

Building an Auditable Mutual Fund Ground-Truth Benchmark

A 20-Fund Study of Source-Grounded LLM Extraction, Human Review, and Abstention

Chang Liu and Xinrong Hao | IEOR E4737 | Columbia University | In collaboration with Ask2AI

Abstract

Mutual fund facts are dispersed across fact sheets, prospectuses, holdings schedules, and shareholder reports. Labels, periods, share classes, and units vary, so a plausible number can still be wrong. We built a stable 92-field template, used an LLM to locate and normalize candidates, and required human review to establish Ground Truth (GT). Across 20 funds, 1,280 of 1,840 targets satisfied a strict Direct Disclosure Only standard.

Core question Can LLM assistance accelerate extraction without sacrificing context control or inventing values when the documents are silent?

Research Objectives

Stable schema: Preserve a common structure even when disclosure is incomplete.

Auditable evidence: Separate direct facts, transparent derivations, and legitimate absence.

Reproducible scoring: Turn review failures into explicit extraction and evaluation rules.

Methodology

Benchmark: Five modules, 92 targets per fund, and 20 heterogeneous funds.

Source mapping: Each document type was assigned a distinct evidentiary role.

LLM assistance: Candidate values were located, normalized, and drafted into the approved schema.

Human validation: Each target was checked for fund, class, period, definition, unit, and provenance.

Evaluation: Direct-only coverage and Ask2AI Portfolioi agreement were measured against human-verified GT.

Benchmark Design

The schema deliberately retains inconsistently disclosed fields, making absence an explicit abstention test.

Template module	Targets
Current fund information	22
Performance	12
Quarter-end snapshot	18
Operations	20
FY financial highlights	20

DIRECT = explicitly disclosed and context-matched; *DERIVED* = auditable calculation; *UNAVAILABLE* = not disclosed, not applicable, or unsupported. *N/A* never means zero.

Key Ground Truth Findings

Overall coverage: 1,280 of 1,840 targets were Direct (69.6%).

Batch dispersion: Xinrong Hao's batch reached 80.0%; Chang Liu's reached 59.1% after the final evidence-status audit.

Availability pattern: Performance and current fund information were strongest; quarter-end snapshots were the main abstention stress test.

Human-AI Workflow

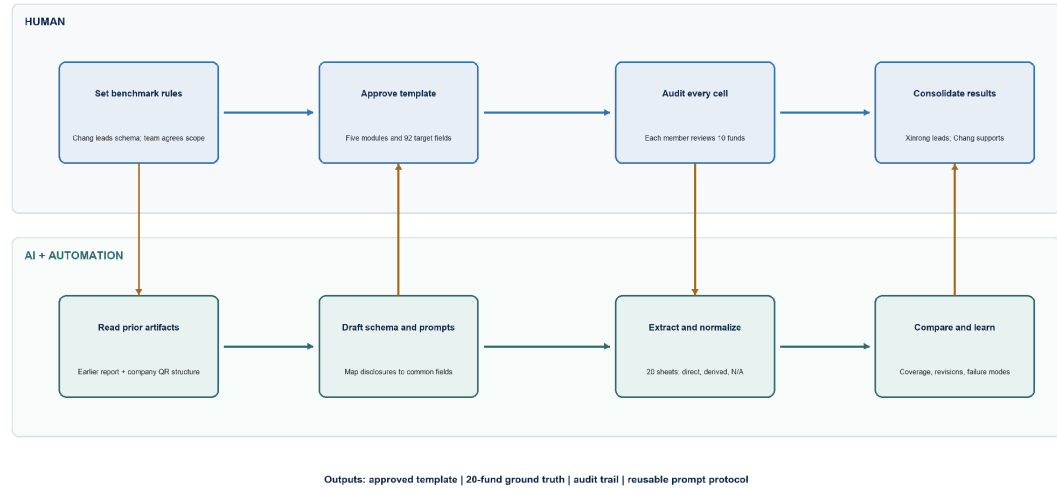


Figure 1. AI supports document comparison, candidate-field location, normalization, and workbook drafting. Humans define the benchmark, approve evidence, audit every cell, and control the final GT decision.

Evidence Protocol

Context lock: Confirm fund, ticker, share class, period, and unit before extraction.

Semantic match: Verify table title, row, column, footnotes, and definition.

Evidence status: Classify every result as Direct, Derived, or Unavailable.

Layout control: Preserve headers across page breaks and escalate ambiguity.

Evidence Sources

Fact sheets supplied identity, returns, expenses, and selected portfolio characteristics; holdings schedules supported quarter-end composition; annual/semiannual reports supplied operations and financial highlights; prospectuses supplied strategy, management, benchmark, and fee context.

