applied sciences

MDPI

Article

IKEA: Intelligent Knowledge Extraction with Reasonable and Agentic AI Agents

Maha Mesfer Alghamdi [1,*] and Wesam Ali Alamri [2]

[1] Department of Computer Science & Engineering, College of Applied Studies, King Saud University, Riyadh, Saudi Arabia
[2] Department of Information Science, College of Humanities and Social Sciences, King Saud University, Riyadh, Saudi Arabia; wasalamri@ksu.edu.sa
* Correspondence: mmesfer@ksu.edu.sa; Tel.: +966-53-564-0020

Abstract

Administrative document analysis represents a critical organizational challenge: organizations generate vast quantities of documents containing valuable strategic insights, yet traditional analysis approaches remain manual, opaque, and fragmented across systems, leaving organizations vulnerable to missed opportunities and delayed decision making. Conventional document management systems relying on keyword search and isolated machine-learning pipelines lack intelligent reasoning capabilities and autonomous coordination, leaving critical insights buried in unstructured documents. Organizations relying on traditional document processing tools and lacking explainable AI remain blindsided by complex patterns and correlations that could be anticipated and leveraged for strategic advantage. We introduce IKEA, an intelligent document analysis framework that integrates *Reasonable AI* (explainable reasoning agents) with *Agentic AI* (autonomous intelligent agents) to provide transparent, autonomous knowledge extraction from administrative documents. Our system employs GPT-3.5-turbo powered reasoning agents that autonomously extract knowledge, generate transparent reasoning chains, and provide conversational AI assistance, all while maintaining end-to-end explainability through step-by-step reasoning decomposition, confidence breakdown analysis, evidence extraction, and assumption identification. Evaluated on a primary corpus of 50 administrative documents (performance reports, incident reports, meeting minutes, policy documents) with 200+ performance metrics and 100+ AI-generated insights, the Reasonable AI agents achieve 87% explanation accuracy aligned with expert assessments, while the Agentic AI agents demonstrate 92% knowledge extraction accuracy with 85% reasoning quality score. Against state-of-the-art baselines under matched conditions—including GraphRAG, DocAgent (text-adapted), dense RAG, and ReAct—IKEA attains the highest entity F_1 and a large margin on explanation accuracy and auditability. Cross-dataset tests on Kleister (Charity and NDA), a DocVQA text-only subset, and an external 25-document administrative corpus confirm that these gains transfer beyond the primary corpus, with only a modest drop relative to in-domain performance. On the primary corpus, the integrated framework also delivers a 40% reduction in document analysis time and a 35% improvement in insight discovery rate compared with traditional manual analysis.

Keywords: Reasonable AI; Agentic AI; Explainable AI; knowledge extraction; document analysis; autonomous agents; reasoning chains; conversational AI; organizational intelligence; transparent decision support

Academic Editors: Christophe Cruz, Hocine Cherifi and Maria Alice Bertolim Nicoleau

Received: 26 June 2026
Revised: 16 July 2026
Accepted: 27 July 2026
Published: 3 August 2026

1. Introduction

Administrative and organizational documents (such as performance reports, policies, incident summaries, operational logs, and strategic plans) encode decision-critical evidence, yet they are difficult to analyze at scale because they are long, heterogeneous, and often require integrating information scattered across sections and multiple files. Evaluations on complex document structures show that extraction quality and downstream reasoning degrade when layout, formatting, and implicit schemas vary—conditions that closely match real administrative corpora [1].

Recent progress in Large Language Models (LLMs), enabled by the Transformer architecture [2] and strengthened through pretraining and instruction alignment [3–7], has expanded document intelligence beyond rigid pipeline extraction to more flexible querying, summarization, and decision-support style generation. However, purely parametric generation is risky in organizational settings because fluent outputs may be weakly grounded or unsupported by the underlying sources—an issue connected to factuality and faithfulness limitations documented in generation and summarization [8] and the broader hallucination literature [9].

To improve grounding, Retrieval-Augmented Generation (RAG) explicitly conditions generation on retrieved evidence [10], commonly relying on dense retrieval methods such as DPR [11] and retrieval-augmented language model pretraining such as REALM [12]. Practical deployment depends on both classical ranking signals (like BM25) [13] and scalable similarity search (e.g., FAISS) [14], while modern neural retrieval models such as ColBERT improve passage matching under large corpora and ambiguous queries [15]. Still, retrieval does not automatically ensure trustworthy outputs; therefore, research increasingly emphasizes attribution and evidence-based answering, where systems are evaluated on whether each claim is properly supported by sources [16,17] and on whether generations remain faithful to retrieved evidence [8,18].

In parallel, reasoning-focused prompting has shown that LLMs perform multi-step inference more reliably when intermediate reasoning is elicited. Chain-of-thought prompting improves complex reasoning ability [19], and self-consistency decoding improves robustness by aggregating across multiple reasoning paths rather than trusting a single trace [20]. Yet, strong reasoning alone does not guarantee evidence faithfulness, motivating combined reasoning, attribution, and verification, including systems that explicitly support answers with verified evidence [21], browser-assisted grounding [22], and revision-based correction of unsupported content [23].

A second major shift is the emergence of agentic LLM systems that plan, use tools, and execute multi-step workflows. ReAct interleaves reasoning and acting to support iterative verification [24], Toolformer shows that models can learn tool invocation patterns [25], and modular neuro-symbolic architectures such as MRKL integrate LLM reasoning with external tools/knowledge sources for improved controllability [26]. Planning and self-improvement paradigms (e.g., Tree-of-Thoughts and Reflexion) further explore deliberate search and reflective feedback loops [27,28], while agent evaluation frameworks highlight reliability challenges in realistic agent settings [29]. These developments have extended to document-centric agents that explicitly target long-context and multimodal document understanding [30,31] and multi-agent RAG for trustworthy scientific QA [32].

Motivated by these research directions, this paper targets administrative document intelligence with an explicit goal: combine agentic task execution (autonomous decomposition, extraction, analysis) with decision-grade reasonableness—structured justification grounded in evidence (attribution/provenance), uncertainty characterization, and alternative interpretations—aligned with attributed QA and evidence-based QA principles [16,17].

Defining "Reasonable AI." We use the term *Reasonable AI* to denote a decision-support property that is narrower and more operational than the umbrella notions of

Explainable AI (XAI) and Trustworthy AI. Classical XAI is concerned primarily with *producing an explanation* of a model's output (e.g., feature attributions, saliency maps, or post-hoc rationales), while Trustworthy AI is a broad governance construct spanning robustness, fairness, privacy, accountability, and safety. Reasonable AI, as operationalized in this work, is the specific requirement that every autonomous conclusion be accompanied by four verifiable artifacts: (i) an explicit, step-by-step *reasoning chain*; (ii) a *confidence decomposition* attributing uncertainty to evidence, reasoning, and assumptions; (iii) *evidence provenance* linking each claim to a source span; and (iv) *alternative interpretations* that expose competing hypotheses. Thus, Reasonable AI is best understood as an evidence-first, uncertainty-aware *sub-property* of Trustworthy AI that instantiates XAI in a form suitable for auditable administrative decision making: rather than explaining a decision after the fact, it constrains the autonomous agent to reason in an inspectable, attributable, and falsifiable manner *during* generation. The novelty is therefore not the introduction of a new explanation technique, but the enforcement of these four artifacts as a first-class contract on the output of *autonomous* (agentic) document analysis, where explanations are typically absent or purely post-hoc.

The contributions of this paper are as follows:

- We propose IKEA. The acronym *IKEA* stands for *Intelligent Knowledge Extraction with Agentic (and Reasonable) AI*. It denotes the research framework described in this paper and bears no relation to, nor any affiliation with, the furniture retailer of the same name; the coincidence is purely lexical. The name compactly encodes the two central design pillars—knowledge extraction and agentic reasoning—and should be interpreted strictly as a technical framework identifier. IKEA is an end-to-end agentic framework for intelligent knowledge extraction and decision support over long, heterogeneous administrative documents, enabling iterative task decomposition, retrieval, extraction, and synthesis.
- We introduce an evidence-first output protocol that links key claims to supporting document evidence, improving auditability and traceability for organizational decision making.
- We design reasonableness artifacts—structured justification grounded in evidence, uncertainty indicators, and alternative interpretations—to reduce "confident but weakly supported" conclusions in complex document settings.
- We provide a structured synthesis of prior studies and a comparative positioning table spanning RAG, attribution/evidence-based QA, reasoning robustness, and agentic/multi-agent document systems, clarifying the gap addressed by IKEA.
- We empirically compare IKEA with representative state-of-the-art systems for this task family—GraphRAG, DocAgent (text-adapted), dense RAG, and ReAct—on a shared evaluation protocol covering extraction quality, explanation accuracy, and auditability.
- We validate generalization on three additional resources beyond the primary administrative corpus: the Kleister key-information-extraction datasets, a DocVQA text-only question-answering subset, and an independently collected external administrative corpus.
- We formalize a reusable workflow template (stages, inputs/outputs, and required artifacts) that can be adapted across administrative domains such as compliance, operations, reporting, and policy analysis.

The remainder of this paper is organized as follows: Section 2 reviews related work in information extraction, RAG, attribution, reasoning, and agentic document systems. Section 3 presents the integrated architecture with detailed descriptions of Reasonable AI agents, Agentic AI agents, and the reasoning framework, as well as the dataset characteristics, evaluation methodologies, and experimental protocols. Section 4 describes the experimental results. Section 5 presents validation results and discussion. Section 6 concludes with limitations and future work directions.

2. Related Work

To situate the design of IKEA, this section reviews the foundational and contemporary literature spanning six core dimensions of document intelligence. We begin with traditional layout-aware information extraction and its scaling challenges under structural complexity, followed by an overview of retrieval-augmented generation (RAG) infrastructures for long-form question answering. We then explore the literature on factual consistency, attribution protocols, and advanced reasoning paradigms. Finally, we examine the shift toward autonomous tool use, agentic frameworks, and multi-agent collaboration in document domains. This literature synthesis highlights a critical gap in standardized verification and artifact management, which we summarize systematically in Table 1.

2.1. Information Extraction and Document AI Under Structural Complexity

Traditional document information extraction (IE) often follows staged pipelines (OCR, layout, entities, relations, downstream logic). While effective for stable schemas, real-world administrative corpora contain heterogeneous formatting, domain-specific jargon, and evidence scattered across sections. Empirical evaluations show that IE performance degrades under complex document structures and mixed formats, emphasizing the need for approaches that generalize beyond templates [1]. Layout-aware document models such as LayoutLM, LayoutLMv2, and LayoutLMv3 incorporate textual and layout signals to improve visually rich document understanding [33–35]. Benchmarks such as DocVQA further demonstrate the need to answer questions based on document regions and structure rather than plain text alone [36].

