BankerToolBench: Evaluating AI Agents in End-to-End Investment Banking Workflows

Elaine Lau[1]*, **Markus Dücker**[1], **Ronak Chaudhary**[1]
Hui Wen Goh[1], **Rosemary Wei**[1], **Vaibhav Kumar**[1]
Saed Qunbar[1], **Guram Gogia**[1], **Yi Liu**[1]
Scott Millslagle[1], **Nasim Borazjanizadeh**[1], **Ulyana Tkachenko**[1]
Samuel Eshun Danquah[1], **Collin Schweiker**[1]*, **Vijay Karumathil**[1]
Asrith Devalaraju[1], **Andrew Peter Martin**[1], **Varsha Sandadi**[1]
Haemi Nam[1], **Punit Arani**[1], **Ray Epps**[1]
Abdullah Arif[1], **Sahil Bhaiwala**[1], **Curtis Northcutt**[1]
Skyler Wang[1,2], **Anish Athalye**[1], **Jonas Mueller**[1]*, **Francisco Guzmán**[1]

[1]Handshake AI [2]McGill University

NeurIPS 2026 · Evaluations and Datasets

Handshake AI

NEURAL INFORMATION PROCESSING SYSTEMS

A banking request becomes a working model

Task

Prompt

You are a healthcare investment banking analyst at xyz bank and you have been tasked with creating an outside in model for Capricor Therapeutics to figure out their valuation through a DCF / SOTP model.

You have finished your revenue model for capricor's lead asset in its lead indication (de-ramiocel in DMD). Next you will build the DCF…

Data Room

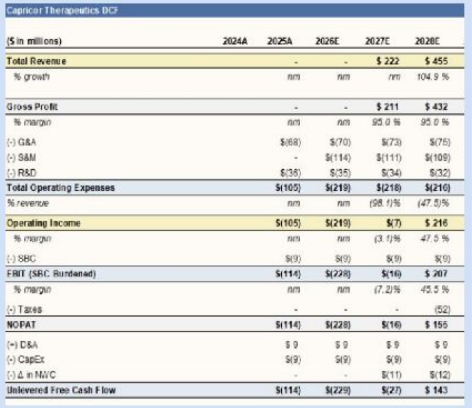
- Provided Files (report, spreadsheet, image)
- APIs to market data platform, SEC EDGAR

LLM Powered Agent →

Expected Output

Deliverable

- FCF loader Excel tab
- P&L Excel tab
- DCF Excel tab



Capricor Therapeutics DCF

($ in millions)	2024A	2025A	2026E	2027E	2028E
Total Revenue		-	-	$ 222	$ 455
% growth		nm	nm	nm	104.9 %
Gross Profit		-	-	$ 211	$ 432
% margin		nm	nm	95.0 %	95.0 %
(-) G&A		$(68)	$(70)	$(73)	$(76)
(-) S&M		-	$(114)	$(111)	$(109)
(-) R&D		$(36)	$(35)	$(34)	$(32)
Total Operating Expenses		$(105)	$(219)	$(218)	$(216)
% revenue		nm	nm	(98.1)%	(47.5)%
Operating Income		$(105)	$(219)	$(7)	$ 216
% margin		nm	nm	(3.1)%	47.5 %
(-) SBC		$(9)	$(9)	$(9)	$(9)
EBIT (SBC Burdened)		$(114)	$(228)	$(16)	$ 207
% margin		nm	nm	(7.2)%	45.5 %
(-) Taxes		-	-	-	(52)
NOPAT		$(114)	$(228)	$(16)	$ 155
(+) D&A		$ 9	$ 9	$ 9	$ 9
(-) CapEx		$(9)	$(9)	$(9)	$(9)
(-) Δ in NWC		-	-	$(11)	$(12)
Unlevered Free Cash Flow		$(114)	$(229)	$(27)	$ 143

