# Technical Report: FDABench: A Benchmark for Data Agents on Analytical Queries over Heterogeneous Data

Ziting Wang*, Shize Zhang†, Haitao Yuan*, Wei Dong*, Jinwei Zhu‡, Gao Cong*

*Nanyang Technological University, †National University of Singapore, ‡Huawei Technologies Co., Ltd

*ziting001@e.ntu.edu.sg, †shize.zhang@u.nus.edu, *{haitao.yuan, wei_dong, gaocong}@ntu.edu.sg, ‡zhujinwei@huawei.com

## I. Prompt Templates In FDABench Dataset Construction

### A. Web Search Prompt

**Web Search Prompt Template**

**Configuration:** Model: Perplexity/Sonar-Pro — Temperature: 0.7 — Top-p: 0.95 — Max Tokens: 6000

**Context:**
- Original query: {original_query}
- Enterprise demonstration: {huawei_example}

Drawing from the provided Huawei enterprise data agent demonstration, perform targeted web searches to support the original query. The demonstration illustrates how data agents integrate structured databases, vector-based policy retrieval, and external validation in real analytical workflows. Following this pattern, focus on gathering external contextual information, current events, regulatory frameworks, and validation sources that complement structured database facts. The retrieved information should provide interpretive insights and real-world context that enable heterogeneous data integration and multi-source reasoning.

**Search Objectives:**
- Identify recent developments, industry benchmarks, and data-driven trends directly related to the original query.
- Retrieve applicable regulatory frameworks, policy guidelines, and institutional standards governing the topic.
- Collect authoritative expert analyses, peer-reviewed research, and technical evaluations providing explanatory context.
- Extract implementation case studies or best practices illustrating how similar analytical tasks were operationalized.
- Gather quantitative indicators, empirical datasets, or performance metrics that enable comparative or temporal assessment.

**Source Requirements:**
- Peer-reviewed journals, conference papers, and academic white papers.
- Industry reports, market analyses, and technical briefs from recognized consulting or data analytics firms.
- Government, inter-agency, or regulatory authority publications relevant to the domain.
- Verified business, finance, or technology sources with demonstrable credibility and traceable citations.

**Quality Thresholds:**
- **Relevance Score:** — retrieved content must directly address the analytical focus of the original query.
- **Authority Score:** — prioritize sources with institutional, academic, or governmental authority.
- **Recency:** prefer information published within the past 24 months unless historical data are essential for longitudinal analysis.
- **Consistency:** cross-validate findings across at least two independent, high-confidence sources before inclusion.

**Output Structure:**
1) **Executive Summary (≈ 200 words)** — concise synthesis of key findings relevant to the original query.
2) **Quantitative Evidence** — tabulated or cited numerical data with explicit source attribution.
3) **Qualitative Insights** — expert commentary, thematic patterns, and emerging trends.
4) **Regulatory & Policy Context** — applicable frameworks, standards, or institutional rules.
5) **Market or Domain Intelligence** — comparative positioning, benchmarks, or regional insights.
6) **Implementation Guidance** — actionable observations or procedural recommendations grounded in the retrieved evidence.

**Variable: {huawei_example}**
Example value: *See Section 5 for detailed Huawei enterprise data agent demonstrations, including expense compliance review and employee contract verification scenarios.*

### B. Vector Database Retrieval Prompt

**Vector Retrieval Prompt Template**

**Configuration:** Embedding Model: text-embedding-3-small — Embedding Dim: 1536 — Batch Size: 50 — Chunk Size: 512 tokens — Similarity: Cosine

**Context:**
- Original query: {original_query}
- Enterprise demonstration: {huawei_example}

Drawing from the provided Huawei enterprise data agent demonstration, retrieve relevant domain knowledge and theoretical frameworks to support comprehensive query analysis. The demonstration illustrates how data agents use vector databases to retrieve policy documents, compliance guidelines, and domain-specific frameworks that provide interpretive context for database facts. Following this pattern, focus on domain-specific knowledge, conceptual foundations, technical speci-

†Intern student at Huawei Technologies Co., Ltd.

fications, and interpretive frameworks that enable reasoning across heterogeneous data sources. The goal is to provide background intelligence that complements structured database facts and external web information, enabling semantic understanding and cross-source integration for the original query.