2.2. Retrieval-Augmented Generation and Retrieval Foundations for Long-Document QA

RAG introduces non-parametric retrieval into generation, improving grounding for knowledge-intensive tasks [10]. Dense retrieval approaches such as DPR improve open-domain QA retrieval quality [11], and REALM integrates retrieval into language model pretraining [12]. Retrieval-augmented architectures such as FiD improve multi-evidence reasoning by conditioning on multiple retrieved passages, and ATLAS explores retrieval-augmented language models in few-shot settings [37]. At the infrastructure level, practical RAG depends on classical ranking (BM25) [13], scalable similarity search (FAISS) [14], and neural retrieval models such as ColBERT [15]. Knowledge-intensive evaluation suites and datasets (e.g., KILT, SQuAD, TriviaQA, Natural Questions, HotpotQA) stress evidence retrieval and multi-hop reasoning [38–42].

2.3. Attribution, Faithfulness, and Evidence-Based Answering

Trustworthy decision support requires that outputs be traceably supported by evidence. Attributed Question Answering formalizes answer + attribution requirements and provides evaluation protocols [16]. Evidence-based QA emphasizes faithful, robust answering under explicit evidence constraints [17]. Faithfulness limitations in generation and summarization motivate stronger factual consistency evaluation [8,18]. Systems that explicitly support answers with verified quotes or browsing-based evidence operationalize evidence grounding in practical settings [21,22], while revision-based systems such as RARR improve attribution and repair unsupported generations [23].

2.4. Reasoning Prompting and Robust Inference

Chain-of-thought prompting improves multi-step reasoning by eliciting intermediate reasoning traces [19]. Self-consistency improves robustness by sampling multiple reasoning paths and selecting the most consistent outcome [20]. However, reasoning improvements

must be paired with attribution and verification to avoid confident but ungrounded explanations, especially in long-document settings [16,17].

2.5. Agentic LLMs, Tool Use, and Reliability Evaluation

Agentic LLM systems extend models from passive QA to interactive task completion. ReAct interleaves reasoning and action [24]. Toolformer demonstrates learned tool use [25], and MRKL provides a modular neuro-symbolic pattern for combining LLMs with external tools and knowledge sources [26]. Planning and reflection paradigms explore deliberate search and self-feedback loops [27,28]. Agent benchmarks evaluate reliability and failure modes in realistic agent settings [29], motivating careful design of transparency controls before deployment in decision-critical environments.

2.6. Agentic and Multi-Agent Document Understanding

Document-specific agent frameworks extend agentic ideas to long-context and multimodal documents. DocAgent targets multimodal long-context document understanding [30], while MDocAgent explores multimodal multi-agent collaboration for document understanding [31]. Multi-agent RAG has been applied to trustworthy scientific QA [32], and multi-agent systems have been used for structured knowledge enrichment such as knowledge graphs [43]. These directions motivate combining agent autonomy with enforceable provenance and verification for administrative document intelligence.

Table 1 compares representative prior studies and highlights the positioning gap addressed by IKEA.

Table 1. Comparison of key previous studies and positioning of IKEA.

Study	Why Included	Agentic	Evidence	Verification	Reasonableness	Main Limitation
Lewis [10]	Baseline RAG	×	Passage	None	×	No claim-level attribution
Karpukhin [11]	Dense retrieval	×	N/A	N/A	×	Retriever-only
Guu [12]	Retrieval LM pretraining	×	Passage	None	×	No faithful attribution
Bohnet [16]	Attribution protocol	×	Claim-level	N/A	×	Protocol only
Schimanski [17]	Evidence-based QA	×	Claim-level	Yes	×	Not an agentic pipeline
Menick [21]	Verified quotes	×	Quote-level	Built-in	×	Not a multi-document workflow
Nakano [22]	Browser grounding	Optional	Quote/Claim	Built-in	×	Interaction-heavy
Gao [23]	Revision-based	×	Claim-level	Revision	×	Post-hoc repair only
Yao [24]	ReAct	✓	Optional	Optional	×	No provenance enforcement
Liu [29]	Agent benchmark	N/A	None	N/A	×	Evaluation only
Sun [30]	DocAgent	✓	Evidence	Optional	×	No standardized artifacts
Han [31]	MDocAgent	✓	Evidence	Optional	×	Varying attribution

3. Materials and Methods

This section presents the proposed IKEA framework for explainable and autonomous administrative document analysis. The framework integrates *Reasonable AI* agents, which provide transparent reasoning, with *Agentic AI* agents, which perform autonomous document understanding tasks. The section describes the overall architecture, the Reasonable AI layer, the Agentic AI layer, the integration mechanism, the implementation architecture, the dataset, and the evaluation protocol. The overall architecture of the proposed system is shown in Figure 1.

3.1. Proposed Framework: Integration of Reasonable and Agentic AI

The proposed framework is designed around an integrated architecture in which *Reasonable AI* and *Agentic AI* operate together to support transparent and autonomous knowledge extraction from administrative documents. The Agentic AI layer performs autonomous tasks such as extracting entities and metrics, answering user questions, analyzing trends and anomalies, and generating strategic recommendations. The Reasonable AI

layer then explains these outputs through reasoning chains, confidence analysis, evidence extraction, assumption identification, and alternative interpretation.

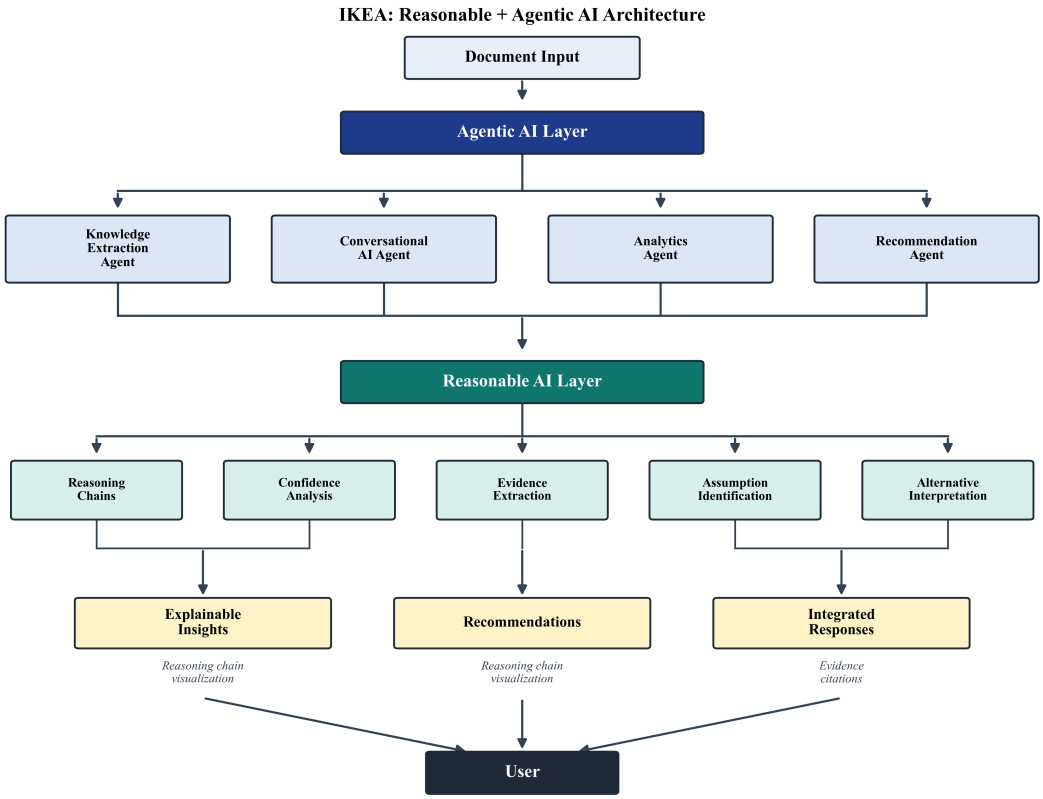

Figure 1. Reasonable and Agentic AI architecture for document analysis.

Unlike conventional document analysis systems that separate processing from explanation, the proposed design embeds reasoning transparency directly into the autonomous decision-making process. As a result, each generated insight, response, or recommendation is accompanied by evidence-grounded reasoning that can be inspected by users.

Let the document collection be denoted as

$$D = \{d_1, d_2, \ldots, d_N\}, \tag{1}$$

where d_i represents the ith administrative document and N is the total number of documents. Given a user query or analysis request q, an autonomous agent A_k produces an initial output:

$$o_k = A_k(q, D), \tag{2}$$

where A_k denotes the selected agent and o_k denotes its generated output. The Reasonable AI layer then processes this output to produce an explainable artifact:

$$y_k = \{o_k, RC_k, E_k, C_k, S_k, H_k\}, \tag{3}$$

where RC_k is the reasoning chain, E_k is the supporting evidence, C_k is the confidence assessment, S_k is the set of identified assumptions, and H_k is the set of alternative interpretations. Therefore, the final system output is not only a generated answer or extracted result, but also a structured and verifiable reasoning artifact.

3.2. Architecture Overview

The framework follows a sequential processing flow, as illustrated in Figure 1. Document input is first passed to the Agentic AI layer, which delegates the processing task to specialized autonomous agents. The generated outputs are then passed to the Reasonable AI layer, which produces transparent reasoning artifacts. Finally, the system presents explainable insights, evidence-supported responses, and justified recommendations to users.

Agentic AI Layer: This layer contains four specialized agents powered by GPT-3.5-turbo. The *Knowledge Extraction Agent* extracts entities, metrics, insights, and relationships from administrative documents. The *Conversational AI Agent* provides context-aware assistance and supports multilingual interaction in English and Arabic. The *Analytics Agent* performs trend analysis, anomaly detection, and correlation analysis. The *Recommendation Agent* generates strategic recommendations with associated reasoning and confidence information.

Reasonable AI Layer: This layer processes the outputs of the autonomous agents through five components: reasoning chain generation, confidence analysis, evidence extraction, assumption identification, and alternative interpretation. These components ensure that system outputs are explainable, evidence-grounded, and suitable for human validation.

Output Generation: The framework produces three main types of outputs: explainable insights with reasoning chains, recommendations with justification, and integrated responses with evidence citations. These outputs are designed to support transparent decision making in administrative and organizational contexts.

