION_FAILED	593	68.71%
Unmet Business Condition or Operation Limit	MODEL_BUSINESS_RULE_REJECTED	161	18.66%
Invalid Tool Arguments	MODEL_ARGUMENT_INVALID	51	5.91%
Tool Call Limit Exceeded	MODEL_TOOL_LIMIT_EXCEEDED	45	5.21%
Write Restriction Triggered in Read-Only Task	MODEL_QUERY_WRITE_FORBIDDEN	5	0.58%
Model Service Request Rejected	MODEL_PROVIDER_REQUEST_REJECTED	5	0.58%
Malformed Tool Call	MODEL_TOOL_CALL_MALFORMED	3	0.35%

Figure 14 Primary Record Categories Across 863 Failed Runs

These 593 runs comprise 428 response check failures and 165 final record check failures. Classification is based on the required-check verdict of the final attempt, with each failed run counted in a single category.

"Unmet Business Condition or Operation Limit" refers to tool requests rejected due to business preconditions or task-level restrictions on allowed objects, operations, or field scopes. This category describes benchmark runtime restrictions and does not imply that HAP's native "Business Rules" feature was triggered.
Categories indicate which stage failed to meet requirements rather than isolating internal model causes; an intercepted operation also does not mean an actual business mutation occurred. See Chapter 7 for real examples.

> New in This Edition (5.6.1–5.6.2): Drawing on raw logs from the 863 failed runs in the ESAB organization's HAP BenchSaaS app (failed_assertions_json, tool_call_log, and model_turn_log), pages 37–38 break down the specific root causes across all 7 failure categories, spanning response assertions, database snapshot persistence, tool-scope interception, parameter validation, and call budgets.

5. Sub–Dimensions & Failure Analysis

5.6.1 Required Assertion Failures (593 Runs) – Raw HAP Breakdown

"Required Assertions" are two sequential verification gates evaluated at the end of every task: Gate 1 (ANSWER_MATCH, mandatory across all 100 tasks) checks the final response for factual accuracy, calculation rules, cited evidence, and mutation receipts; Gate 2 (RESULT_CONSISTENCY, triggered on write tasks) inspects the frozen HAP database snapshot. Raw record inspection reveals that 100% of the 165 Gate 2 failures had already passed Gate 1 – their written reply sounded complete, yet the database snapshot exposed NULL required fields, empty status arrays [], or claims contradicting stored records.

I. Gate 1: Answer Failed Verification (ANSWER_MATCH – 428 Runs, 72.18% | 245 Read–Only + 183 Write)

Failure Sub–Category	Observed Behavior & Business Impact in Raw HAP Records	Runs / Share
Read–Only: Filter, Rule, or Calc Error READ – Filter / Rule / Calc	Missed or over–included target records (e.g., listing records when the valid filter set was empty, or finding only 5–9 of 15 statements); violated cross–unit rules ('quantities in different UoMs must not be summed'); or mixed 'Pending' into 'Confirmed' amounts.	**156 (36.4%)**
Write: Omitted Required Receipt Elements WRITE – Receipt Omitted	After mutating records, the reply was overly brief (e.g., merely stating 'registered/appended') and omitted mandatory confirmation items: failing to state preserved handler ownership, quote appended notes verbatim, or confirm other lines were untouched.	**97 (22.7%)**
Read–Only: Self–Contradiction, Missing Evidence, or Wrong Unit/ID READ – Contradiction / Unit / ID	Line–item table was right but summary header contradicted it (38 runs); omitted document status, extremes, underlying voucher IDs, or follow–up advice (31 runs); or output wrong UoMs / raw internal UUIDs instead of business codes (20 runs).	**89 (20.8%)**
Write: Incomplete Mutation or Wrong Target Scope WRITE – Scope / Incomplete	Batch updates missed qualifying rows or modified the wrong record (23 runs); reply admitted newly created document status remained blank instead of 'Pending Confirmation/Inbound' (20 runs); or missed designated applicant/dept or returned empty text (43 runs).	**86 (20.1%)**