**Retrieval Strategy:**

1) Generate vector embedding for the original query and retrieve the top-50 most similar document candidates from the vector database.
2) Apply a cosine similarity threshold of 0.75 to ensure semantic relevance.
3) Expand the query with domain-specific synonyms, abbreviations, and related terminology to capture latent contextual matches.
4) Filter retrieved candidates by domain relevance (threshold: 0.8) to exclude off-topic or cross-domain noise.
5) Apply cross-encoder reranking to refine semantic proximity among the remaining documents.
6) Select and output the final top-20 documents ranked by aggregated semantic relevance and contextual diversity.

**Document Priorities:**

- Foundational theoretical frameworks, conceptual models, and domain methodologies.
- Technical specifications, standards, or structured documentation directly relevant to the query's domain.
- Case studies, empirical analyses, and historical datasets providing contextual depth.
- Domain-specific knowledge bases and best practices capturing expert consensus.
- Regulatory and institutional documents offering compliance or policy perspectives.

**Output Requirements:**

- Document chunks with associated **relevance scores** (0–1) and ranked order.
- Source metadata including **title, publication date, document type, and origin**.
- Extracted **key concepts, entities, and semantic relations** supporting the original query.
- A concise explanation of each document's **contribution to query resolution or contextual enrichment**.

**Variable: {huawei_example}**
Example value: *See Section 5 for detailed Huawei enterprise data agent demonstrations, including expense compliance review and employee contract verification scenarios.*

### C. Dataset Construction Agent Prompt

**Dataset Generation Prompt Template**

**Configuration:** anthropic/claude-sonnet-4 — Temperature: 0.7 — Top-p: 0.95 — Frequency Penalty: 0.3 — Max Tokens: 3000
**Context Provided:**

- Original query: {original_query}
- SQL result: {gold_result}
- Web summary: {web_summary}
- Vector content: {vector_summary}

**Task Requirements:**
*1) Query Construction*

- Preserve the *original analytical intent* while extending scope to require heterogeneous data integration.
- Request *specific* database metrics (fields, aggregations, temporal/categorical breakdowns) needed for quantitative evidence.
- Specify external context requirements (theoretical frameworks, policy background, domain knowledge, validation sources).
- Design questions such that database facts alone are *insufficient* without interpretive context from unstructured sources.
- Avoid revealing numerical answers or conclusions within the question itself to prevent information leakage.

*2) Heterogeneous Integration Requirements*

- Design the task so that it *cannot be solved from any single data source*. Each source must provide essential, non-redundant information.
- **Structured Database (SQL):** Provides quantitative evidence, precise metrics, temporal/categorical breakdowns, and statistical patterns.
- **Vector Database:** Supplies theoretical frameworks, domain knowledge, conceptual models, technical specifications, and interpretive guidelines.
- **Web Search:** Provides current events, regulatory updates, policy context, external validation, and real-world corroboration.
- Ensure complementary roles: SQL answers ”what/how much,” Vector explains ”why/how to interpret,” Web validates ”current context/applicability.”

*3) Multi-Source Reasoning Chain*

1) **SQL Execution:** Identify required database tables, fields, aggregations, and constraints. Extract quantitative patterns without revealing numeric answers in the question.
2) **Vector Retrieval:** Surface relevant theoretical frameworks, domain concepts, and interpretive guidelines that contextualize database findings.
3) **Web Search:** Retrieve current events, policy updates, and external validation sources. Identify corroborating or contradictory information requiring resolution.
4) **Cross-Source Integration:** Establish connections between SQL quantitative patterns, Vector conceptual frameworks, and Web contextual validation.
5) **Framework Application:** Apply retrieved domain methodologies to reconcile information conflicts and establish causal interpretations.
6) **Synthesis:** Produce evidence-backed conclusions integrating all three sources, with clear attribution and reasoning transparency.