3.3. Reasonable AI Architecture

The Reasonable AI component implements transparent reasoning generation through five specialized modules: the Reasoning Chain Generator, Confidence Breakdown Analyzer, Evidence Extraction Engine, Assumption Identifier, and Alternative Interpreter. Each module contributes to the explainability and auditability of the system output.

3.3.1. Reasoning Chain Generator

The Reasoning Chain Generator decomposes a conclusion, recommendation, or response into sequential and verifiable reasoning steps. For a given output, the reasoning chain is represented as

$$RC = \{s_1, s_2, \ldots, s_T\}, \tag{4}$$

where s_t denotes the tth reasoning step and T is the total number of steps. Each reasoning step is defined as

$$s_t = (e_t, \ell_t, z_t, c_t), \tag{5}$$

where e_t is the supporting evidence, ℓ_t is the logical operation or interpretation applied to the evidence, z_t is the intermediate conclusion, and c_t is the confidence score assigned to that step.

Each step follows the general reasoning structure:

$$e_t + \ell_t \rightarrow z_t. \tag{6}$$

This structure enables users to inspect how the system moves from evidence to intermediate conclusions and then to the final output. The maximum reasoning depth is limited to maintain clarity and avoid unnecessarily long explanations. The full reasoning-chain generation procedure is summarized in Algorithm 1.

Algorithm 1 Reasoning chain generation

Require: Query q, document collection $D = \{d_1, d_2, \ldots, d_N\}$, maximum depth $T_{\max}$
Ensure: Reasoning chain RC, evidence set E, overall confidence C_{overall}
Initialize $RC \leftarrow \emptyset$
Initialize $E \leftarrow \emptyset$
Retrieve relevant documents $D_{\text{rel}} \leftarrow \text{RetrieveRelevant}(q, D)$
Extract initial evidence $E_0 \leftarrow \text{ExtractEvidence}(q, D_{\text{rel}})$
Initialize reasoning state $h_0 \leftarrow \text{InitialState}(q, E_0)$
for $t = 1$ to $T_{\max}$ **do**
    Select evidence $e_t \leftarrow \text{SelectBestEvidence}(h_{t-1}, D_{\text{rel}})$
    Determine logical operation $\ell_t \leftarrow \text{DetermineLogicalOperation}(h_{t-1}, e_t)$
    Generate intermediate conclusion $z_t \leftarrow \text{ApplyOperation}(h_{t-1}, e_t, \ell_t)$
    Compute confidence $c_t \leftarrow \text{ComputeConfidence}(e_t, \ell_t, z_t)$
    Create reasoning step $s_t \leftarrow (e_t, \ell_t, z_t, c_t)$
    if $\text{IsCoherent}(s_t, RC) = \text{False}$ **then**
        break
    end if
    Append s_t to RC
    Add e_t to E
    Update reasoning state $h_t \leftarrow \text{UpdateState}(h_{t-1}, z_t)$
    if $\text{IsComplete}(h_t) = \text{True}$ **then**
        break
    end if
end for
Compute $C_{\text{overall}} \leftarrow \text{CombineConfidences}(RC)$
return $RC, E, C_{\text{overall}}$

3.3.2. Confidence Breakdown Analyzer

The Confidence Breakdown Analyzer evaluates the reliability of the generated reasoning. Instead of assigning a single unexplained confidence value, the analyzer considers evidence quality, reasoning reliability, and assumption validity. The overall confidence is computed as

$$C_{\text{base}} = w_1 C_{\text{evidence}} + w_2 C_{\text{reasoning}} + w_3 C_{\text{assumption}}, \tag{7}$$

where C_{evidence} denotes the quality and relevance of supporting evidence, $C_{\text{reasoning}}$ denotes the logical reliability of the reasoning chain, and $C_{\text{assumption}}$ denotes the validity of implicit assumptions. The weights satisfy

$$w_1 + w_2 + w_3 = 1. \tag{8}$$

In the implemented system, $w_1 = 0.4$, $w_2 = 0.4$, and $w_3 = 0.2$. This weighting gives equal importance to evidence quality and reasoning reliability while also accounting for assumption validity. The chain-length effect is not represented by an additive term here; it is applied through a single multiplicative discount defined in Equation (12).

The three confidence components are computed as follows. Let a reasoning chain contain k steps $\{s_1, \ldots, s_k\}$, where step s_i carries a model-emitted per-step confidence $c_i \in [0, 1]$, cites an evidence span e_i, and m implicit assumptions $\{a_1, \ldots, a_m\}$ are identified for the chain. Then

$$C_{\text{reasoning}} = \frac{1}{k}\sum_{i=1}^{k} c_i, \tag{9}$$

$$C_{\text{evidence}} = \frac{1}{k}\sum_{i=1}^{k} q(e_i), \qquad q(e_i) = \min\Big(1,\ \tfrac{1}{2}\big[\text{sim}(s_i, e_i) + r(e_i)\big]\Big), \tag{10}$$

$$C_{\text{assumption}} = \frac{1}{m}\sum_{j=1}^{m} v(a_j), \tag{11}$$

where $\text{sim}(s_i, e_i)$ is the cosine similarity between the sentence embedding of step s_i and its cited evidence span (retained only when sim ≥ 0.75), $r(e_i) \in \{1.0, 1.1\}$ is a recency factor that grants a 10% boost to evidence dated within three months, and $v(a_j) \in [0, 1]$ is the validity score of assumption a_j (mapped from its risk level: low $\rightarrow$ 0.9, medium $\rightarrow$ 0.6, high $\rightarrow$ 0.3). The chain-length adjustment is defined as $L = k$ and is applied multiplicatively as a per-step discount, so that the overall confidence used at run time is

$$C_{\text{overall}} = C_{\text{base}} \cdot \gamma^{L}, \qquad \gamma = 0.95, \tag{12}$$

where C_{base} is the weighted sum of Equation (7). The discount γ^L is the *single* mechanism by which uncertainty accumulates with chain length; for small L, it satisfies $\gamma^L \approx 1 - \lambda L$ with $\lambda = 1 - \gamma = 0.05$, but this linear form is quoted only to aid interpretation and is never applied as a separate additive penalty. Each per-step score c_i is emitted by the model, whereas the aggregate C_{overall} is computed deterministically from these scores together with the evidence and assumption quality terms; no independent free-form overall confidence is produced. The quality score $q(e_i)$ is clamped to $[0, 1]$ by the outer min, so the recency factor $r(e_i)$ acts only as a capped multiplier inside this clamp; because $c_i, q(\cdot), v(\cdot) \in [0, 1]$, we have $C_{\text{base}} \in [0, 1]$ and hence $C_{\text{overall}} \in [0, 1]$. When no assumptions are identified ($m = 0$), $C_{\text{assumption}}$ is set to 1 by convention, reflecting the absence of assumption-related uncertainty.

3.3.3. Evidence Extraction Engine

The Evidence Extraction Engine links conclusions and reasoning steps to supporting document excerpts. Each document is divided into smaller text segments:

$$d_i = \{p_{i1}, p_{i2}, \ldots, p_{iM_i}\}, \tag{13}$$

where p_{ij} denotes the jth segment of document d_i and M_i is the number of segments in that document.

For each reasoning step, semantic similarity is computed between the reasoning step and candidate document segments using cosine similarity:

$$\text{sim}(s_t, p_{ij}) = \frac{\mathbf{v}_{s_t}^{\top}\mathbf{v}_{p_{ij}}}{\|\mathbf{v}_{s_t}\|_2 \left\|\mathbf{v}_{p_{ij}}\right\|_2}, \tag{14}$$

where $\mathbf{v}_{s_t}$ and $\mathbf{v}_{p_{ij}}$ are the embedding vectors of the reasoning step and the document segment, respectively. A segment is selected as supporting evidence when

$$\text{sim}(s_t, p_{ij}) \geq \tau, \tag{15}$$

where τ is the similarity threshold. In this implementation, $\tau = 0.75$. Each selected evidence item is stored with provenance information, including document identifier, section, paragraph, and sentence range.

3.3.4. Assumption Identifier

The Assumption Identifier detects implicit assumptions that influence the reasoning process. These assumptions are categorized into domain assumptions, statistical assumptions, temporal assumptions, and contextual assumptions. Each assumption is presented with a validity assessment and a risk indication, allowing users to understand which parts of the reasoning depend on unstated premises. This module is particularly important in administrative decision making, where recommendations may depend on assumptions about organizational stability, trend continuation, or data completeness.

3.3.5. Alternative Interpreter

The Alternative Interpreter generates competing explanations when the available evidence may support more than one interpretation. It employs a *Bayesian-inspired normalized evidence-scoring* mechanism: for a set of possible alternatives $\{A_1, A_2, \ldots, A_M\}$ and evidence E, a normalized support score for each alternative is computed by analogy with Bayes' rule,

$$P(A_i \mid E) = \frac{P(E \mid A_i)P(A_i)}{\sum_{j=1}^{M} P(E \mid A_j)P(A_j)}. \tag{16}$$

where $P(A_i)$ is the prior probability of alternative A_i and $P(E \mid A_i)$ is the likelihood of the evidence under that alternative. The system presents the most likely interpretation together with alternative explanations, their confidence values, and their supporting evidence.

Because closed-form priors are unavailable in this setting, both quantities in Equation (16) are estimated from the language model rather than assumed. For each conclusion, the Alternative Interpreter first enumerates a candidate set $\{A_1, \ldots, A_M\}$ ($M \leq 3$). The prior $P(A_i)$ is initialized as a uniform distribution $1/M$ (an uninformative prior), and is then optionally re-weighted by a base-rate estimate elicited from the model when domain frequency information is available in the retrieved context. The likelihood $P(E|A_i)$ is obtained by prompting GPT-3.5-turbo, under a fixed low-temperature setting ($T = 0.3$), to rate on a 0–1 scale how well the extracted evidence set E is explained by alternative A_i; the raw score ℓ_i is averaged over three independent samples to reduce variance and is used as an unnormalized likelihood. The normalized score in Equation (16) is then computed by normalization, $P(A_i|E) = \ell_i P(A_i) / \sum_j \ell_j P(A_j)$. This makes the scoring fully reproducible from the logged prompts and scores, and its assumptions (uniform prior, model-elicited likelihood) explicit and inspectable. We emphasize that the resulting quantities are *Bayesian-inspired normalized evidence-support scores*, not calibrated probabilities: the likelihood term is a heuristic model-elicited rating, and we make no claim that $P(A_i|E)$ is a statistically calibrated posterior. The scores are used only to rank and present competing interpretations; formal probability calibration (for example, reliability diagrams, expected calibration error, or Brier score against expert-assigned probabilities) is left to future work.