II. Gate 2: Final DB Record Failed Check (RESULT_CONSISTENCY – 165 Runs, 27.82% | 147 Create + 18 Update)

Failure Sub–Category	Observed Behavior & Business Impact in Raw HAP Records	Runs / Share
Required Field or Table Relation Left NULL in DB NULL Required Field / Link	Header record was created, but required line–item or header fields remained NULL in HAP (models often assumed linking a product auto–populates text and stock). Top gaps: available stock (17 runs, Case 7.4), line product code/name (13), header handler/dept/rep (11), return unit cost (9), and physical relation link (7).	**58 (35.2%)**
Wrong Field Value, Code Prefix, or Mutation Scope Wrong Code / Value / Scope	Count/transfer lines wrote inventory code KC–* into product code SP–* (22 runs); notes missed a period, added extra bracketed text, or used full–width digits (8); wrong stock qty or swapped transfer–in/out warehouses (7); wrong dept or invalid relation ID (6); missed batch rows or duplicate creation (4).	**47 (28.5%)**
Core Record Correct, but Reply Contradicted DB Snapshot Claim vs. DB Mismatch	Core database fields met criteria, but additional claims in the model's reply conflicted with the frozen snapshot ('reply–to–ledger mismatch') – e.g., reporting tax–inclusive subtotal as tax amount, deducting pending inbound qty from remaining receivable early, or misstating original payable.	**33 (20.0%)**
Status Option Left Empty [] or Premature Actual Timestamp Empty Status [] / Timestamp	Reply claimed status was set to 'Pending Confirmation / Inbound / Outbound', but the API write omitted the option key so HAP stored an empty array [] (API writes do not trigger UI form defaults); or wrongly set status to Confirmed or filled actual inbound timestamp on a pending order.	**27 (16.4%)**

Methodology Note: Breakdown derived from mutually exclusive attribution across all 593 official V1.0 MODEL_REQUIRED_ASSERTION_FAILED runs in the HAP BenchSaaS application (428 + 165 = 593; examples are anonymized with zero disclosure of hidden task IDs or sensitive test entity data).

5. Sub–Dimensions & Failure Analysis

5.6.2 Remaining Six Failure Categories (270 Runs) – Raw HAP Breakdown

> Beyond required assertion failures (593 runs), the remaining 270 failed runs (31.29% of 863) resulted from harness tool–scope rejections, argument schema validation errors, 30–call budget exhaustion, and read–only write guardrails. Below is the exact empirical breakdown verified from HAP run records, tool_call_log, and model_turn_log.

Failure Category & Stage	Empirical Breakdown & Root–Cause Evidence in Raw HAP Logs	Runs / Share
Business Rule / Scope Rejected MODEL_BUSINESS_RULE_REJECTED 161 runs (18.66%) – TOOL_EXECUTION	160/161 terminated after two consecutive rejections with zero DB mutations (135 runs persisted granular tool_call_log): – 75 runs (46.6%): create_record submitted read–only or non–whitelisted fields (55 logged runs 100% hit FIELD_OUT_OF_SCOPE); – 74 runs (46.0%): worksheet structure (58) / list (15) / detail (1) reads passed invalid worksheet identifiers (66 of 70 logged runs passed short aliases like spxx, ddmx instead of 24–char worksheet IDs; 4 queried out–of–scope tables); – 12 runs (7.5%): get_app_info (7) / get_app_worksheets_list (5) probed unauthorized cross–app metadata.	**161 (18.66%)**
Invalid Tool Arguments MODEL_ARGUMENT_INVALID 51 runs (5.91%) – ARGUMENT_VALIDATION	51/51 (100%) failed pre–execution schema validation twice consecutively (33 runs persisted granular tool_call_log): – 41 runs (80.4%): get_record_list argument errors (31 logged runs: 19 pagination cursor _runner_read.cursor conflicts, 6 filter.children passed as JSON object instead of array, 6 missing required properties like appId; plus 10 unlogged runs); – 9 runs (17.6%): create_record field value JSON type mismatch (e.g., wrapping relation ID as {"item":...} object); – 1 run (2.0%): get_record_relations missing required parameter appId.	**51 (5.91%)**
Tool Call Limit Exceeded MODEL_TOOL_LIMIT_EXCEEDED 45 runs (5.21%) – LOOP_CONTROL	43 runs hit the 30–call cap (2 stopped at 21/29 calls); 39/45 (86.7%) had zero tool errors throughout (tools succeeded but search failed to converge): – 19 runs (42.2%): read–only tasks looped on pagination or filter retries (get_record_list >= 20 calls); – 16 runs (35.6%): read–only multi–table exploration sprawl (14 cross–table scans + 2 per–row get_record_details scans); – 8 runs (17.8%): write tasks exhausted budget on pre–write queries before issuing any write call; – 2 runs (4.4%): write tasks executed the write, then looped re–verifying until hitting the 30–call cap.	**45 (5.21%)**
Write in Read–Only Task MODEL_QUERY_WRITE_FORBIDDEN 5 runs (0.58%) – TOOL_SELECTION	100% occurred in read–only decision–support tasks and were blocked pre–execution (executed = false, zero data mutation): – 3 runs (60.0%): called create_record in a read–only procurement advisory task to create a purchase order; – 2 runs (40.0%): called update_record or write tools during a read–only customer statement follow–up assessment.	**5 (0.58%)**
Provider Request Rejected MODEL_PROVIDER_REQUEST_REJECTED 5 runs (0.58%) – Judge–Stage Rejection	Audit clarification: candidate models had already delivered final answers (answer_available = true) before the automated Judge call was rejected; after manual judge supplement (22 runs upgraded to PASS), these 5 runs were confirmed substantive failures: – 3 runs (60.0%): prematurely abandoned Top–3 aggregation across paginated records; – 2 runs (40.0%): misapplied filter rules on an empty–set task and falsely listed 5–10 unrelated statements.	**5 (0.58%)**
Malformed Tool Call MODEL_TOOL_CALL_MALFORMED 3 runs (0.35%) – MODEL_RESPONSE	3/3 occurred in write tasks after 5–15 successful tool calls, emitting a corrupted tool–call payload immediately after a valid tool result: – 2 runs (66.7%): emitted malformed tool calls mid–flow during multi–step return/count creation; – 1 run (33.3%): emitted an unparseable payload on turn 5 right after a valid turn–4 update_record call.	**3 (0.35%)**