*4) Answer Formats*

- **Single-choice:** 4 options with exactly 1 correct; distractors must be plausible and mutually exclusive.
- **Multiple-choice:** 8 options with 2–4 correct; each correct option supported by distinct evidence.
- **Report:** structured sections (*Objective, Data/Evidence, Analysis, External Context, Conclusion, Limitations*); include evaluation criteria for correctness and completeness.

**Quality Checklist:**

- Multi-source requirement verified (SQL + Web + Vector all necessary).
- Integration necessity confirmed (single-source solution is insufficient by design).
- Answer deterministic and verifiable (clear evidence path; reproducible from sources).
- Realistic scenario (domain-appropriate assumptions; no artificial constraints).
- No information leakage (prompt does not expose numeric results or final conclusions).
- Complete reasoning chain (all steps specified and source-

aligned).

## II. System Configuration

### Configuration Parameters

| | |
|---|---|
| **Question Generation Model** | |
| Model | anthropic/claude-sonnet-4 |
| Temperature | 0.7 |
| Top-p | 0.95 |
| Frequency Penalty | 0.3 |
| **Web Search Model** | |
| Model | perplexity/sonar-pro |
| Context Length | 200,000 tokens |
| Max Output Tokens | 6,000 tokens |
| Temperature | 0.7 |
| Top-p | 0.95 |
| **Embedding Model** | |
| Model | text-embedding-3-small |
| Dimensions | 1,536 |
| Batch Size | 50 |
| API Provider | OpenAI |
| **Vector Index** | |
| Type | VectorStoreIndex (LlamaIndex) |
| Metric | Cosine Similarity |
| Top-k Retrieval | 25 |
| Chunk Size | 512 tokens |
| Splitter | SentenceSplitter |
| **Validation** | |
| Expert Pool | 6 experts |

## III. Expert Verification Protocol

### Expert Verification Guidelines

**Expert Pool:** 6 PhD-level experts in database systems and analytics, each with 5+ years of SQL proficiency across multiple dialects and familiarity with enterprise analytical scenarios.

**Verification Criteria:**

**1. Integration Necessity (1-5 Likert scale, threshold $\geq$ 4):**

- Whether multi-source integration provides semantic enrichment
- Presence of causal relationships across data sources
- Contextual validation beyond single-source facts
- Interpretive frameworks requiring heterogeneous data

**2. Answer Correctness:**

- SQL re-execution for numerical verification
- Source cross-referencing for factual accuracy
- Reasoning chain validation for logical consistency
- Any discrepancy triggers REVISE with specific feedback

**3. Realism Assessment (1-5 Likert scale, threshold $\geq$ 4):**

- Genuine business need or research scenario
- Industry relevance and current applicability
- Appropriate complexity for target difficulty level
- Domain consistency with real-world analytical tasks

**Decision Protocol:**

- **ACCEPT:** All criteria met (integration necessity $\geq$ 4, answer correct, realism $\geq$ 4)
- **REVISE:** Improvements needed (provide specific feedback $\mathcal{C}_k$)
- **DISPOSE:** Fundamental flaws (single-source solvable, unrealistic, non-convergent)

**Annotation Disagreements:**

- Majority voting with $\geq$ 2/3 agreement required
- Senior expert arbitration for unresolved conflicts
- Inter-annotator agreement validated through Cohen's $\kappa > 0.75$

## IV. Difficulty Standard and Data Selection

### A. *Difficulty Classification*

Following BIRD [1], we classify task difficulty based on multiple dimensions that comprehensively assess the complexity of heterogeneous data integration tasks. Unlike traditional text-to-SQL difficulty that focuses primarily on SQL syntactic complexity, FDABench difficulty considers the entire analytical workflow including question comprehension, external knowledge reasoning, data complexity, and query construction.