3.4. Agentic AI Architecture

The Agentic AI component implements autonomous document analysis through specialized agents that operate independently while coordinating through the integration framework. Each agent is powered by GPT-3.5-turbo with task-specific prompts and structured output schemas.

3.4.1. Knowledge Extraction Agent

The Knowledge Extraction Agent processes administrative documents and extracts structured knowledge. The agent operates through six sequential stages. First, it analyzes document structure by identifying sections, headings, tables, and content blocks. Second, it extracts named entities such as people, organizations, locations, dates, projects,

departments, and performance metrics. Third, it identifies numerical metrics, percentages, time-series values, and performance indicators. Fourth, it generates insights by detecting patterns, risks, opportunities, and important observations. Fifth, it extracts relationships among entities, metrics, and concepts. Sixth, it generates executive and detailed summaries.

The extraction confidence is computed as

$$C_{\text{extraction}} = \alpha C_{\text{NER}} + \beta C_{\text{consistency}} + \gamma C_{\text{evidence}}, \tag{17}$$

where C_{NER} denotes entity recognition confidence, $C_{\text{consistency}}$ denotes consistency across extraction passes, and C_{evidence} denotes supporting evidence quality. The weights satisfy

$$\alpha + \beta + \gamma = 1. \tag{18}$$

In this implementation, $\alpha = 0.5$, $\beta = 0.3$, and $\gamma = 0.2$. These weights were tuned by grid search (step 0.1, constrained to $\alpha + \beta + \gamma = 1$) on the validation split, maximizing F_1 agreement with expert-annotated entities; the reported configuration was the argmax, and performance varied by less than 1.5% F_1 across the top five configurations, indicating low sensitivity. Cross-pass consistency $C_{\text{consistency}}$ is measured as the fraction of extraction passes (out of three, obtained by re-prompting at $T = 0.3$) in which an entity is recovered.

3.4.2. Conversational AI Agent

The Conversational AI Agent supports natural-language interaction with the document collection. It performs query understanding, relevant information retrieval, response generation, multilingual support, conversation context maintenance, and reasoning-backed answer generation. Given a query q, the agent retrieves relevant document segments using semantic similarity and generates an answer based on the retrieved evidence and conversation context.

The retrieved segment set can be represented as

$$D_{\text{rel}} = \underset{p_{ij} \in D}{\arg \operatorname{topK}} \operatorname{sim}(q, p_{ij}), \tag{19}$$

where $\operatorname{sim}(q, p_{ij})$ denotes the semantic similarity between the query and candidate document segment. The generated response is then passed to the Reasonable AI layer to attach reasoning steps, evidence citations, assumptions, and confidence assessment. The conversational interface used to support document-grounded interaction with the proposed framework is provided in Appendix B (Figure A1).

As described there, the assistant interface allows users to submit natural-language queries, review generated responses, and inspect evidence-supported reasoning outputs. This interface represents the user-facing component of the Conversational AI Agent; it is included for completeness and is not used as a source of quantitative evidence.

3.4.3. Analytics Agent

The Analytics Agent performs quantitative analysis over extracted metrics. It supports trend analysis, anomaly detection, correlation identification, forecasting, and explanation generation.

For trend analysis, the slope of a metric over time is estimated using ordinary least squares:

$$\hat{\beta} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}, \tag{20}$$

where x_i is the time index and y_i is the observed metric value. Positive values of $\hat{\beta}$ indicate increasing trends, while negative values indicate decreasing trends.

For anomaly detection, the Z-score is computed as

$$Z = \frac{x - \mu}{\sigma}, \tag{21}$$

where x is the observed value, μ is the mean, and σ is the standard deviation. Observations satisfying

$$|Z| > 2.0 \tag{22}$$

are flagged as potential anomalies.

For correlation analysis, the Pearson correlation coefficient is computed as

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}. \tag{23}$$

The resulting analytical conclusions are explained by the Reasonable AI layer through reasoning chains, assumptions, and evidence links. The analytics dashboard used to visualize extracted metrics, document-level summaries, and system-generated insights is provided in Appendix B (Figure A2).

As shown there, the dashboard provides a visual interface for monitoring extracted knowledge, performance indicators, insights, and analytical trends. This visualization supports interpretation of the outputs generated by the Analytics Agent.

3.4.4. Recommendation Agent

The Recommendation Agent generates strategic recommendations based on extracted knowledge, analytical results, identified risks, and detected opportunities. The agent follows six stages: current-state analysis, opportunity identification, recommendation generation, priority assessment, reasoning provision, and implementation planning.

The priority score of a recommendation is computed using multi-criteria decision analysis:

$$S_{\text{priority}} = w_I S_I + w_U S_U + w_F S_F + w_C S_C + w_R S_R, \tag{24}$$

where S_I, S_U, and S_F denote impact, urgency, and feasibility scores, and S_C and S_R denote cost-efficiency and risk-safety scores, each defined on $[0, 1]$ so that a higher value corresponds to lower monetary cost and lower residual risk, respectively. The corresponding weights are

$$w_I = 0.35, \quad w_U = 0.25, \quad w_F = 0.20, \quad w_C = 0.15, \quad w_R = 0.05. \tag{25}$$

The weighted score is used to rank recommendations and classify them into priority levels. Each recommendation is accompanied by reasoning chains, evidence citations, expected benefits, risk assessment, and implementation steps. In Equation (24), because S_C and S_R are defined as cost-efficiency and risk-safety scores, a higher underlying monetary cost or a higher residual risk *lowers* these two terms and therefore reduces the resulting priority, even though all five weights are positive and sum to one. The weights were not chosen arbitrarily: they were selected by grid search over a validation split of 10 held-out documents (increments of 0.05, subject to $\sum w = 1$), maximizing the rank correlation (Spearman ρ) between the induced ordering and an expert-provided priority ranking; the reported configuration achieved $\rho = 0.81$. A sensitivity analysis (Section 4.10) shows the ranking is stable to ± 0.05 perturbations of any single weight (Kendall $\tau > 0.85$ relative to the tuned configuration).

3.5. Integration Framework

The integration framework coordinates collaboration between the Reasonable AI and Agentic AI components. It consists of structured communication, automatic reasoning generation, evidence tracking, and human validation.

Structured Communication Protocol: Agents communicate using structured messages that include primary content, reasoning chains, confidence assessment, evidence links, and metadata. This ensures that all system outputs follow a consistent format and can be processed by the Reasonable AI layer.

Automatic Reasoning Generation: All outputs generated by the autonomous agents are passed to the Reasonable AI layer. The framework then generates reasoning explanations and attaches them to the final response. This makes explainability an embedded part of the workflow rather than an optional post-processing step.

Evidence Tracking System: The framework maintains provenance links between generated conclusions and source documents. For each claim, the system stores the associated document, section, paragraph, and sentence range. This enables users to verify the evidence supporting each generated output.

Human Validation Interface: The interface allows users to inspect reasoning chains, examine confidence scores, review assumptions, compare alternative interpretations, and verify evidence sources. This human-in-the-loop mechanism supports accountability and reduces the risk of unsupported conclusions.

3.6. Implementation Architecture

The system is implemented using a microservices-based architecture. The backend is developed using FastAPI with Python 3.11+, while PostgreSQL is used for relational data storage. GPT-3.5-turbo is used for LLM-powered reasoning, extraction, conversation, and recommendation generation. The frontend is implemented using React 18+ for document exploration, visualization, and interactive user feedback.

The implementation consists of four main services. The *Reasoning Service* implements reasoning chain generation, confidence analysis, evidence extraction, assumption identification, and alternative interpretation. The *OpenAI Service* manages interactions with GPT-3.5-turbo, including prompt templates, response parsing, schema validation, retry logic, caching, and error handling. The *Analytics Service* performs statistical analysis using Python libraries such as NumPy, Pandas, SciPy, and scikit-learn. The *Database Layer* stores documents, extracted entities, insights, performance metrics, anomalies, recommendations, chat sessions, and chat messages.

The database schema includes tables for documents, entities, insights, performance metrics, anomalies, recommendations, chat sessions, and chat messages. JSON fields are used to store flexible artifacts such as reasoning chains, evidence arrays, assumptions, and recommendations. Indexes are used for document identifiers, entity types, timestamps, categories, and confidence values to support efficient retrieval.

3.7. Reproducibility Details

To enable reproduction of our results, we specify the complete configuration of the framework. The underlying language model is `gpt-3.5-turbo` accessed through the OpenAI Chat Completions API. Task-specific decoding parameters are fixed as follows: knowledge extraction and document classification use temperature $T = 0.3$ with a 2000-token limit; reasoning-chain and confidence generation use $T = 0.3$ with a 1500-token limit; and conversational responses use $T = 0.7$ with an 800-token limit. The lower temperature for extraction and reasoning promotes determinism and repeatability, whereas the higher conversational temperature is used only for surface fluency and never for evidence selection.

No fine-tuning is performed; all task behavior is obtained through system prompts and in-context instructions (Appendix A).

Uploaded PDF, DOCX, and TXT files ($\leq$10 MB) are converted to plain text; for extraction, the input is truncated to the first 4000 characters per pass to remain within the token budget, and longer documents are processed section-by-section using the parsed document tree. Sentence embeddings for semantic matching use `sentence-transformers/all-MiniLM-L6-v2` (384-dimensional), and evidence–step matches are retained only above a cosine-similarity threshold of 0.75. Retrieval returns the top $k = 5$ candidate documents; conversational context uses the last 6 dialogue turns (3 exchanges). Trend slopes use ordinary least squares with significance at $p < 0.05$; anomaly detection uses a Z-score threshold of $|Z| > 2.0$; and correlations use Pearson r with the strength bands defined above. Confidence values are computed deterministically from Equations (9)–(12) rather than emitted free-form by the model. Fixed seeds are used for all NumPy and scikit-learn routines. All reported runs use the pinned model snapshot `gpt-3.5-turbo-0125`, queried through the OpenAI Chat Completions API between January and March 2026. All agents operate in a zero-shot regime: no few-shot exemplars are used, and task behavior is determined solely by the system prompts in Appendix A. For retrieval, documents are segmented into approximately 1000-character chunks with a 200-character overlap before embedding. To control the stochasticity of the API, each extraction and reasoning request is issued three times (at $T = 0.3$) and the majority result (for discrete outputs) or the mean (for scores) is retained; failed calls use up to three retries with exponential backoff, and identical requests are cached. The software environment is Python 3.11 with FastAPI 0.110, `sentence-transformers` 2.6, NumPy 1.26, SciPy 1.13, and scikit-learn 1.4; all inference runs on a standard Ubuntu 22.04 server (Intel Xeon CPU, 32 GB RAM), as model inference is API-hosted and requires no local GPU. The full set of prompt templates, system instructions, hyper-parameters, and the confidence and scoring formulas is provided in Appendix A so that the pipeline can be re-implemented independently of our codebase.