Methodology Note: 270 = 161 + 51 + 45 + 5 + 5 + 3, summing with 593 to all 863 official FAIL runs in V1.0. Some earlier evaluation batches did not persist granular tool_call_log payloads (135 of 161 and 33 of 51 persisted); exact logged sample sizes are noted in each row.

6. Execution Stability, Efficiency & Cost

6.1 Execution Stability

Model – Tier	3/3 Passed Tasks	3/3 Valid Tasks	Stable Pass Rate	Excluded Tasks
Gemini 3.8 Flash · Med	90	94	95.74%	6
Kimi K3.1 Pro · High	92	99	92.93%	1
DeepSeek V4 Pro · Med	90	97	92.78%	3
DeepSeek V4 Pro · Low	87	94	92.55%	6
Gemini 3.8 Flash · High	85	93	91.40%	7
Gemini 3.8 Flash · Low	84	93	90.32%	7
GLM–5.3 · High	85	95	89.47%	5
GPT–6.1–Terra · High	86	98	87.76%	2
Qwen 3.8 Max · Med	85	97	87.63%	3
GPT–6.1–Terra · Med	85	98	86.73%	2
GPT–6–Luna · High	82	96	85.42%	4
GPT–6.1–Terra · Low	83	99	83.84%	1
Qwen 3.8 Max · High	81	97	83.51%	3
Kimi K3.1 Pro · Med	81	97	83.51%	3
Claude Sonnet 5.5 · Med	82	99	82.83%	1
Claude Sonnet 5.5 · Low	81	98	82.65%	2
GPT–6–Luna · Med	81	99	81.82%	1
DeepSeek V4 Pro · High	79	97	81.44%	3
Qwen 3.8 Max · Low	81	100	81.00%	0
GLM–5.3 · Med	79	98	80.61%	2
DeepSeek V4.1 Flash · Med	77	96	80.21%	4
Claude Sonnet 5.5 · High	80	100	80.00%	0
DeepSeek V4.1 Flash · High	74	95	77.89%	5
DeepSeek V4.1 Flash · Low	75	97	77.32%	3
Kimi K3.1 Pro · Low	73	97	75.26%	3
GLM–5.3 · Low	62	97	63.92%	3
MiniMax M3.1 Flash · Low	57	92	61.96%	8
GPT–6–Luna · Low	54	94	57.45%	6
MiniMax M3.1 Flash · Med	48	88	54.55%	12
MiniMax M3.1 Flash · High	45	87	51.72%	13

Figure 15 D6 Execution Stability: Share of Tasks Passed Across All Three Trials

Each configuration uses tasks with three valid trials as its denominator; the resulting task sets may differ across configurations.

6. Execution Stability, Efficiency & Cost

6.2 Joint View of Completion Quality & Resource Input

Each row corresponds to one model. Within each metric, bars from top to bottom represent Low (blue), Medium (green), and High (orange) tiers. Resolution Rate is plotted from 70%; duration and cost are plotted from zero. Exact values are listed on the next page.