Graded by Verifier →

Rubric

Criterion 1

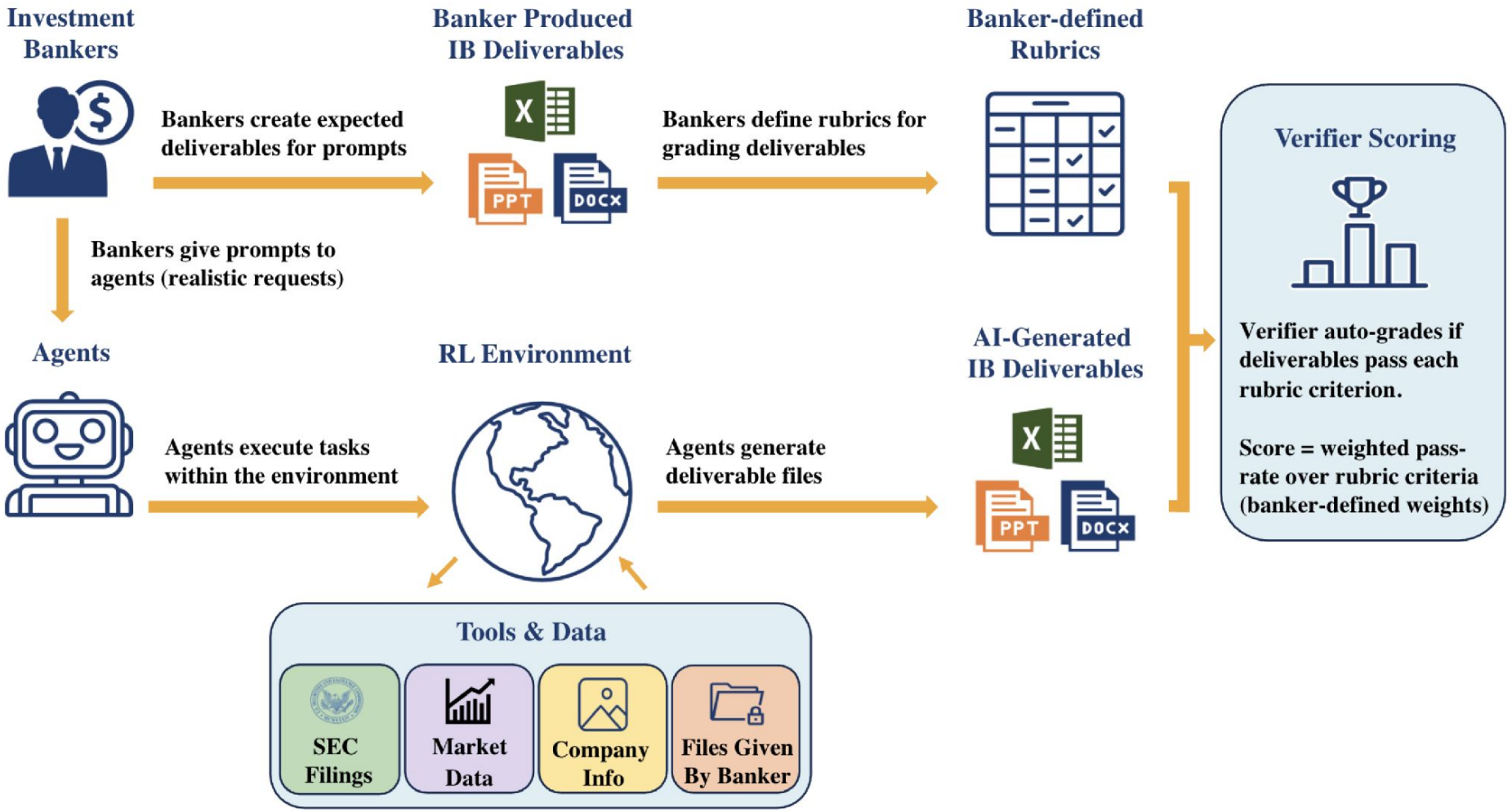
Instruction Following
Active loader scenario box/switch is present on Loader tab allowing user to toggle …
Weight: 10

Criterion 2

Model Methodology & Financial Accuracy
In P&L tab: All revenue components are linked to Loader tab (not hardcoded)
Weight: 10

