RESEARCH BRIEF · EXPLORATORY PROJECT

BizHallu: Auditing Evidence Binding Errors in LLM-Generated Business Analysis

Yuchi Wang · MS student, Business Analytics and Artificial Intelligence
Johns Hopkins Carey Business School · Accounting and supply-management background

Research problem

Can we distinguish correctly copied data from a correct business relationship? BizHallu audits generated retail claims against transaction evidence. The current contribution is an inspectable workflow and a retrospective evaluation audit; an independent relation verifier remains proposed.

> **A concrete example: April 2011**
>
> Qwen assigns WOODEN UNION JACK BUNTING to rank 3 at GBP 4,173.18. The product and amount match source row 2, but the product ranks 7 in the eight shown rows. Rank 3 belongs to PAPER CHAIN KIT EMPIRE at GBP 6,619.51. Inspect q_0064.
>
> Curated evidence check; not an independent verifier prediction or a population error rate.

Dataset and method

UCI Online Retail; 100 questions across seven types; local Qwen3-0.6B answers; 205 AI-assisted provisional spans from 35 of 36 dev/test answers. Fifteen selected spans received additional assistant review. Independent human annotation and agreement remain pending.

Transaction evidence → questions → answers → pre-identified spans → token alignment → detector scores. Automatic extraction from new responses and whole-answer accuracy are outside the current evaluation.

Evidence status. Labels come from an outcome-informed, error-enriched queue. Dev/test share periods and exact evidence-row payloads. Thresholds were fitted on dev, but headline signals were selected after test comparison: exploratory, not confirmatory.

Research question and immediate pilot

How should complete entity-rank-amount relationships be annotated and checked without counting one binding error several times?

Specific request: feedback on one relation schema and one worked case, to define a small calibration exercise with a second reviewer. No independent review is yet complete.

A proposed evidence-aware verifier would use the question, answer, metric contract and evidence, excluding gold answers and evaluation labels from prediction. Report extraction coverage, abstention and errors separately. The current review schema is label-derived, not that verifier.

The design must account for shared contexts and different information available to token-time uncertainty and completed-answer checks. Early list-marker scores do not establish confidence in the later relationship.

Longer-term study design

The 48-context / 96-question plan is estimation-focused, design-only and not execution-ready. Three of seven execution gates are complete. Confirmation remains sealed. Protocol and remaining gates.

Historical B1 results

103 pre-identified test spans from 18 questions; provisional labels.

Reference / signal	Test AP	Test F1
Top-2 margin	0.835	0.752
Token entropy	0.809	0.779
Flag every span	0.592	0.744

Entropy minus flag-every-span F1: +0.0355; exploratory paired question-bootstrap 95% interval [-0.0356, 0.1047] crosses zero. Shared periods and test-based selection limit inference.

AP is tied-score-aware. AP/F1 maxima come from different signals. The dev fact-type prior (F1 0.836) is a composition control: supplied categories can reveal correctness.

September 5 update. B2 sensitivity adds 9 assistant-provisional q_0048 dev atoms. Dev-only refitting changes entropy F1 from 0.779 to 0.752 on the same 103 old test spans; top-2 margin is unchanged. Original B1 results remain intact. This is not fresh held-out evidence or independent review. B2 methods appendix.

Full methods, controls and source replay

Related work and intended distinction

- **FActScore** (Min et al., EMNLP 2023). Atomic factual precision motivates explicit units; business relationships also require grouping related facts.
- **TabFact** (Chen et al., ICLR 2020). Table-based fact verification motivates structured evidence; this proposal targets transaction scope and generated rankings.
- **Semantic Entropy** (Farquhar et al., Nature 2024). Meaning-level uncertainty requires multiple generations; the current study evaluates token entropy.

Related work only; none is an evaluated BizHallu baseline.

Contribution and provenance. Yuchi Wang directed the project with AI-assisted implementation and review across the workflow, provisional annotation, analysis and presentation. Independent human validation remains pending.

Business scope. Negative transaction value is not verified physical returns; merchandise uses a code-based heuristic. No production readiness, realized savings or inventory optimization is established.

Research brief revised September 6, 2026 · Author profile · Reproducibility guide · yuchi-wang02.github.io/bizhallu