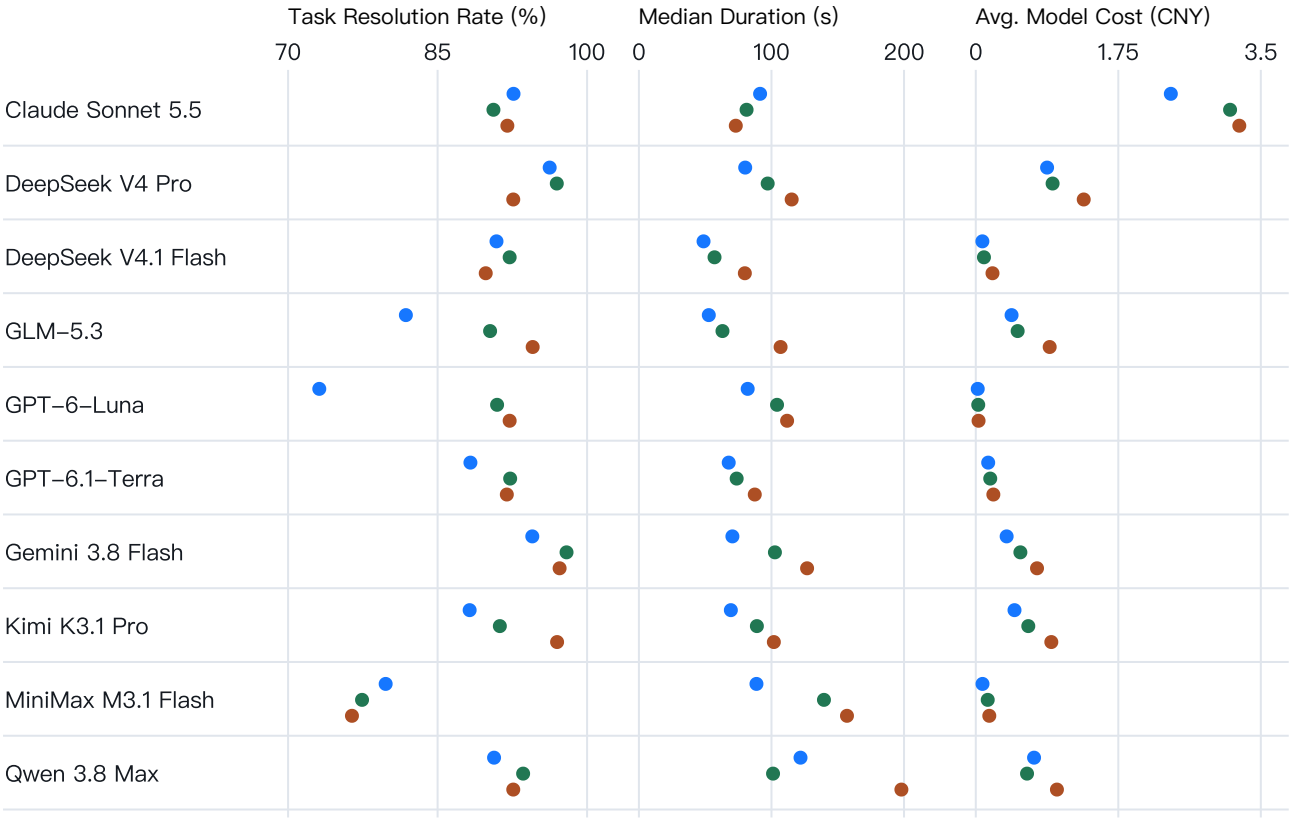


Figure 16 Completion Quality, Model Cost, and Duration Across 30 Configurations

The three metrics are measured independently. When selecting a configuration, first align on required completion quality, then compare duration and cost.

6. Execution Stability, Efficiency & Cost

Completion Quality & Execution Input – Numerical Comparison

Each row corresponds to a model and each column to its reasoning tier. Cells list Task Resolution Rate (%), Median Duration (s), and Avg. Model Cost (CNY) from left to right, enabling direct row–wise comparison across tiers. Duration covers model generation and intermediate tool operations; see Section 6.3 for start/end boundaries.

Model	Low: Res. Rate / s / CNY	Med: Res. Rate / s / CNY	High: Res. Rate / s / CNY
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

For Gemini 3.8 Flash (Medium), Task Resolution Rate is 97.94%, median duration is 102.60 s, and average cost is 0.5476 CNY. Supplementary Figure 22 ranks configurations separately by duration and cost.

6. Execution Stability, Efficiency & Cost

6.3 Measurement Definitions & Missing Data

Response duration is measured from the Runner–logged start of the evaluated model's first generation call to the end of its final generation call, including multi–turn processing, tool queries, data writes, and wait times. It excludes pre–call queuing/setup and post–response grading, verification, and environment restoration. Only valid runs with a final response and complete turn timestamps are included.