⋮

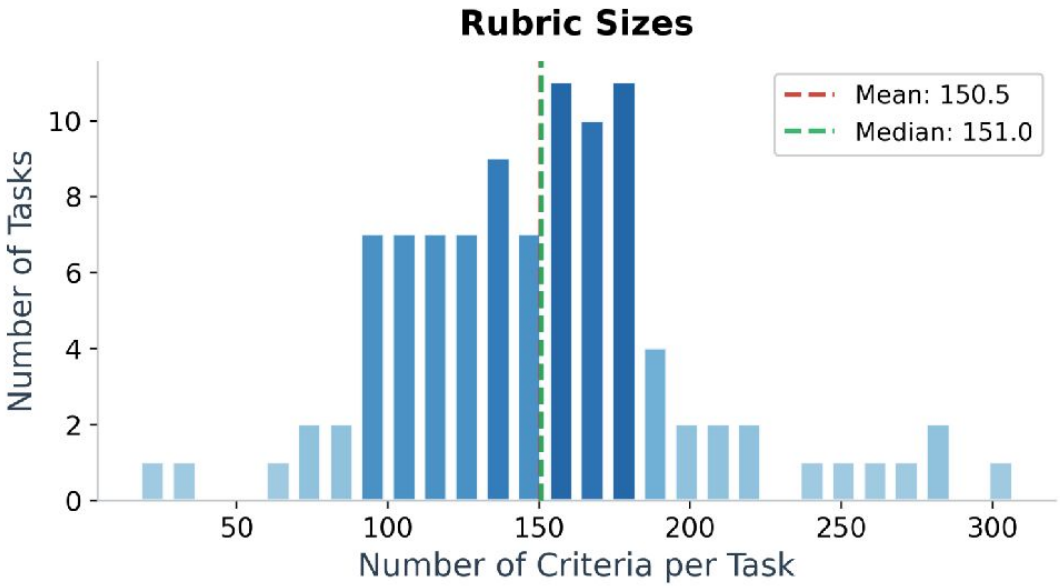
BankerToolBench evaluates the whole workflow