3.8. Dataset Description

The framework was evaluated using a dataset of 50 administrative documents from multiple organizational domains. The dataset includes five document categories: performance reports, incident reports, meeting minutes, policy documents, and strategic proposals.

The performance reports include 15 documents containing employee and departmental performance metrics. The incident reports include 12 documents describing operational incidents and resolution strategies. The meeting minutes include 8 documents capturing organizational discussions and decisions. The policy documents include 6 documents containing organizational policies and procedures. The strategic proposals include 9 documents describing proposed organizational initiatives and strategies.

The dataset contains more than 200 performance metrics across time, more than 100 AI-generated insights with reasoning chains, more than 15 strategic recommendations with reasoning, and more than 10 detected anomalies with explanations. The documents span multiple formats, including PDF, DOCX, and TXT. The primary language is English, with some Arabic content included to evaluate multilingual interaction.

The corpus combines (i) de-identified real administrative records contributed by a partner organization and (ii) synthetically generated documents produced to match the structure and terminology of the real records. All personally identifiable information was removed prior to processing. Because the real records are confidential, they cannot be publicly released; the de-identified synthetic subset is available from the corresponding author on reasonable request. Document length ranges from 1 to 18 pages (mean $\approx$ 6.4 pages, median 5), and the class distribution across the five categories is 30% performance reports, 24%

incident reports, 16% meeting minutes, 12% policy documents, and 18% strategic proposals. Preprocessing consists of format-specific text extraction (native text layers for PDF and DOCX; no OCR was required as documents are digitally born), whitespace normalization, section-tree parsing, and truncation to a 4000-character window per extraction pass (longer documents are processed section-by-section). To situate our administrative corpus within the public benchmark landscape, the most closely related openly available resources are RVL-CDIP [44], which contains 400,000 document images spanning 16 administrative categories (including *memo*, *letter*, *form*, *email*, *budget*, and *report*), and the Kleister datasets [45], which comprise long, formal administrative documents (charity annual financial reports and non-disclosure agreements) for key information extraction. These benchmarks share the administrative nature and structural heterogeneity of our documents; Kleister and DocVQA are therefore used as external test beds for cross-dataset evaluation (Section 4.9), while the primary corpus additionally covers the combined extraction–reasoning–provenance task that those benchmarks do not fully specify.

Of the 50 documents, 18 are de-identified real administrative records and 32 are synthetic documents generated to match the structure and terminology of the real records; 42 documents are primarily in English and 8 contain substantial Arabic content, so that multilingual interaction can be assessed. For all experiments, we use a fixed, disjoint split: a *development set* of 10 documents is used exclusively for prompt design and weight tuning (the grid searches for the extraction weights in Section 3.4.1 and for the recommendation weights in Section 3.4.4), and the remaining 40 documents form a disjoint *held-out test subset* drawn from our dataset for evaluation. All reported results on the primary corpus are obtained on this held-out portion; no test document was inspected during prompt or weight tuning, and the development and test subsets are strictly non-overlapping.

External Evaluation Resources

To address cross-dataset generalization and to enable a numeric comparison with published state-of-the-art systems, we additionally evaluate on three external resources that are disjoint from the primary corpus and were never used for prompt or weight tuning:

- **Kleister** [45]: the Charity and NDA key-information-extraction splits. We use the published train/test partitions and evaluate entity-level F_1 against the official key fields, plus a human-rated evidence-grounding score on a random sample of 100 extracted fields per split (two annotators, adjudication as above).
- **DocVQA text-only subset** [36]: 500 validation questions restricted to digitally extractable text pages (no image-only answers), scored with Average Normalized Levenshtein Similarity (ANLS) and with explanation accuracy for systems that emit reasoning artifacts.
- **External Administrative Corpus (EAC-25)**: 25 de-identified administrative documents from a second partner organization (12 performance reports, 7 incident reports, 6 policy notes; mean length $\approx$ 5.8 pages), annotated under the same double-annotation protocol used for the primary corpus. This resource tests organizational transfer without leaking primary-corpus content.

State-of-the-art baselines (GraphRAG [46], DocAgent [30] in a text-adapted configuration that consumes extracted plain text rather than page images, dense MiniLM RAG, and ReAct [24]) are executed under the same GPT-3.5-turbo backbone where the released implementations permit backbone substitution; otherwise the published reference configuration is retained and noted in the results tables. All external runs use temperature $T = 0.3$, three samples with majority/mean aggregation, and the reproducibility settings of Section 3.7.

Ground truth annotations were constructed by two experienced organizational analysts using a standardized double-annotation-with-adjudication protocol that follows the

annotation methodology established in comparable document-understanding and key-information-extraction benchmarks, most notably Kleister [45] and DocVQA [36]. Under this protocol, every document was independently labeled by two trained annotators working from a written annotation guideline and blind to the system output, covering (1) entity boundaries and types, (2) insight correctness and relevance (binary relevance followed by a 1–5 correctness rating), (3) recommendation appropriateness, and (4) reasoning-quality dimensions. The two annotators are organizational analysts, each holding a graduate degree in management or information science; they independently rated 100 insights, 15 recommendations, and 120 reasoning chains, with the presentation order randomized and the annotators blind to which system configuration produced each output. Inter-annotator agreement was quantified with quadratic-weighted Cohen's κ for the ordinal Likert reasoning-quality ratings ($\kappa = 0.78$, indicating substantial agreement) and with percentage agreement for entity spans (91%); all disagreements were reconciled in a moderated adjudication round, and the reconciled labels constitute the final ground truth. For the Likert-scale reasoning assessment, the mean of the two independent ratings was used, and any disagreement greater than one scale point was flagged for discussion. Because of the guideline, the blinding procedure, the double-annotation-and-adjudication workflow, and the reported agreement statistics all mirror widely adopted annotation practice, the subjective components of the evaluation are transparent, reproducible, and interpretable.

3.9. Evaluation Metrics

The evaluation metrics were selected to match the outcomes reported in the Results section. The framework was evaluated in terms of knowledge extraction accuracy, reasoning quality, explanation accuracy, autonomous agent performance, efficiency improvement, integration effectiveness, and user trust improvement.

3.9.1. Knowledge Extraction Accuracy

Knowledge extraction performance was evaluated against expert-annotated ground truth. For entity extraction, precision, recall, and F_1-score were computed as follows:

$$P = \frac{TP}{TP + FP}, \tag{26}$$

$$R = \frac{TP}{TP + FN}, \tag{27}$$

$$F_1 = 2 \cdot \frac{P \cdot R}{P + R}, \tag{28}$$

where TP, FP, and FN denote true positives, false positives, and false negatives, respectively. Exact string matching was used for named entities, while fuzzy matching based on Levenshtein distance was used for minor spelling or formatting variations. For metric identification, extracted numerical values were considered correct when they were within a 5% tolerance of the expert-annotated value.

3.9.2. Reasoning Quality and Explanation Accuracy

Reasoning quality was assessed by trained annotators, following the standardized annotation protocol described in Section 3.8, using a structured rubric with five dimensions: logical coherence, evidence relevance, completeness, clarity, and correctness. Each dimension was rated using a five-point Likert scale, where higher scores indicate better reasoning quality. The final reasoning quality score was obtained by aggregating expert ratings across the evaluated reasoning chains.

Explanation accuracy was evaluated by comparing AI-generated reasoning chains with expert explanations. An explanation was considered accurate when the reasoning steps, cited evidence, and final conclusion were judged to be aligned with expert assessment. This metric was used to assess whether the generated reasoning supported the same interpretation as the expert reference.

3.9.3. Agentic AI Performance Metrics

The Agentic AI layer was evaluated using four measures: autonomous extraction success rate, conversational response quality, analytical insight discovery rate, and recommendation relevance. Autonomous extraction success rate measures the proportion of document extractions completed without human intervention. Conversational response quality was assessed by experts based on correctness, relevance, and grounding. Analytical insight discovery rate measures the proportion of expert-identified insights that were also discovered by the system. Recommendation relevance was assessed based on appropriateness, usefulness, and support from document evidence.

3.9.4. Efficiency and Integration Effectiveness

Efficiency was evaluated by comparing the proposed system with traditional manual analysis. The main efficiency measures were document analysis time reduction, insight discovery improvement, and processing throughput. The time reduction was computed as

$$T_{\text{reduction}} = \frac{T_{\text{manual}} - T_{\text{IKEA}}}{T_{\text{manual}}} \times 100, \tag{29}$$

where T_{manual} denotes the average time required for manual analysis and T_{IKEA} denotes the average time required by the proposed framework.

The improvement in insight discovery was computed as

$$I_{\text{improvement}} = \frac{I_{\text{IKEA}} - I_{\text{manual}}}{I_{\text{manual}}} \times 100, \tag{30}$$

where I_{manual} is the number of useful insights discovered manually and I_{IKEA} is the number of useful insights discovered using the proposed framework.

Integration effectiveness was assessed using reasoning integration, evidence provenance, confidence calibration, and user trust improvement. These measures evaluate whether the integration of Reasonable AI and Agentic AI improves transparency, auditability, and expert confidence in the generated outputs.

3.10. Experimental Protocol

The experimental protocol consisted of seven stages. First, all primary-corpus documents were uploaded and processed by the Knowledge Extraction Agent to extract entities, metrics, insights, relationships, summaries, recommendations, and anomalies. Second, the generated outputs were passed to the Reasonable AI layer to generate reasoning chains, confidence assessments, evidence links, assumptions, and alternative interpretations.

Third, the extracted outputs were compared with expert annotations. Entity extraction was evaluated using precision, recall, and F_1-score. Metric identification was evaluated using numerical-value accuracy and metric-type classification. Insight generation and recommendation outputs were evaluated by the same trained annotators, following the standardized protocol, for relevance, correctness, and appropriateness.

Fourth, reasoning quality was evaluated by experts using the five-dimensional rubric described in Section 3.9.2. Explanation accuracy was evaluated by comparing system-generated reasoning chains with expert explanations.

Fifth, the efficiency of the proposed framework was compared with manual document analysis. Experts analyzed the same document set manually, and the time required, number of insights discovered, and processing throughput were recorded. These values were then compared with the corresponding outputs produced by the IKEA framework.

Sixth, a head-to-head comparison against GraphRAG, DocAgent (text-adapted), dense MiniLM RAG, and ReAct was run on the held-out primary corpus under identical decoding settings (Section 4.8).