Keywords: Mutual Funds, Large Language Models, Ground Truth, Provenance, Human Review, Abstention

Ground Truth Coverage

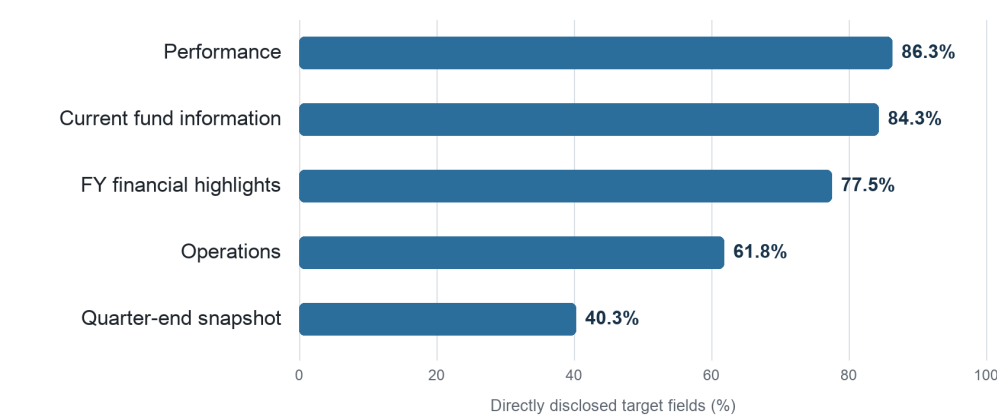


Figure 2. Direct Disclosure Only coverage across 20 funds. Performance (86.3%) and current fund information (84.3%) were strongest; quarter-end snapshots were weakest (40.3%).

Human Review Audit

1,124 review changes 457 in Chang Liu's 10-fund batch and 667 in Xinrong Hao's. Edits covered values, statuses, formulas, context, N/A decisions, and source notes; they are not counts of independent model errors.

Ask2AI Portfolioi Evaluation

Direct Match Rate by fund

Field-level agreement across 92 direct-extraction fields per fund.

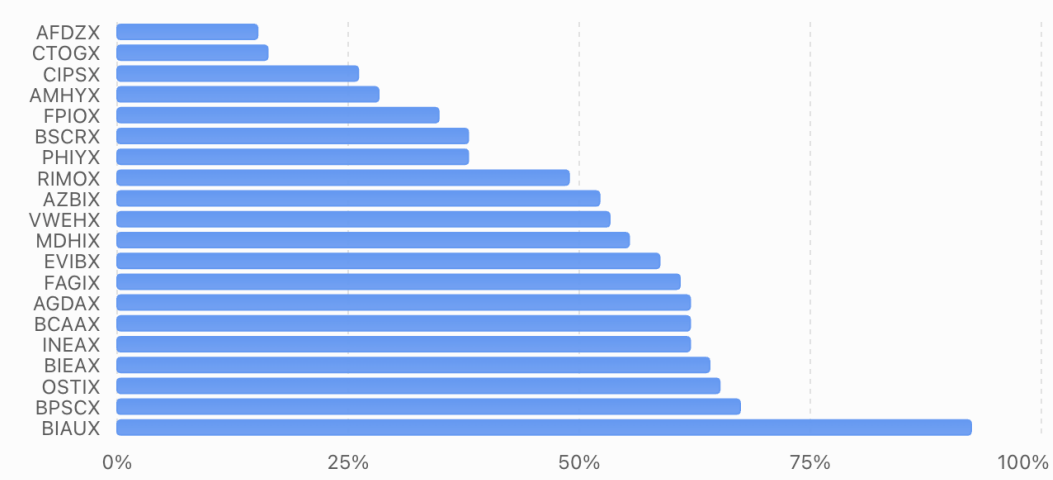


Figure 3. Direct field-level workbook agreement across 92 fields per fund. Match rates ranged from 15.2% (AFDZX) to 92.4% (BIAUX), a 77.2 percentage-point spread.

Key evaluation results 921 of 1,840 direct fields matched GT (50.1%). Among 919 mismatches, 457 were omissions (49.7% of mismatches). Formula fields matched at 44.0% (308 of 700), with 389 omitted.

Failure Modes

Period mismatch: Six-month or nearby-date values substituted for the requested fiscal period.

Context mismatch: Similar labels accepted without checking table title, definition, share class, or unit.

Unsupported calculation: Summed holdings or proxy metrics presented as direct disclosure.

Cross-page layout: Rows and values associated without continuation-header checks.

Over-extraction: At least 118 extra Portfolioi entries appeared across seven funds and require source review.

Implications and Next Iteration

Structured provenance: Retain value, unit, period, class, source, and page/table context.

Abstention by design: Use Not Disclosed when direct matching evidence is absent.

Risk-based review: Route context conflicts, low-confidence outputs, and split-page tables to human auditors.

Limitations

The 20-fund sample does not represent every asset class or disclosure format. Inter-reviewer agreement was not measured, and the 50.1% result reflects workbook-level agreement rather than value-only precision.

Conclusion

A plausible number is not sufficient evidence. Automation expands search and structure; human auditors preserve financial meaning, provenance, and the right to abstain. The strongest result is an auditable division of labor rather than a fully automated spreadsheet.