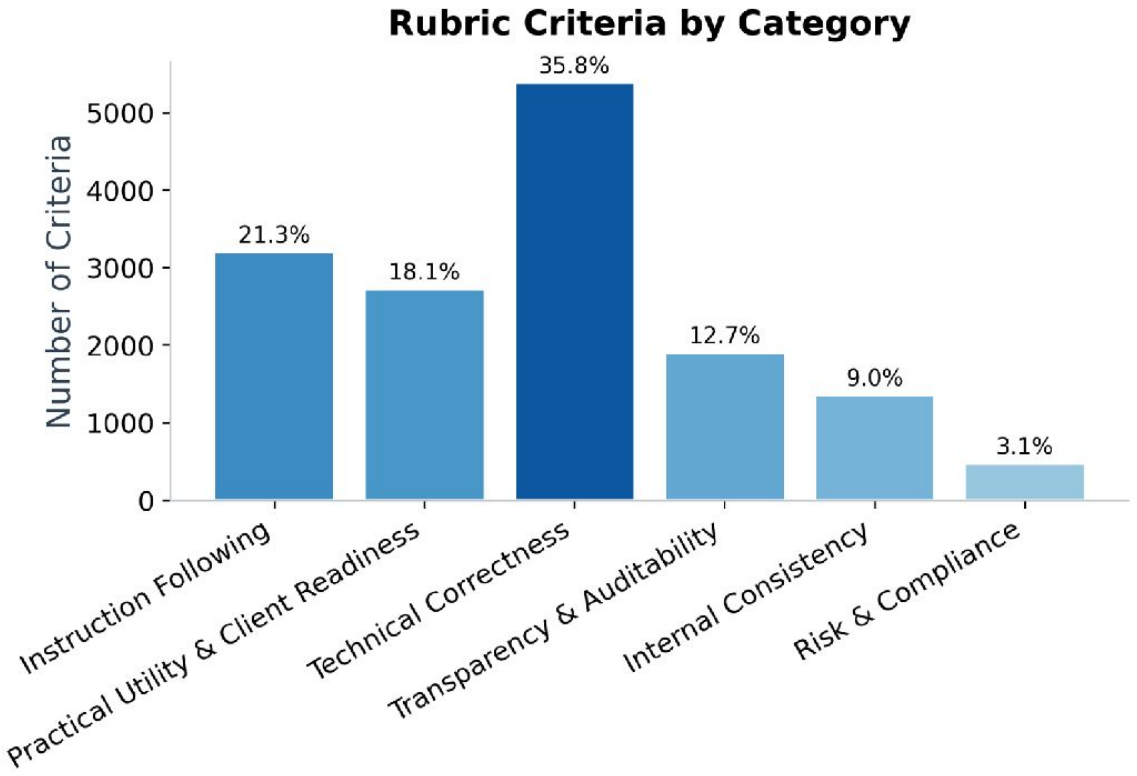


100 tasks, 502 banker contributors, 5 hours average human effort

Expert rubrics score the deliverables

Bankers define task-specific pass/fail checks, weighted by importance.





Verifier: Gandalf, an agent-as-judge grader for files

CRITERION “CAPEX is linked from the Mar'25 fiscal year end key items source file.”

1

Locate CapEx in deliverable

```
ws = wb['Operating Model']
for r in range(22, 28):
  print(r, ws.cell(r, 1).value,
        ws.cell(r, 2).value)
```

→ **Found B25 = 18.52**

ELF LBO Model.xlsx · Operating Model

	A	B
24	EBIAT	–
25	**(–) CapEx**	**18.52**
26	% of Revenue (1)	–

2

Read the footnote claim

```
ws['A78'].value

# footnote reference
# on the CapEx row
```

→ **“(1) CapEx based on Mar'25…”**

ELF LBO Model.xlsx · A78

AGENT'S FOOTNOTE CLAIM

“CapEx based on Mar'25 fiscal year end…”

↓ **Unverified — fetch the source**

3

Fetch the source file

```
vdr.pull(
  symbol="ELF",
  data_types="cashflow_annual"
)

# MCP tool, same one
# the rollout agent had
```

→ **ELF_Statement.xlsx**

MCP TOOL · VDR

vdr.pull ➡ **VDR FILE** **ELF_Statement .xlsx**

Same tool the agent used

4

Compare against Mar'25

```
for row in source.rows:
  if row.label == \
    'Capital Expenditures':
    print(row['MAR 25'])
```

→ **−18.52 (Mar'25 column)**

ELF_Statement.xlsx · VDR source

	MAR ’25	MAR ’24
Sales	1,313.5	1,023.9
EBITDA	202.5	183.0
CapEx	**−18.52**	**−8.66**

5

Verdict

✓ **MET**

Cross-file link confirmed.

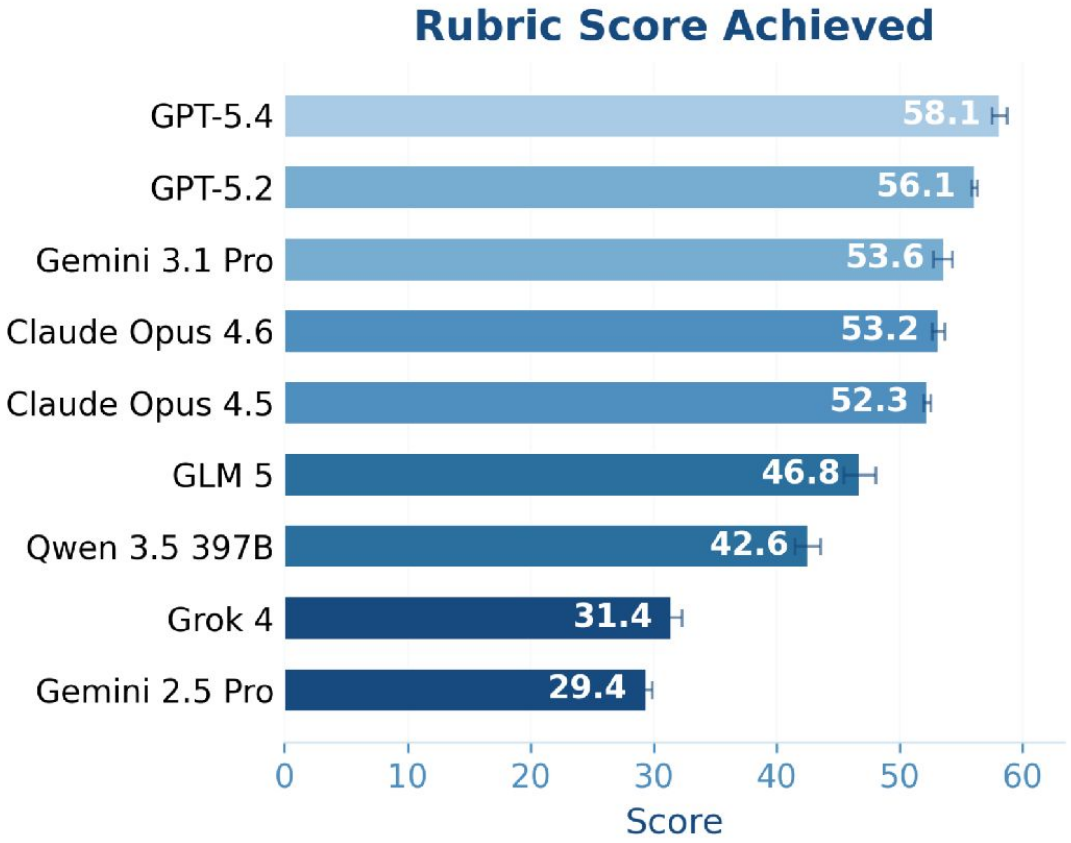
Deliverable B25 traces back to VDR Mar'25 cash flow data via the rollout agent's MCP tool.

