University of St.Gallen

FINLLM @ IJCAI 2026 · BREMEN · 15 AUGUST 2026

Can Open-Weight Models Compete on **Financial Text Comprehension?**

Benchmarking 20 frontier & open-weight LLMs on long-context annual-report QA

Jan Spörer · University of St. Gallen, Switzerland · jan.spoerer@whu.edu

From insight to impact.

Motivation & Research Question

- Open-weight LLMs, especially from Chinese labs (**GLM 5, Qwen3, DeepSeek, Kimi K2.6**), caught up to proprietary frontier models within months. Their **reliability on real financial tasks is largely untested.**
- Annual reports are a demanding, high-value test bed:
  - **Economically consequential**: markets and analysts act on them.
  - **Long & heterogeneous**: many exceed 700 pages (~165k tokens on average).
  - **Open-domain**: American, European, Chinese & Indian filings, many jurisdictions.
- The **2026 contamination problem:** public benchmarks leak into training corpora. FinancialTouchstone is kept **private until publication** → results stay on genuinely unseen data.

Research question: Can open-weight models match proprietary frontier models on long-context financial document comprehension?

Data: FinancialTouchstone v1.2

- **495 annual reports** from publicly traded companies · **22 countries · 20 stock exchanges** · all GICS industry classes.
- **2,967** manually annotated **question-context-answer triplets** · **> 83 million tokens** · reporting years 2021 to 2023.
- **Six analyst-style question types:** key financials, cash flow, revenue, revenue growth, business segments, company type.
- **Human baseline** (dual annotation): **82.8% accuracy**, **2.8% hallucination**: a best-case upper bound, not typical practice.
- v1.2 adds 89 questions & 15 reports over v1.1, with a two-round quality re-check (67 answers & 15 hallucinations revised).

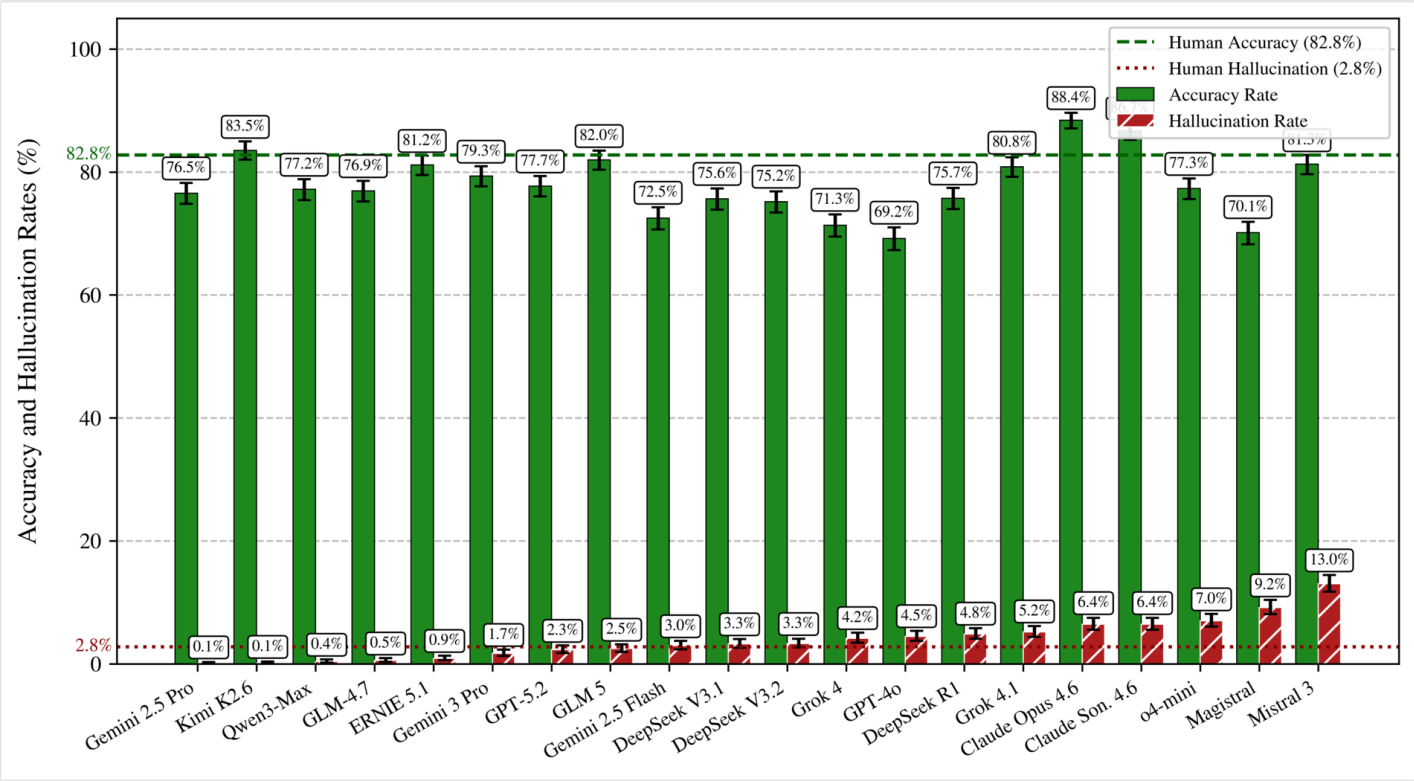
495	2,967	22	83 M+
annual reports	Q-C-A triplets	countries · 20 exchanges	tokens of financial text