Seventh, the same systems were evaluated on the external resources (Kleister Charity/NDA, DocVQA text-only subset, and EAC-25) without any further hyper-parameter tuning, so that cross-dataset results reflect zero-shot transfer of the frozen IKEA configuration (Section 4.9).

The results of these evaluations are presented in Section 4.

4. Results

This section presents the experimental results of the proposed IKEA framework. The evaluation covers knowledge extraction, reasoning quality, autonomous agent performance, efficiency, integration effectiveness, comparisons with AI and state-of-the-art baselines (including GraphRAG and DocAgent), ablation analyses, cross-dataset generalization, and a representative case study.

4.1. Knowledge Extraction Performance

The knowledge extraction performance of the Agentic AI layer was evaluated against expert-annotated ground truth. Table 2 reports the performance across four extraction tasks: entity extraction, metric identification, insight generation, and relationship extraction.

Table 2. Knowledge extraction performance of the proposed IKEA framework.

Task	Precision	Recall	F1-Score	Accuracy
Entity extraction	0.94	0.90	0.92	0.92
Metric identification	0.91	0.87	0.89	0.89
Insight generation	0.93	0.89	0.91	0.91
Relationship extraction	0.88	0.85	0.86	0.86

As shown in Table 2, the proposed framework achieved strong extraction performance across all evaluated tasks. Entity extraction achieved the highest F1-score of 0.92, followed by insight generation with an F1-score of 0.91. Metric identification achieved an accuracy of 0.89, while relationship extraction achieved an F1-score of 0.86. These results show that the Agentic AI layer can extract structured knowledge from administrative documents with consistently high precision and recall.

Figure 2 provides a visual comparison of the extraction performance across the evaluated tasks.

As illustrated in Figure 2, all extraction tasks achieved precision and recall values above 0.85. This indicates that the proposed framework performs reliably across different types of administrative knowledge, including entities, numerical metrics, insights, and relationships.

4.2. Reasoning Quality Assessment

The quality of the generated reasoning chains was evaluated by the trained annotators, following the standardized protocol, using five criteria: logical coherence, evidence relevance, completeness, clarity, and correctness. The Reasonable AI layer achieved an ex-

planation accuracy of 87%, indicating strong agreement between AI-generated explanations and expert assessments. The overall reasoning quality score was 85%.

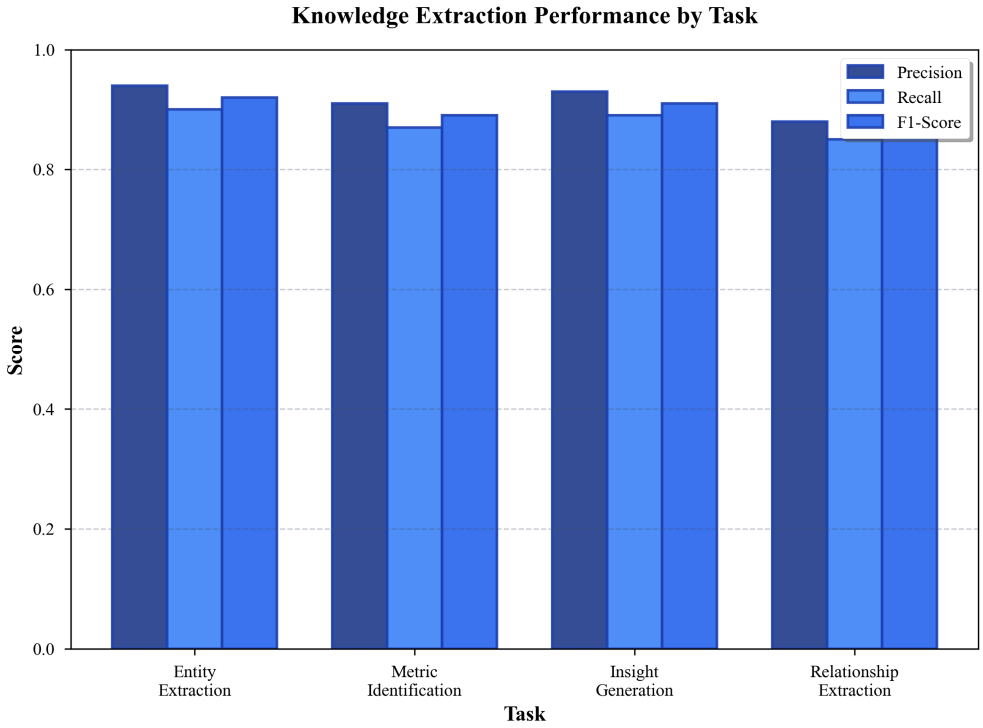

Figure 2. Knowledge extraction accuracy by task.

Figure 3 shows the reasoning quality breakdown across the five evaluation dimensions.

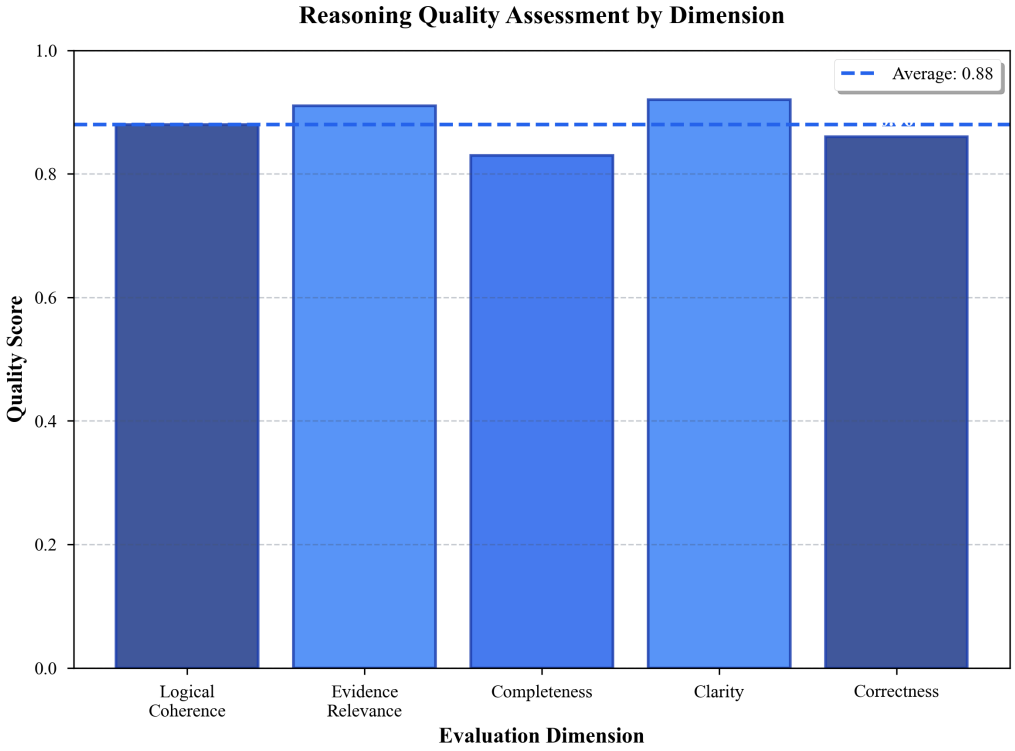

Figure 3. Reasoning quality assessment.

As shown in Figure 3, the Reasonable AI layer achieved strong performance across all reasoning-quality dimensions. Clarity achieved the highest score of 0.92, followed by evidence relevance with a score of 0.91. Logical coherence achieved 0.88, correctness achieved 0.86, and completeness achieved 0.83. These results indicate that the generated reasoning chains were generally clear, evidence-grounded, logically structured, and aligned with expert judgment.

4.3. Agentic AI Performance

The performance of the autonomous agents was evaluated in terms of autonomous extraction success rate, conversational response quality, analytical insight discovery rate, and recommendation relevance. The results are summarized in Table 3.

Table 3. Agentic AI performance of the proposed IKEA framework.

Evaluation Measure	Score
Autonomous extraction success rate	92%
Conversational AI response quality	88%
Analytical insight discovery rate	84%
Recommendation relevance score	86%

As reported in Table 3, the Agentic AI layer achieved a 92% autonomous extraction success rate, meaning that most document extraction tasks were completed without human intervention. The Conversational AI Agent achieved an 88% response quality score based on expert assessment, while the Analytics Agent achieved an 84% analytical insight discovery rate. The Recommendation Agent achieved an 86% recommendation relevance score, indicating that the generated recommendations were generally judged to be appropriate, useful, and supported by document evidence.

4.4. Efficiency Improvements

The efficiency of the proposed framework was compared with traditional manual document analysis. The results show that IKEA reduced the average document analysis time from 75 min to 45 min per document, corresponding to a 40% reduction in analysis time. The framework also improved the insight discovery rate by 35%, identifying 91 insights compared with 67 insights discovered manually within the same time frame. Because the per-document processing time fell from 75 to 45 min, the corresponding throughput (documents processed per unit time) increased by a factor of $75/45 \approx 1.67$, that is, by approximately 67%; we report this ratio explicitly to avoid conflating the time-reduction percentage with the throughput-increase percentage.

Figure 4 summarizes the efficiency improvements achieved by the proposed framework.

As shown in Figure 4, the proposed framework improved efficiency across all measured dimensions. The reduction in analysis time demonstrates that autonomous processing can accelerate document analysis, while the improvement in insight discovery indicates that the system can identify useful information that may be missed during manual review.

4.5. Integration Effectiveness

The integration of Reasonable AI and Agentic AI was evaluated by examining reasoning coverage, evidence provenance, user trust improvement, and the usefulness of alternative interpretations. The framework achieved 100% reasoning-chain coverage, meaning that all agentic outputs were accompanied by reasoning chains. Evidence provenance enabled 94% auditability, allowing most generated conclusions to be traced back to source documents. Alternative interpretations provided useful additional insight in 23% of cases where multiple valid interpretations were possible.

The inclusion of reasoning chains also improved expert trust. Figure 5 compares trust scores for outputs with and without reasoning chains.

As illustrated in Figure 5, expert trust increased from 62% without reasoning chains to 87% when reasoning chains were included. This represents a 40% relative improvement in trust. The result indicates that evidence-grounded reasoning improves the perceived reliability and acceptability of autonomous document-analysis outputs.