*1) Difficulty Dimensions:* Our difficulty assessment evaluates four key dimensions:

1) **Question Understanding:** Assesses the ambiguity and cognitive load required to comprehend the analytical intent.
   - Level 1: Straightforward question with clear intent
   - Level 2: Requires interpretation and domain context
   - Level 3: Highly ambiguous, requiring extensive clarification
2) **Knowledge Reasoning:** Evaluates the external knowledge required to bridge question intent and data sources.
   - Level 1: No external knowledge needed, straightforward mapping
   - Level 2: Requires domain knowledge for contextual understanding
   - Level 3: Requires extensive theoretical frameworks and causal reasoning
3) **Data Complexity:** Measures the structural complexity of schemas and the heterogeneity of data sources.
   - Level 1: Simple schema with clear relationships
   - Level 2: Complex schema requiring database documentation
   - Level 3: Highly complex with multiple heterogeneous sources
4) **SQL Complexity:** Evaluates the syntactic and logical complexity of required queries.
   - Level 1: Simple SELECT with basic aggregations
   - Level 2: Multiple JOINs and nested queries
   - Level 3: Complex subqueries, window functions, and CTEs

*2) Final Difficulty Assignment:* Each task receives a difficulty score based on expert voting and reasoning token analysis:

- **Easy (Level 1):** Straightforward question understanding; no external knowledge required; simple schema; basic SELECT with aggregations. Single dimension scores primarily at Level 1.
- **Medium (Level 2):** Requires interpretation and domain context; domain knowledge needed; complex schema requiring documentation; multiple JOINs and nested queries. Mixed dimension scores with majority at Level 2.

- **Hard (Level 3):** Highly ambiguous requiring clarification; extensive theoretical frameworks needed; highly complex with multiple heterogeneous sources; complex CTEs and window functions. Multiple dimensions scoring at Level 3.

The final difficulty assignment combines expert votes weighted by reasoning token distribution from LLM analysis. Final distribution achieved: 20.68% Easy (415/2,007), 32.84% Medium (659/2,007), 46.49% Hard (933/2,007), with inter-annotator agreement validated through Cohen’s $\kappa > 0.75$.

### *B. Unstructured Data Collection Criteria*

*1) Selection Process:* For each candidate query from the original datasets containing over 23,900 queries, we manually curated relevant unstructured files by searching authoritative sources (e.g., Google Scholar, arXiv) based on the database topic and query intent, yielding over 1,600 files across 50+ domains.

*2) Retention Criteria:* Files are retained if they provide essential context for query interpretation, domain knowledge for result validation, or causal explanations connecting database patterns to real-world events. Specifically:

- **Complementarity:** Providing causal explanations or contextual validation not derivable from database facts alone
  - Financial regulations for transaction analysis
  - Medical guidelines for patient outcome assessment
  - Market reports explaining sales trends and anomalies
  - Policy frameworks for compliance verification
- **Authority:** Peer-reviewed publications or institutional documents only
  - Academic papers from Google Scholar, arXiv, IEEE Xplore, ACM Digital Library
  - Government publications and regulatory documents
  - Institutional white papers from recognized organizations
  - Exclude user-generated content without editorial oversight
- **Coverage:** Balanced representation across domains
  - Finance, healthcare, retail, manufacturing sectors
  - Multiple perspectives and methodologies per domain
  - Regional contexts for cross-cultural analytical scenarios

*3) Data Source Categories:* The collected corpus includes:

- **Academic Literature:** Theoretical frameworks and domain knowledge from peer-reviewed sources
- **Industry Documentation:** Market analyses, regulatory documents, and best practices
- **Multimedia Content:** Transcripts from educational videos and expert seminars
- **Technical Specifications:** API documentation and system architecture descriptions
- **Enterprise Knowledge:** Huawei’s real application cases with analytical workflow patterns

*4) Processing Pipeline:* Collected unstructured data undergoes standardized preprocessing:

1) **Format Normalization:** Convert multimedia content to text transcripts, extract text from PDFs, and normalize encoding.
2) **Chunking Strategy:** Segment documents into 512-token chunks using sentence-boundary-aware splitting to preserve semantic coherence.
3) **Embedding Generation:** Generate vector representations using text-embedding-3-small (1,536 dimensions) with batch processing for efficiency.
4) **Index Construction:** Build VectorStoreIndex using LlamaIndex with cosine similarity metric and metadata preservation for source attribution.
5) **Quality Verification:** Sample 10% of chunks for manual review to ensure embedding quality and retrieval relevance.