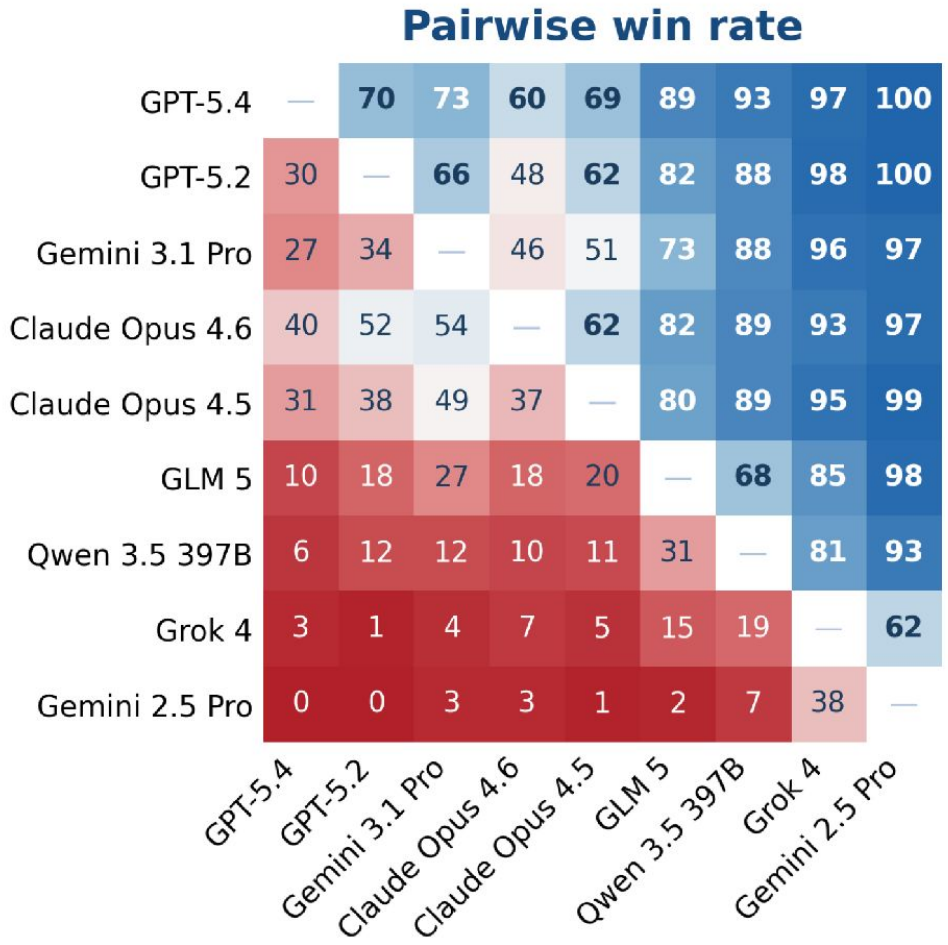
EVIDENCE TRAIL

- B25 (deliverable) = 18.52
- A78 footnote cites Mar'25
- VDR MAR ’25 col = −18.52
- |18.52| = |−18.52| ✓