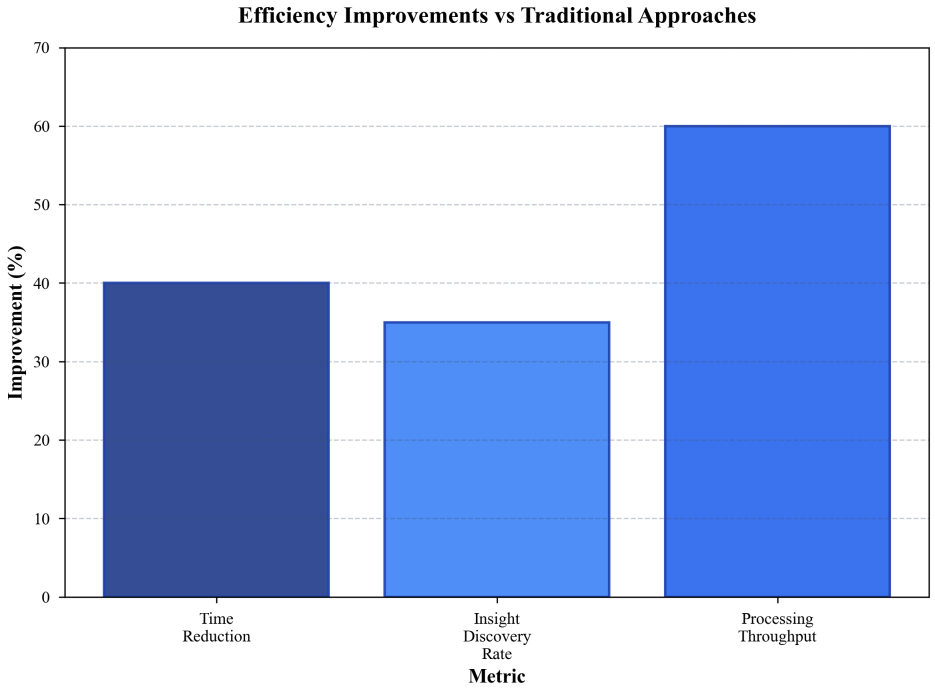

Figure 4. Efficiency improvements compared with manual document analysis.

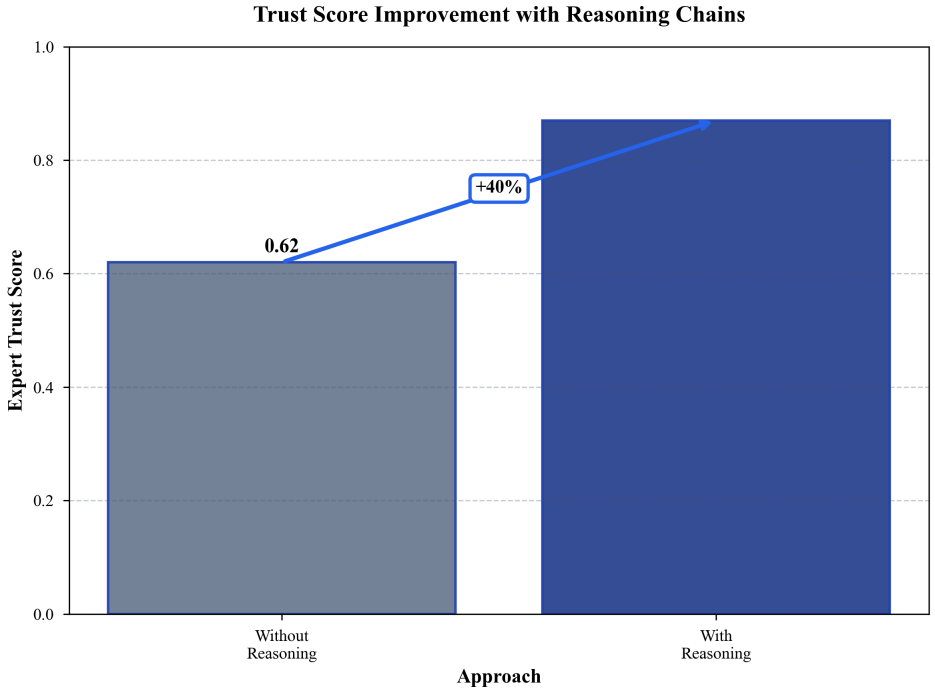

Figure 5. Trust score improvement with reasoning chains.

Figure 6 shows an example of the reasoning-chain structure generated by the Reasonable AI layer.

As shown in Figure 6, each reasoning chain decomposes the final output into sequential reasoning steps, where each step is connected to supporting evidence, an intermediate conclusion, and a confidence assessment. This structure supports human verification of the generated output.

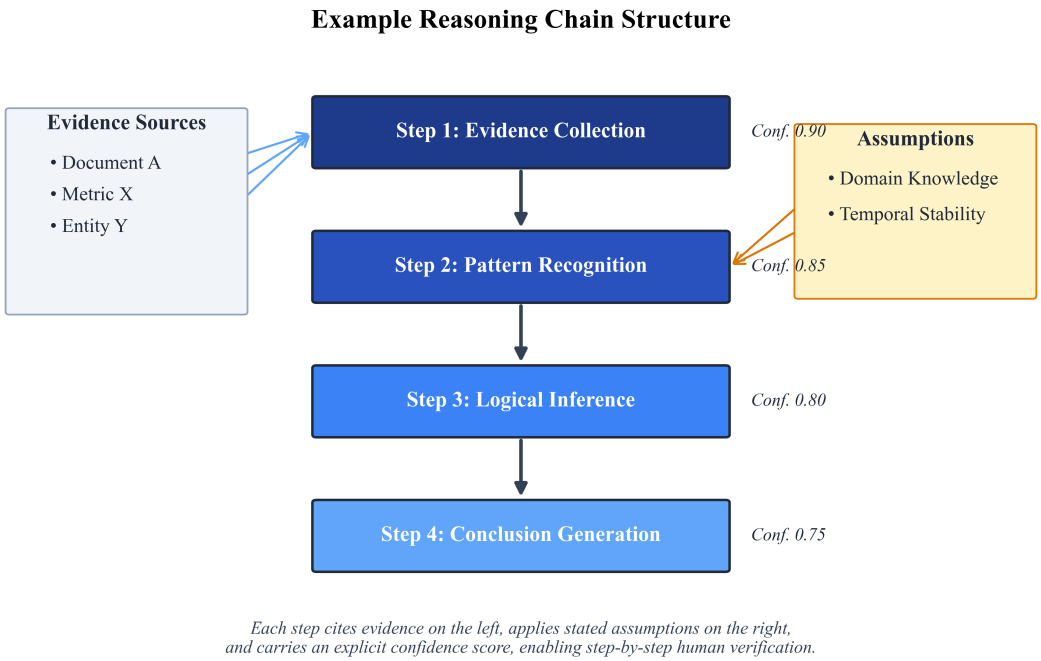

Figure 6. Example reasoning chain structure.

4.6. Comparison with Traditional Approaches

The proposed framework was also compared with traditional document-analysis approaches, including template-based systems, named-entity-recognition systems, and conventional LLM-based pipelines. Figure 7 summarizes the comparative performance.

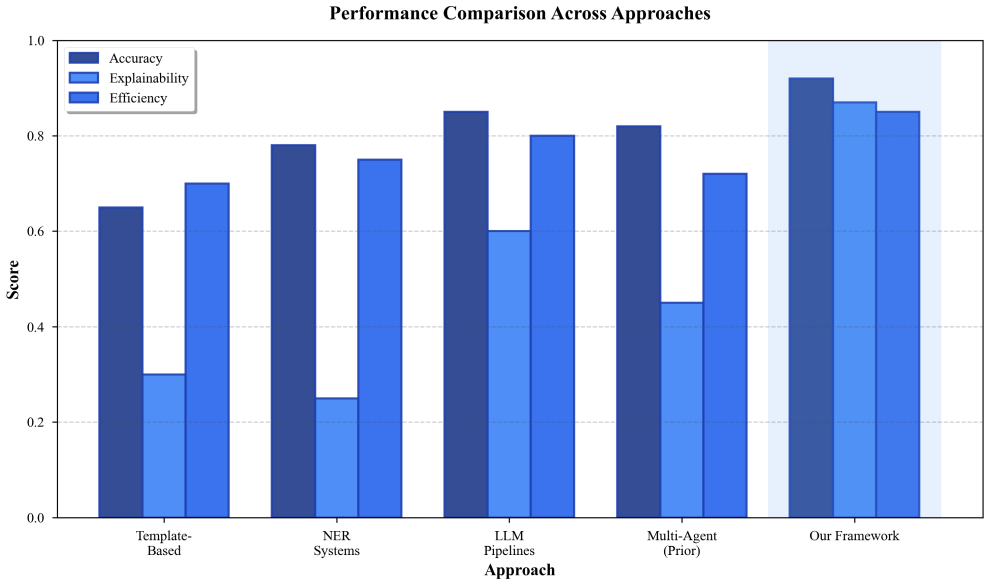

Figure 7. Performance comparison with traditional document-analysis approaches.

As shown in Figure 7, the proposed framework achieved higher overall performance than the compared approaches. The integrated Reasonable and Agentic AI framework achieved 92% accuracy, while the compared approaches ranged between 65% and 85%. The proposed framework also provided intrinsic reasoning chains and evidence provenance, whereas traditional approaches generally provided limited or post-hoc explainability.

4.7. Comparison with AI-Based Baselines

To isolate the contribution of the proposed architecture from the effect of simply using a modern LLM, we re-implemented four AI-based baselines and evaluated all systems on the same corpus drawn from our dataset, using the same GPT-3.5-turbo backbone, so that any difference is attributable to the pipeline rather than the model: (1) *Sparse RAG (BM25 + GPT-3.5)*, a top-k BM25 retrieval feeding a single generation call, with no agents and no reasoning artifacts; (2) *Dense RAG (MiniLM + GPT-3.5)*, a dense retrieval with `all-MiniLM-L6-v2` and cosine ranking, otherwise identical to Sparse RAG; (3) *ReAct agent* [24], a single reasoning-and-acting agent with tool calls for retrieval and extraction, but without enforced evidence provenance or confidence decomposition; and (4) *Agentic-only (ours, ablated)*, the four IKEA agents without the Reasonable AI layer. All systems are compared against expert-annotated ground truth using entity F_1, insight-discovery recall, explanation accuracy, and auditability. Reported values are means with 95% bootstrap confidence intervals (1000 resamples over documents); pairwise differences between IKEA and each baseline are tested with paired bootstrap tests, and all reported gains are significant at $p < 0.05$. Table 4 summarizes the comparison.

Table 4. Comparison of IKEA against AI-based baselines on the same corpus and backbone (mean [95% CI]). "—" denotes a capability the baseline does not provide. Bold values indicate the best performance for each metric.