Methodology

- **20 models across 10 providers** (up from 11): 16 reasoning + 4 non-reasoning:
  - **Proprietary:** Gemini 2.5 Pro / 2.5 Flash / 3 Pro, GPT-5.2, o4-mini, GPT-4o, Claude Opus 4.6 / Sonnet 4.6, Grok 4 / 4.1.
  - **Open-weight (focus):** DeepSeek R1 / V3.1 / V3.2, Zhipu GLM-4.7 / GLM 5, Qwen3-Max, Kimi K2.6, ERNIE 5.1, Mistral 3 / Magistral.
- **Deliberately simple, fixed RAG**: `1000`-token chunks (200 overlap), `text-embedding-3-small`, FAISS, **top-5**.
- **Same retrieved context for every model** → variation in accuracy is attributable to the model, not the retriever. Retrieval is a fixed upstream condition, not a confounder.
- **LLM-as-judge** grading (Claude Opus 4.6; the four last-evaluated models graded by Opus 4.8): per-question-type correctness thresholds and an explicit **incorrect vs. hallucinated** distinction (rounding / unit tolerances).

Results · Accuracy & Hallucination

- **Claude Opus 4.6 tops accuracy: 88.4%**, then Sonnet 4.6 (86.7%) and **Kimi K2.6 (83.5%)**. **The top three all beat the 82.8% human baseline.**
- **Gemini 2.5 Pro: lowest hallucination, 0.08%**. Below any competitor and below the 2.8% human rate. **But only 13th in accuracy (76.6%).**
- **Kimi K2.6 = best all-round:** 3rd in accuracy *and* 0.13% hallucination. The strongest accuracy + groundedness combination in the field.



Accuracy (green) & hallucination (red) with 95% CIs, excluding retriever errors. Ranked by hallucination rate. Dashed lines = human baselines.

Results · Open-Weight & Non-Reasoning Closed the Gap

- **Three of the top five are open-weight**: Kimi K2.6 (3rd), GLM 5 (4th), Mistral 3 (5th), outranking GPT-5.2, Gemini 3 Pro, o4-mini and DeepSeek R1.
- **Reasoning is no longer a prerequisite.** Mistral 3 is **non-reasoning** yet beats many reasoning models outside the top four.
- The reverse also holds: **Magistral** (Mistral's reasoning model) ranks **19th, 14 places below** its non-reasoning sibling Mistral 3; Grok 4 is 18th.

#	MODEL	ACC.	HALL.	TYPE
1	Claude Opus 4.6	**88.4%**	6.4%	closed
2	Claude Sonnet 4.6	**86.7%**	6.4%	closed
3	Kimi K2.6 · Moonshot	**83.5%**	0.13%	open
	human baseline 82.8% accuracy			
4	GLM 5 · Zhipu	**82.0%**	2.5%	open non-reas.
5	Mistral 3	**81.3%**	13.0%	open non-reas.
6	Grok 4.1	**80.9%**	5.2%	closed
13	Gemini 2.5 Pro · lowest hall.	**76.6%**	0.08%	closed

Accuracy ranking (retriever errors excluded). Open-weight rows highlighted.

Results · Content-Filter Refusals: a New Deployment Risk

- **18 refusals persisted in the benchmark data (0.08%)**. All from **Chinese-provider models**. **No non-Chinese model refused any question.**
- Triggers = politically sensitive content in the *retrieved report*: Chinese leaders (a Kweichow Moutai report quotes Xi Jinping → "Content Exists Risk"), a state-owned bank, and Hong Kong protest mentions (even a US 10-K: Ralph Lauren).
- **Refusals are route-dependent, not just model-dependent:** Kimi K2.6 via the native Moonshot API refused 20 questions as "high risk". **The identical requests through a proxy were all answered.**