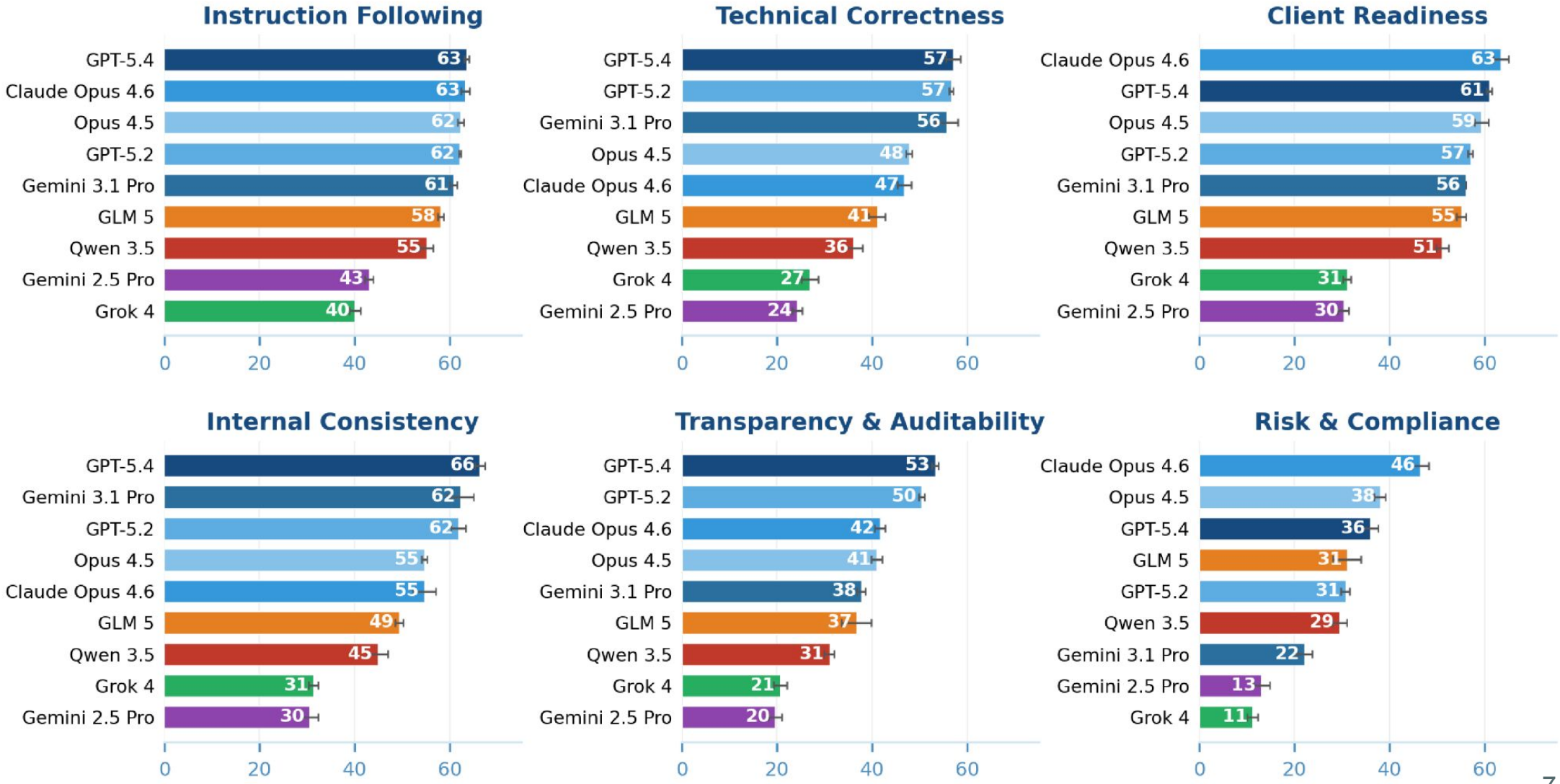
https://github.com/handshake-AI-Research/gandalf-the-grader