Model cost is calculated at official pricing using CNY cost records retained by the Runner, averaged exclusively over runs with recorded cost data.

Metric	Recorded Runs	Runs Missing Metric	Summary Statistic
Response Duration	8584	266	Median 86.55 s
Model Cost	8818	32	Average 0.7111 CNY

Total recorded model cost across valid runs is 6,270.12 CNY. Cost and duration each use their respective available samples, excluding missing records. This sum covers recorded evaluated–model costs only, excluding platform and judge–model costs.

7. Representative Business Task Case Studies

The following five case studies illustrate record lookup, financial aggregation, data maintenance, incomplete document creation, and cross-app lineage tracing. Diagrams reconstruct the business process from a single actual run, grouped by business stage rather than raw log screenshots, ending with benchmark verification. Cases illustrate task execution and grading mechanics rather than comparing models or estimating failure shares.

Each diagram depicts a single run; the page footer summarizes all 90 runs for the task across 30 configurations (3 trials each).

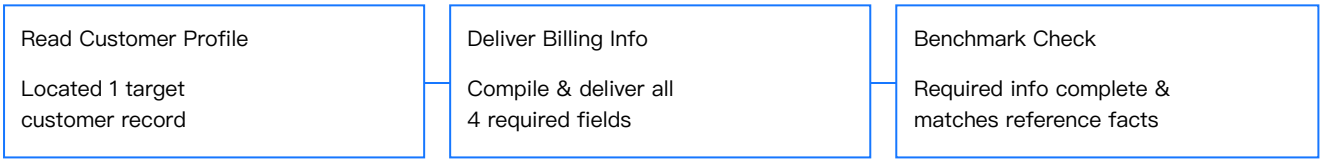
7.1 Query Customer Billing Information

Model: DeepSeek V4 Pro (Medium)

Business Requirement: Retrieve the billing name, tax ID, registered address, and registered phone for a specified customer for invoicing.

Customer Billing Info: All Required Information Delivered

Business Requirement: Retrieve billing name, tax ID, registered address, and phone for a target customer.
Model: DeepSeek V4 Pro (Medium)



Actual records grouped by business stage; the final step is benchmark verification, while prior steps reflect model processing.

Check Item	Task Requirement	Run Outcome
Billing Name & Tax ID	Matches specified customer	Delivered & verified accurate
Registered Address & Phone	Use invoicing-required details	Delivered & verified accurate
Information Completeness	All 4 required fields present	Requirement met

> Benchmark Verdict: PASS Required details are complete and consistent with reference business facts.

Business process diagram compiled from actual run logs and verification results.

Figure 17 Business Execution Process & Result Verification: Customer Billing Information

Verdict: PASS. All required details were delivered completely and matched reference facts, fulfilling the billing lookup requirement.

Aggregate results for this task across 30 configurations: 88 passes, 2 failures, 0 invalid runs; valid pass rate: 97.78%.

BATCH-0134 run_v3_a378a0e6ab77bcec16536869

7. Representative Business Task Case Studies

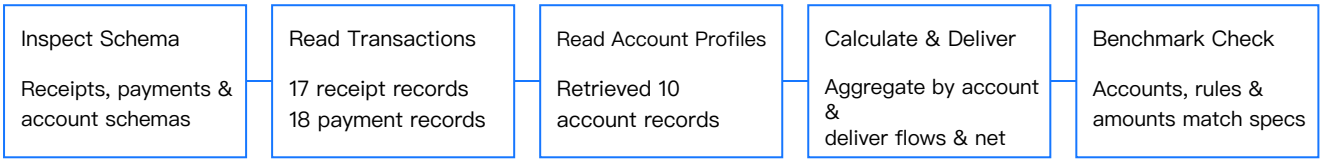
7.2 Calculate Monthly Account Net Cash Flow

Model: DeepSeek V4 Pro (Medium)

Business Requirement: Aggregate actual receipts and actual payments for each fund account in the specified month, and calculate net cash flow (receipts minus payments).

Account Net Cash Flow: Receipts & Payments Delivered per Business Rules

Business Requirement: Calculate receipts, payments, and net cash flow per account from actual monthly transactions.
Model: DeepSeek V4 Pro (Medium)



Actual records grouped by business stage; the final step is benchmark verification, while prior steps reflect model processing.

Check Item	Task Requirement	Run Outcome
Period & Payment Status	By actual date & confirmed status	Response rule check passed
Account Receipts & Payments	Deliver amounts by account	Delivered & verified accurate
Net Cash Flow	Computed as receipts minus payments	Delivered & verified accurate

> Benchmark Verdict: PASS Calculation rules, accounts, and amounts meet task requirements.