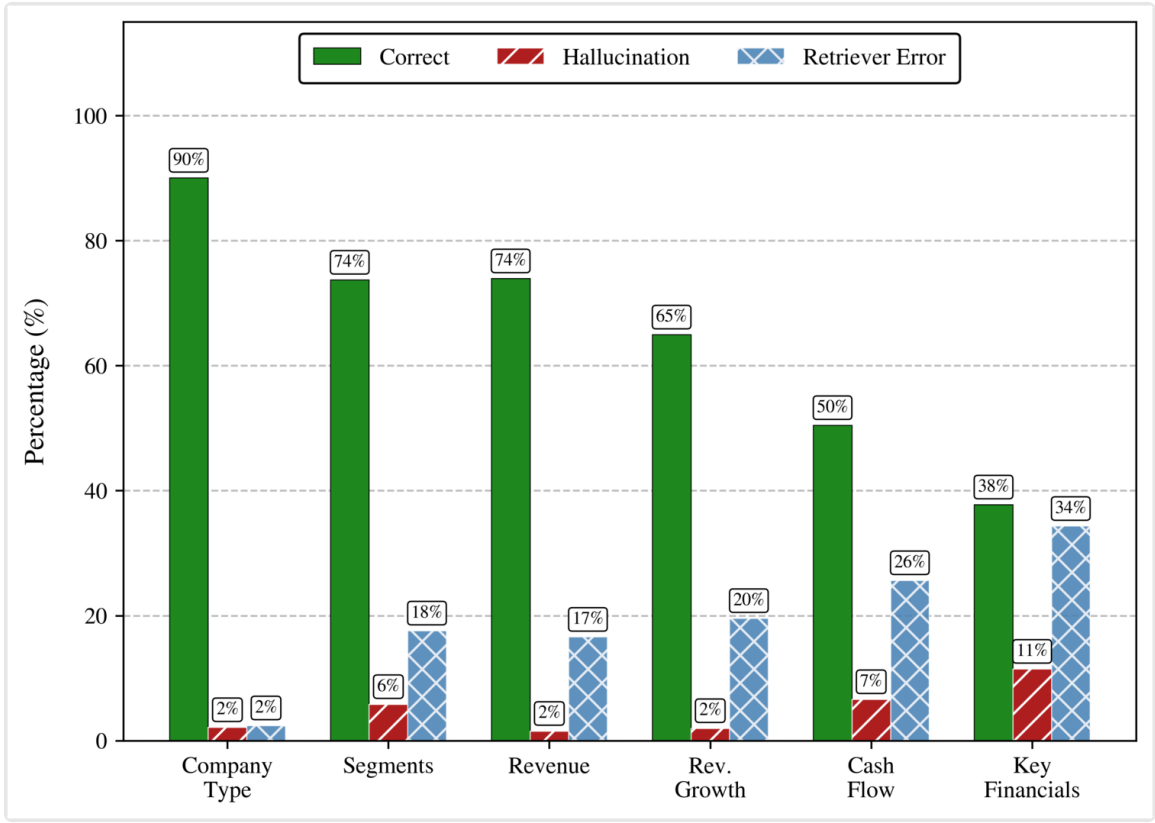
New for benchmark design: Refusals must be attributed to **model-route pairs**, not models alone. A proxy can mask a vendor's content filter.

Model	Route	Ref.	Companies
Kimi K2.6	Moonshot	20	Kw. Moutai, Nintendo, Ind. Bank, R. Lauren
Kimi K2.6	NanoGPT*	0	same 20 Q: all answered
DeepSeek R1	DeepSeek	5	Kw. Moutai (CN)
DeepSeek V3.1	DeepSeek	5	Kw. Moutai (CN)
DeepSeek V3.2	DeepSeek	5	Kw. Moutai (CN)
GLM 5	Zhipu	3	R. Lauren (US), PTT Public (TH)
GLM-4.7	OpenRouter*	0	none
Qwen3-Max	OpenRouter*	0	none
ERNIE 5.1	NanoGPT*	0	none

18 persisted / 23,736 attempts (0.08%). * = third-party proxy; zeros on proxy routes are route-conditional.

Results · Retrieval Bottleneck & Contamination Controls

- **Retrieval = 48.9% of all failures** (vs. 44.7% model error, 9.1% standalone hallucination). It hurts most on the hardest questions: key financials & cash flow.
- **Contamination controls** confirm models **read, not recall:**
  - Retrieval dependence: accuracy **77.9%** → **12.0%** when retrieval fails: a > 6× drop.
  - Temporal robustness: 76.8 / 77.4 / 81.0% for 2021 / 22 / 23: **no decline** on older reports.
- → Rankings reflect **comprehension, not memorization**; the retriever is the remaining lever.



Correct (green), model error (red), retriever error (blue) by question type: retriever error dominates the hardest questions.

Conclusion & Outlook

- **Claude, not Gemini, now tops accuracy**: Opus 4.6 (88.4%) & Sonnet 4.6 (86.7%) surpass the human baseline; open-weight **Kimi K2.6 takes 3rd (83.5%).**
- **Open-weight models have closed the gap**: Kimi K2.6 (3rd), GLM 5 (4th) & Mistral 3 (5th) crack the top five; the non-reasoning Mistral 3 outranks every reasoning model but the top four, while its reasoning sibling Magistral falls to 19th.
- **Content filters are a new, measurable deployment risk**: Chinese models only (0.08%), absent from all Western models, and **route-dependent.**
- **Retrieval is still the bottleneck** (48.9% of errors; 77.9% → 12.0% when it fails): the clearest lever for the next gain.
- **Outlook:** cross-lingual filings · broader question set (169 analyst archetypes) · GraphRAG · **ensembling** (Claude Opus 4.6 + Gemini 2.5 Pro + GLM 5).

Full dataset & evaluation framework are public → **financial-touchstone.datascience-nlp.ai** · **jan.spoerer@whu.edu**