The best model scores 58.1 out of 100

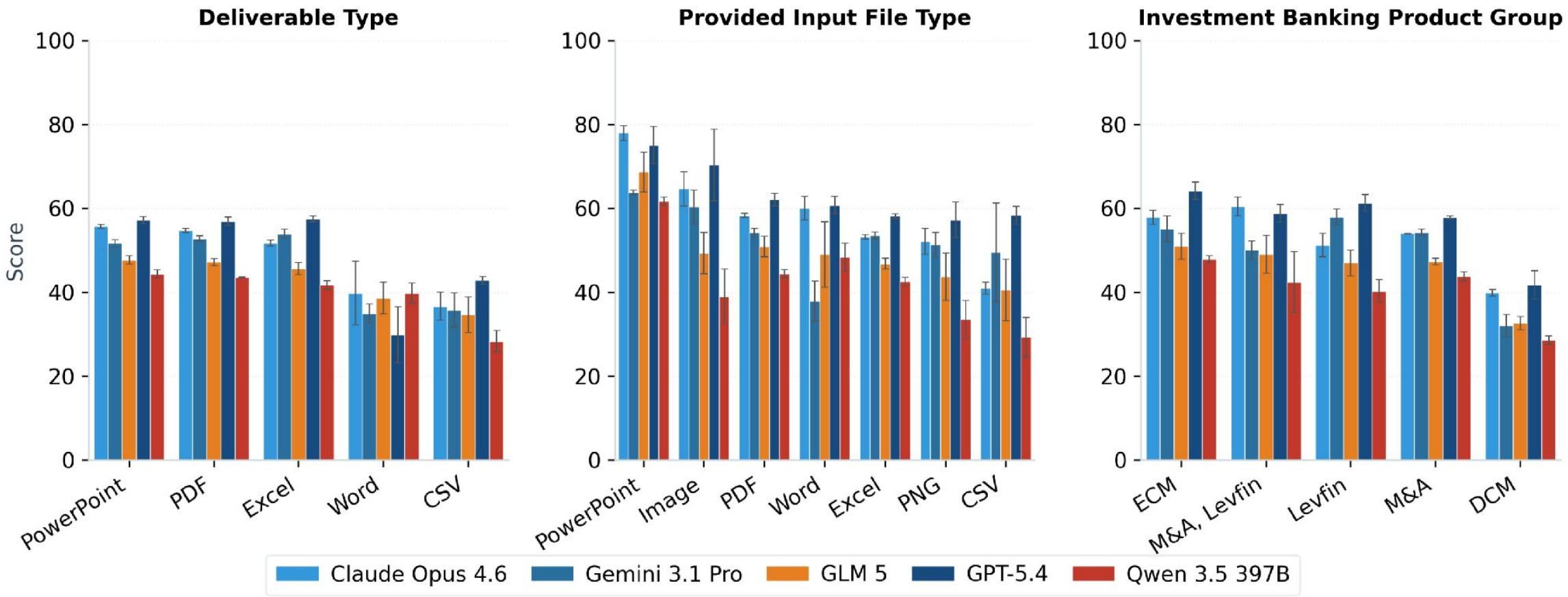




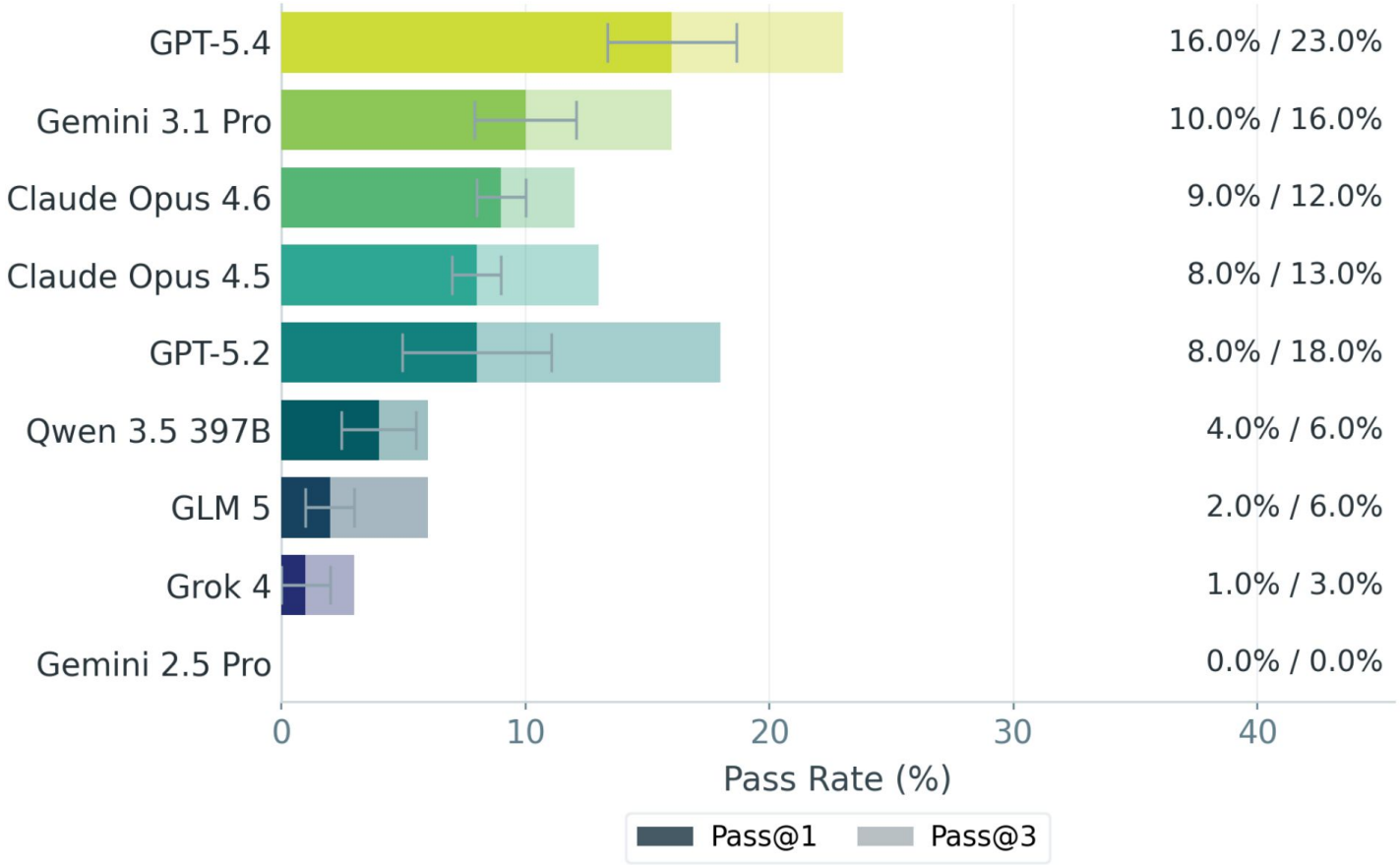
Model strengths vary across all six categories



Performance varies across banking workflows



Higher scores still leave a readiness gap



0%

rated sendable as-is

Human review of GPT-5.4 and Claude Opus 4.6

Errors in logic and consistency break the work

Cost savings enter the wrong financial line

Failing Rubric Criterion: Synergy assumptions must flow through pro forma correctly.
Verifier Rationale: Agent added $75M cost synergies to revenue line instead of reducing operating expenses. This inflates combined revenue by an amount that represents cost savings, which is categorically incorrect.

The same metric disagrees within the pitchbook

Failing Rubric Criterion: LTM Revenue in Key Stats quad matches LTM Revenue in Financial Statistics table.
Verifier Reasoning: LTM Revenue in Key Stats quad ($189.5bn) does not match the Revenue in the Financial Statistics table ($201.0bn for 2025A).

BTB measures complete professional work

Realistic workflows expose gaps in financial logic, file integrity, and consistency.

Open-source tasks, tools, and rubrics for agent evaluation and training.

Code github.com/Handshake-AI-Research/bankertoolbench

Dataset huggingface.co/datasets/handshake-ai-research/bankertoolbench