Business process diagram compiled from actual run logs and verification results.

Figure 18 Business Tool Call Trace & Verification: Monthly Account Net Cash Flow

Verdict: Pass. Required statistics for all accounts meet task criteria. The evaluation checks both calculation logic and results, not just whether a single net figure was given.

Aggregate results across 30 model configurations: 78 passed, 9 failed, 3 invalid; valid pass rate: 89.66%.

BATCH-0134 run_v3_327ebd4d4ca44b0f771d9e6b

7. Representative Business Task Case Studies

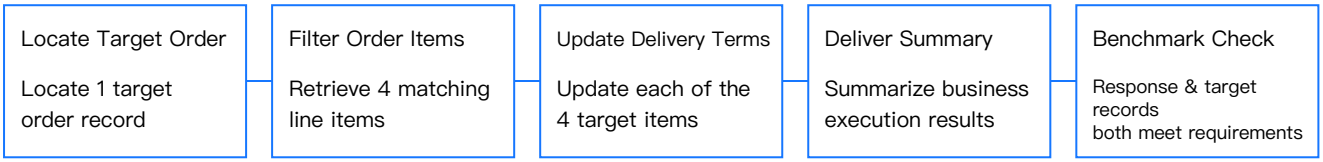
7.3 Updating Delivery Terms on Customer Orders

Executed Model: Qwen 3.8 Max (Low)

Business Requirement: Append delivery terms to price-qualified line items in a specified order while preserving required original content.

Order Delivery Terms: Target Line Items Updated as Required

Business Requirement: Append delivery terms to price-qualified line items in a specified order while preserving original content.
Model: Qwen 3.8 Max (Low)



Actual records grouped by business stage; the final step is benchmark verification, while prior steps reflect model processing.

Check Item	Task Requirement	Run Outcome
Order & Line Item Scope	Filter by target order & price criteria	Retrieved 4 matching line items
Target Line Item Content	Update delivery terms as required	Final record check passed
Completion Summary	Consistent with actual execution	Response check passed

Benchmark Verdict: PASS Both the completion summary and final target records meet requirements.

Business process diagram compiled from actual run logs and verification results.

Figure 19 Business Tool Call Trace & Verification: Updating Customer Order Delivery Terms

Verdict: Pass. Both the completion summary and the final state of target records meet task requirements.

Aggregate results across 30 model configurations: 83 passed, 7 failed, 0 invalid; valid pass rate: 92.22%.

BATCH-0145 run_v3_a71faa7d68dc2bf670272bf7

7. Representative Business Task Case Studies

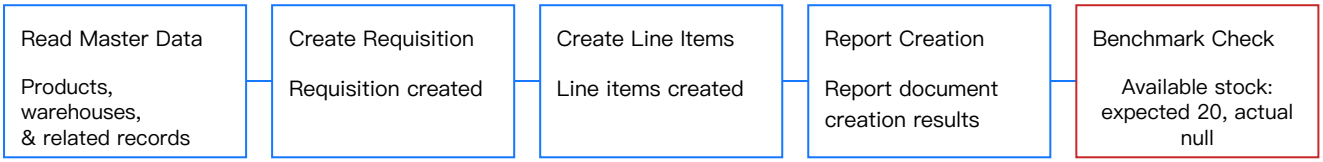
7.4 Creating a Draft Purchase Requisition

Executed Model: Qwen 3.8 Max (Medium)

Business Requirement: Create a draft purchase requisition with complete header and line items under current purchasing rules, without placing an order or receiving stock.

Purchase Requisition: Header Created, Line Item Info Incomplete

Business Requirement: Create a complete draft purchase requisition and line items without ordering or receiving.
Model: Qwen 3.8 Max (Medium)



Actual records grouped by business stage; the final step is benchmark verification, while prior steps reflect model processing.

Check Item	Task Requirement	Run Outcome
Purchase Requisition	Create compliant requisition header	Created, meets requirements
Requisition Line Items	Create and link line items	Records created
Line Item: Available Stock	Record stock value 20 as required	Actual is null, requirement unmet

Verdict: Fail Required stock value missing; final record check failed.

Business process diagram derived from actual execution and inspection results. Required stock values are based on benchmark inventory records.
Figure 20 Purchase Requisition: Business Execution vs. Result Verification

Verdict: Fail. Response check passed, but final record check failed: requisition header and items were created, yet the required available stock field was left empty. A successful creation call does not equate to full task resolution.

Aggregate results across 30 model configurations: 39 passed, 49 failed, 2 invalid; valid pass rate: 44.32%.