This systematic collection and processing ensures FD-ABench’s unstructured data sources provide genuine semantic value for heterogeneous analytical tasks, reflecting real-world scenarios where data agents must integrate structured databases with external knowledge.

## V. Huawei’s Enterprise Data Agent Experience Example

The following example demonstrates the real-world enterprise data agent scenario from Huawei that guides our dataset construction process:

**Huawei Expense Reimbursement Compliance Review Data Agent**

Huawei’s Expense ERP Data Agent conducts intelligent compliance reviews through heterogeneous data integration. The reimbursement review data agent operates through the following analytical workflow:

**Structured Database Query:**

- Retrieve recent employee expense claim information from the relational database
- Query fields include employee ID, grade level, expense amount, submission timestamp, expense category, and approval status
- Execute aggregations to identify statistical patterns and outliers

**Vector Database Retrieval:**

- Based on employee grade level and expense category, query the vector database for applicable reimbursement policies
- Retrieve relevant policy documents, compliance guidelines, and historical precedents
- Extract allowable expense limits, required documentation, and approval workflows

**External Web Validation:**

- Cross-validate supporting documentation and transaction timestamps with regulatory databases
- Check compliance with internal policy updates and external audit requirements
- Verify merchant information and transaction authenticity against external sources

**Heterogeneous Data Integration:**

- **SQL provides** quantitative patterns: expense amounts, submission timing, approval rates
- **Vector database provides** policy context: allowable limits, required documentation, compliance rules
- **Web sources provide** validation: regulatory updates, merchant verification, audit trail

**Anomaly Detection and Reporting:**
The data agent identifies compliance issues including:

- Late submissions exceeding policy deadlines
- Over-claiming beyond grade-specific limits
- Missing or insufficient supporting documentation
- Policy violations based on recent regulatory updates

**Final Output:**
The agent generates structured compliance summaries that:

- Link quantitative database facts (expense amounts, timestamps) with policy violations
- Provide causal explanations grounded in retrieved policy documents
- Include external validation from regulatory sources
- Recommend corrective actions based on compliance frameworks

This enterprise example demonstrates genuine heterogeneous data integration where **no single source is sufficient**. Database queries alone cannot determine policy compliance without policy documents. Policy documents alone cannot identify violations without transaction data. External validation is required to ensure regulatory compliance. This pattern guides FDABench's dataset construction to ensure tasks require authentic multi-source reasoning.

### Huawei Employee Contract Verification Data Agent

Huawei's Employee Contract Verification Data Agent conducts intelligent inspections of labor agreements across the organization through heterogeneous data integration. The contract verification agent operates through the following analytical workflow:

**Structured Database Query:**

- Retrieve employee information from the HR relational database
- Query fields include employee ID, grade level, department, employment date, and contract renewal history
- Execute joins to correlate employee grades with applicable regulatory requirement categories

**Vector Database Retrieval:**

- Based on employee grade level, query the vector database for grade-specific regulatory knowledge bases
- Each grade level has corresponding policy requirements: executive grades require enhanced compensation standards, junior grades follow standard labor templates
- Retrieve relevant compliance requirements: minimum wage thresholds, maximum probation lengths, mandatory leave policies, and termination notice periods

**External Web Validation:**

- Cross-validate contract clauses with recent labor law amendments from government labor bureau websites
- Check compliance with latest court rulings on employment disputes and regulatory policy briefings
- Verify minimum wage standards against regional labor authority announcements

**Heterogeneous Data Integration:**

- **Structured HR database provides** employee facts: grade levels, employment dates, department assignments
- **Unstructured contract documents provide** legal terms: compensation clauses, probation durations, termination conditions
- **Vector database provides** grade-specific standards: required salary ranges, permissible probation lengths, mandatory benefits
- **Web sources provide** current legal context: recent law changes, regulatory updates, judicial interpretations

**Compliance Issue Detection and Reporting:**
The data agent identifies contract violations including:

- Below-standard compensation clauses violating grade-specific minimum wage requirements
- Excessive probation periods exceeding legal limits for the employee's grade level
- Missing mandatory legal disclosures required by recent labor law amendments
- Outdated contract templates not reflecting current regulatory requirements