BATCH-0156 run_v3_41c132a9fdcb93040ce4cd26

7. Representative Business Task Case Studies

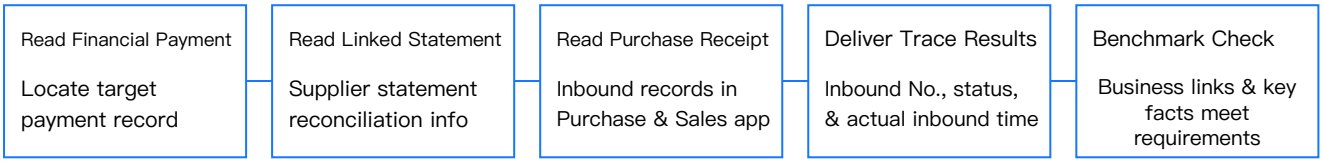
7.5 Tracing from Payment to Purchase Receipt

Model: DeepSeek V4 Pro (Medium)

Business Requirement: Trace a specified payment record to its corresponding purchase receipt and report the receipt number, inbound status, and actual inbound time.

Payment–to–Receipt Traceability: Cross–App Information Fully Delivered

Business Requirement: Trace a specified payment to its purchase receipt and report inbound No., status, and actual time.
Model: DeepSeek V4 Pro (Medium)



Actual records grouped by business stage; the final step is benchmark verification, while prior steps reflect model processing.

Check Item	Task Requirement	Run Outcome
Receipt Record & Business Link	Trace via payment and statement links	Delivered matching inbound No.
Inbound Status	Retrieved matching receipt status	Delivered & verified accurate
Actual Inbound Time	Used actual time from inbound record	Delivered & verified accurate

> Benchmark Verdict: PASS Both cross–app links and receipt details meet task requirements.

Business process diagram compiled from actual run logs and verification results.

Figure 21 Business Tool Call Trace & Verification: Tracing Payment to Purchase Receipt

Verdict: Pass. The delivered inbound number, status, and actual time meet inspection criteria; this case illustrates the execution trace across linked applications.

Aggregate results across 30 model configurations: 89 passed, 0 failed, 1 invalid; valid pass rate: 100.00%.

BATCH–0134 run_v3_1a7f453d18b6f7de6c60226b

8. Conclusions & Scope of Applicability

8.1 Key Takeaways

Gemini 3.8 Flash (Medium) ranks first by Task Resolution Rate (285/291, 97.94%); out of 300 Planned Runs per configuration, Kimi K3.1 Pro (High) achieved the most Valid Passes (290/300, 96.67%). Across the 52 inventory tasks, the aggregate pass rate is 90.02%, led by GPT–6.1–Terra – Low at 98.72%.

On the 38 common tasks where all 30 configurations obtained 3 Valid Runs (114 runs each), DeepSeek V4 Pro (Low) passed 113 runs (99.12% resolution rate). This control set checks the impact of sample differences on rankings.

Text responses, tool call returns, and final business states must be evaluated separately. In Section 7.4, the purchase requisition was created and the text response passed inspection, but the line item's current available stock was left empty instead of 20, failing the task. Execution traces help distinguish whether to refine response prompts, fix business constraints, or complete target records.

Repeatability and resource investment must be interpreted together. Execution consistency, latency, and model cost capture distinct dimensions; lower cost does not imply higher reliability, nor do reasoning tiers map to equal resource consumption across models. Configurations should be chosen by combining Chapters 4–6 results with actual business needs.

8.2 How to Apply These Results

This report supports initial configuration screening, execution bottleneck diagnosis, and joint quality–cost evaluation for enterprise business tasks. In practice, align operational goals and constraints first, then consult the matching domain, business type, and subgroup breakdowns.

Business Question	Supporting Evidence
Which configuration best fits a workload?	Check matching business type & domain first, then inspect specific tasks
Which business step failed?	Cross–check failure types, response assertions, and post–call database records
Can it execute reliably across repeats?	Check 3–of–3 pass counts and denominators in Table D6
What investment does quality require?	Jointly review Task Resolution Rate, Median Latency (s), and Avg. Model Cost (CNY)

8. Conclusions & Scope of Applicability

8.3 Coverage & Comparison Boundaries

Task-group outcomes do not translate into standalone capability scores. Differences in valid sample sizes, unverified initial environment equivalence, and judge model variance constrain interpretations of narrow leads or root failure causes. Items lacking full verification evidence are not scored separately.

Sample sizes vary by domain. The 9,000 runs derive from 100 tasks; repeated runs on the same task measure outcome variance.

8.4 Data Sources

Tasks originate from the BenchSaaS benchmark suite. Pass, fail, and invalid verdicts were retrieved from HAP evaluation run states on November 19, 2026; latency, model cost, and case traces were extracted from Runner logs for the corresponding run IDs.

Supplementary Charts

Separate Rankings by Response Latency and Avg. Model Cost

Rank	Model Configuration	Median Duration (s)	Model Configuration	Avg. Model Cost (CNY)
1	DeepSeek V4.1 Flash · Low	48.75	GPT-6-Luna · Low	0.0241
2	GLM-5.3 · Low	52.73	GPT-6-Luna · Med	0.0302
3	DeepSeek V4.1 Flash · Med	57.03	GPT-6-Luna · High	0.0341
4	GLM-5.3 · Med	63.04	DeepSeek V4.1 Flash · Low	0.0809
5	GPT-6.1-Terra · Low	67.75	MiniMax M3.1 Flash · Low	0.0825
6	Kimi K3.1 Pro · Low	69.35	DeepSeek V4.1 Flash · Med	0.1003
7	Gemini 3.8 Flash · Low	70.52	MiniMax M3.1 Flash · Med	0.1465
8	Claude Sonnet 5.5 · High	73.09	GPT-6.1-Terra · Low	0.1521
9	GPT-6.1-Terra · Med	73.75	MiniMax M3.1 Flash · High	0.1660
10	DeepSeek V4.1 Flash · High	79.91	GPT-6.1-Terra · Med	0.1775
11	DeepSeek V4 Pro · Low	80.18	DeepSeek V4.1 Flash · High	0.2052
12	Claude Sonnet 5.5 · Med	81.22	GPT-6.1-Terra · High	0.2157
13	GPT-6-Luna · Low	82.06	Gemini 3.8 Flash · Low	0.3789
14	GPT-6.1-Terra · High	87.40	GLM-5.3 · Low	0.4393
15	MiniMax M3.1 Flash · Low	88.67	Kimi K3.1 Pro · Low	0.4754
16	Kimi K3.1 Pro · Med	88.99	GLM-5.3 · Med	0.5141
17	Claude Sonnet 5.5 · Low	91.42	Gemini 3.8 Flash · Med	0.5476
18	DeepSeek V4 Pro · Med	97.14	Qwen 3.8 Max · Med	0.6293
19	Qwen 3.8 Max · Med	101.19	Kimi K3.1 Pro · Med	0.6447
20	Kimi K3.1 Pro · High	101.78	Qwen 3.8 Max · Low	0.7155
21	Gemini 3.8 Flash · Med	102.60	Gemini 3.8 Flash · High	0.7519
22	GPT-6-Luna · Med	104.19	DeepSeek V4 Pro · Low	0.8756
23	GLM-5.3 · High	106.88	GLM-5.3 · High	0.9074
24	GPT-6-Luna · High	111.78	Kimi K3.1 Pro · High	0.9273
25	DeepSeek V4 Pro · High	115.23	DeepSeek V4 Pro · Med	0.9430
26	Qwen 3.8 Max · Low	121.91	Qwen 3.8 Max · High	0.9964
27	Gemini 3.8 Flash · High	126.87	DeepSeek V4 Pro · High	1.3254
28	MiniMax M3.1 Flash · Med	139.59	Claude Sonnet 5.5 · Low	2.3949
29	MiniMax M3.1 Flash · High	156.97	Claude Sonnet 5.5 · Med	3.1224
30	Qwen 3.8 Max · High	198.01	Claude Sonnet 5.5 · High	3.2358

Figure 22 Separate Rankings by Response Latency and Avg. Model Cost

明道云
mingdao.com

Copyright & Partner Invitation

Authorship & Copyright

Author: Mingdao Cloud AI Engineering Team

(c) 2026 Mingdao Cloud. All rights reserved.

Any reproduction or citation of text, charts, or evaluation results from this report must credit the source: "Mingdao Cloud AI Engineering Team, Enterprise Application LLM Benchmark Report (BenchSaaS, V1.0)".

Call for Partners

1. Model Service Provider Partnership

We welcome model service providers to sponsor token resources in support of ongoing BenchSaaS evaluations for enterprise applications, expanding model and task coverage.

2. Model Developer Partnership

We welcome model developers to partner with us for access to the public benchmark suite and diagnostic reports, offering insights into model performance and failure modes on enterprise application tasks to guide capability improvements.

Official Website

Visit the official project website to explore the interactive leaderboard, download the latest report, and submit partnership inquiries:

mingdao.com/landing/benchsaas