**Final Output:**
The agent generates structured contract review summaries that:

- Link employee grade data with specific contract clauses and applicable regulations from grade-specific knowledge bases
- Provide clause-by-clause compliance analysis referencing grade-specific policy requirements
- Include external legal validation from labor authority updates and court rulings
- Recommend contract amendments with specific regulatory citations

This enterprise example demonstrates multi-modal heterogeneous integration where **each data source provides essential non-redundant information**. Structured employee data identifies which regulatory standards apply. Unstructured contract documents contain the actual terms requiring review. Grade-specific vector knowledge bases provide the compliance benchmarks. External web sources ensure alignment with current legal requirements.

## VI. Quality Control and Validation Results

### A. Overall Quality Control

The agent-expert collaboration framework accepted 2,007 from 4,127 generated candidates, achieving a 48.6% acceptance rate. This stringent acceptance rate reflects the rigorous quality control ensuring only tasks requiring genuine heterogeneous data integration enter FDABench.

*1) Iterative Refinement Statistics:* The iterative expert verification process demonstrates high efficiency:

**Iterative Refinement Statistics**

| | |
|---|---|
| **First-Iteration Acceptance** | |
| Accepted Without Revision | 1,523 tasks |
| First-Iteration Rate | 75.9% (1,523/2,007) |
| **Second-Iteration Acceptance** | |
| Accepted After One Revision | 367 tasks |
| Second-Iteration Rate | 18.3% (367/2,007) |
| **Third+ Iteration Acceptance** | |
| Accepted After 2+ Revisions | 117 tasks |
| Third+ Iteration Rate | 5.8% (117/2,007) |
| **Efficiency Metrics** | |
| Average Iterations per Task | 1.9 iterations |
| Average Review Time per Task | 8.3 minutes |

The high first-iteration acceptance rate (75.9%) demonstrates that the dataset construction agent, guided by Huawei's enterprise demonstrations $\mathcal{E}$, can produce high-quality drafts efficiently.

*2) Rejection Analysis:* Detailed analysis of the 2,120 rejected candidates (4,127 generated - 2,007 accepted) reveals:

**Rejection Reasons Distribution**

| **Rejection Reason** | **Count** | **Percentage** |
|---|---:|---:|
| Single-Source Solvability | 1,058 | 49.9% |
| SQL Errors | 558 | 26.3% |
| Insufficient Integration | 322 | 15.2% |
| Unrealistic Scenarios | 182 | 8.6% |
| **Total Rejected** | **2,120** | **100%** |

The primary rejection reason—single-source solvability (49.9%)—confirms that expert verification successfully filters tasks that do not require genuine heterogeneous data integration, addressing Challenge **C1** (verifying ground truth correctness) and **C3** (ensuring diverse task types with varying difficulty levels).

### B. Final Dataset Distribution

**Task Type Distribution**

| **Task Type** | **Count** | **Percentage** |
|---|---:|---:|
| Report | 668 | 33.28% |
| Single-Choice | 579 | 28.90% |
| Multiple-Choice | 760 | 37.87% |
| **Total** | **2,007** | **100%** |

**Difficulty Distribution**

| **Difficulty Level** | **Count** | **Percentage** |
|---|---:|---:|
| Easy | 415 | 20.68% |
| Medium | 659 | 32.84% |
| Hard | 933 | 46.49% |
| **Total** | **2,007** | **100%** |

The final dataset contains 2,007 tasks across 50+ domains spanning 139 structured databases. The distribution reflects the natural complexity of real-world analytical scenarios, with a higher proportion of hard tasks (46.49%) representing the sophisticated reasoning required for genuine heterogeneous data integration. The balanced task type distribution ensures comprehensive evaluation across precise answers (single-choice), complex reasoning (multiple-choice), and comprehensive analysis (report).

## References

[1] J. Li *et al.*, "Can LLM already serve as A database interface? A big bench for large-scale database grounded text-to-sqls," in *NeurIPS 2023*, 2023.