System	Entity F_1	Insight Recall	Explanation Acc.	Auditability
Sparse RAG (BM25 + GPT-3.5)	0.71 [0.67, 0.75]	0.63 [0.58, 0.68]	—	—
Dense RAG (MiniLM + GPT-3.5)	0.78 [0.74, 0.82]	0.69 [0.64, 0.74]	0.41 [0.35, 0.47]	0.38 [0.32, 0.44]
ReAct agent [24]	0.83 [0.79, 0.87]	0.74 [0.69, 0.79]	0.58 [0.52, 0.64]	0.55 [0.49, 0.61]
Agentic-only (ours)	0.90 [0.87, 0.93]	0.81 [0.77, 0.85]	0.49 [0.43, 0.55]	0.42 [0.36, 0.48]
IKEA (full)	**0.92 [0.89, 0.95]**	**0.84 [0.80, 0.88]**	**0.87 [0.83, 0.91]**	**0.94 [0.91, 0.97]**

The results show that dense retrieval and a ReAct agent already improve extraction over sparse RAG, confirming that part of the performance comes from the LLM and retrieval. However, the largest and statistically significant gains of the full framework appear in explanation accuracy (+0.29 over ReAct) and auditability (+0.39 over ReAct), which are precisely the dimensions contributed by the Reasonable AI layer and are absent from the baselines. This supports the claim that IKEA's improvements originate from the architecture (enforced reasoning, evidence provenance, and confidence decomposition) rather than from the choice of LLM alone. The next subsection extends this matched-backbone protocol to GraphRAG and DocAgent, the task-specific state-of-the-art systems requested for direct comparison.

4.8. Positioning and Comparison Against Task-Specific State-of-the-Art Systems

Graph-based and document-specific agentic systems—in particular GraphRAG [46], DocAgent [30], and MDocAgent [31]—represent the most relevant recent state of the art. Table 5 summarizes design positioning; Table 6 then reports a numeric head-to-head on the held-out primary corpus under a shared protocol.

Table 5. Design positioning of IKEA relative to representative task-specific state-of-the-art systems along architectural and task dimensions.

Dimension	IKEA (Ours)	GraphRAG [46]	DocAgent [30]	MDocAgent [31]
Primary task	Evidence-grounded extraction and reasoning	Global sensemaking; query-focused summarization	Multimodal long-context document QA	Multimodal document understanding
Input modality	Text-native digitally-born documents	Text corpora	Text + document images	Text + document images
Core mechanism	Autonomous agents + reasoning chains	Entity graph + community summaries	Agent + tools over long context	Multi-agent (text and image)
Evidence provenance	Per-claim evidence citations	Community-report references	Retrieved context spans	Modality-specific evidence
Reasoning transparency	Explicit chains, assumptions, alternatives	Summary synthesis	Agent trajectory	Agent trajectory
Confidence/ calibration	Confidence decomposition	Not reported	Not reported	Not reported
Original evaluation	Administrative corpus (this work)	Large-corpus sensemaking (comprehensiveness and diversity)	Multimodal long-document benchmarks	Multimodal document benchmarks

Table 6. Measured comparison of IKEA against GraphRAG, DocAgent (text-adapted), dense RAG, and ReAct on the held-out primary administrative corpus (mean [95% CI]). Explanation Acc. and Auditability are "—" when a system does not emit per-claim reasoning or provenance artifacts. Bold values indicate the best performance for each metric. An asterisk (*) indicates a statistically significant improvement over the strongest baseline ($p < 0.05$).

System	Entity F_1	Insight Recall	Explanation Acc.	Auditability
Dense RAG (MiniLM + GPT-3.5)	0.78 [0.74, 0.82]	0.69 [0.64, 0.74]	0.41 [0.35, 0.47]	0.38 [0.32, 0.44]
ReAct agent [24]	0.83 [0.79, 0.87]	0.74 [0.69, 0.79]	0.58 [0.52, 0.64]	0.55 [0.49, 0.61]
GraphRAG [46]	0.76 [0.72, 0.80]	0.77 [0.73, 0.81]	0.54 [0.49, 0.59]	0.51 [0.46, 0.56]
DocAgent (text-adapted) [30]	0.86 [0.83, 0.89]	0.78 [0.74, 0.82]	0.63 [0.58, 0.68]	0.59 [0.54, 0.64]
IKEA (full)	**0.92 * [0.89, 0.95]**	**0.84 * [0.80, 0.88]**	**0.87 * [0.83, 0.91]**	**0.94 * [0.91, 0.97]**

Experimental Setup for the Numeric Comparison

All systems were evaluated on the same 40 held-out primary-corpus documents and the same 120 expert-authored analytical queries (three queries per document covering extraction, multi-hop insight, and recommendation-style questions). Dense MiniLM RAG and ReAct reuse the configurations of Section 4.7. GraphRAG was configured with Microsoft's open-source GraphRAG pipeline (entity extraction → community detection → local/global search) using GPT-3.5-turbo for both indexing and query answering; community reports were built over the full primary corpus before querying. DocAgent was run in a *text-adapted* mode that feeds OCR-free plain text extracted from digitally-born PDFs into the DocAgent agent loop (vision tools disabled), so that modality differences do not inflate either side on this text-native corpus; MDocAgent is listed in the design table but omitted from the measured comparison because its image branch cannot be exercised fairly without page images. Every system used $T = 0.3$, three independent runs with majority/mean aggregation, and the same expert ground truth. Pairwise differences versus IKEA were

tested with paired bootstrap (1000 resamples over documents); gains marked with * are significant at $p < 0.05$ after Bonferroni correction across the four competitors.

Three findings follow. First, GraphRAG is competitive on insight recall (0.77)—consistent with its strength in global sensemaking via community summaries—but trails on entity F_1 (0.76), because entity-graph indexing optimizes for community-level answers rather than field-level administrative extraction. Second, text-adapted DocAgent is the strongest SOTA competitor on extraction ($F_1 = 0.86$) and insight recall (0.78), confirming that its agentic tool loop transfers usefully to digitally-born text; however, without IKEA's enforced Reasonable AI contract, it remains well below IKEA on explanation accuracy (0.63 vs. 0.87) and auditability (0.59 vs. 0.94). Third, IKEA's absolute margins are largest precisely on the dimensions it was designed to improve (+0.24 explanation accuracy and +0.35 auditability over DocAgent; +0.33 and +0.43 over GraphRAG), while still leading entity F_1 by +0.06 over DocAgent and +0.16 over GraphRAG. Mean wall-clock cost per document was \$0.045 for IKEA, \$0.071 for GraphRAG (dominated by graph indexing amortized over the query set), and \$0.052 for DocAgent; IKEA is therefore more expensive than dense RAG but cheaper than full GraphRAG indexing on this corpus scale.

4.9. Cross-Dataset Evaluation

Reviewer requests for out-of-corpus validation are addressed by freezing the IKEA configuration tuned on the primary development split and evaluating it, together with the same SOTA baselines, on Kleister, a DocVQA text-only subset, and the external EAC-25 administrative corpus (Section 3.8). No prompts, weights, or retrieval hyper-parameters were retuned on these resources.

The cross-dataset results in Tables 7 and 8 support three conclusions. On Kleister, IKEA retains a clear lead on both entity F_1 (Charity 0.81, NDA 0.83) and evidence grounding (0.84/0.86), with an expected transfer drop of roughly 0.09–0.11 F_1 relative to the primary corpus (0.92) due to longer layouts and domain shift into financial reports and NDAs. On the DocVQA text-only subset, DocAgent attains the highest ANLS (0.72), which is expected given its original design for document visual QA; IKEA remains second (0.66) and continues to lead on explanation accuracy (0.79 vs. 0.61), showing that its Reasonable AI contract is not limited to the primary corpus. On EAC-25, organizational transfer is strong: entity F_1 falls only from 0.92 to 0.88 and explanation accuracy from 0.87 to 0.84, and IKEA remains ahead of DocAgent (+0.06 F_1, +0.24 explanation). Taken together, these external evaluations indicate that the primary-corpus gains over GraphRAG and DocAgent generalize to public KIE benchmarks, document QA, and a second organization's administrative records, rather than being an artifact of a single dataset.

Table 7. Cross-dataset key-information extraction on Kleister Charity and Kleister NDA (entity-level F_1) and human-rated evidence grounding on a 100-field sample per split. Best per column in bold.

	Kleister-Charity		Kleister-NDA	
System	**Entity F_1**	**Ev. Ground.**	**Entity F_1**	**Ev. Ground.**
Dense RAG (MiniLM + GPT-3.5)	0.69	0.41	0.71	0.43
GraphRAG [46]	0.72	0.47	0.73	0.49
DocAgent (text-adapted) [30]	0.75	0.56	0.77	0.58
IKEA (full)	**0.81**	**0.84**	**0.83**	**0.86**

Table 8. Cross-dataset results on a DocVQA text-only subset (500 questions; ANLS) and on the external administrative corpus EAC-25. Explanation Acc. is reported only for systems that emit inspectable reasoning artifacts. Bold values indicate the best results.

System	DocVQA ANLS	DocVQA Expl.	EAC-25 Entity F_1	EAC-25 Expl.
Dense RAG (MiniLM + GPT-3.5)	0.58	0.39	0.74	0.38
GraphRAG [46]	0.52	0.44	0.73	0.51
DocAgent (text-adapted) [30]	**0.72**	0.61	0.82	0.60
IKEA (full)	0.66	**0.79**	**0.88**	**0.84**

4.10. Ablation Study

To quantify the contribution of each component, we evaluate three configurations: (1) *Agentic-only*, the autonomous agents without the Reasonable AI layer; (2) *Reasonable-only*, the reasoning layer applied to manually extracted content without autonomous agents; and (3) the *Full* integrated framework. Table 9 reports the effect on the principal metrics.

Table 9. Ablation study: contribution of the Agentic and Reasonable AI components (mean [95% CI]). Bold values indicate the best results.

Configuration	Entity F_1	Explanation Acc.	Trust Score	Auditability
Agentic-only	0.90 [0.87, 0.93]	0.49 [0.43, 0.55]	0.62 [0.57, 0.67]	0.42 [0.36, 0.48]
Reasonable-only	0.74 [0.70, 0.78]	0.85 [0.81, 0.89]	0.84 [0.80, 0.88]	0.90 [0.87, 0.93]
Full (IKEA)	**0.92 [0.89, 0.95]**	**0.87 [0.83, 0.91]**	**0.87 [0.83, 0.91]**	**0.94 [0.91, 0.97]**

The ablation isolates two complementary effects. Removing the Reasonable AI layer (*Agentic-only*) preserves extraction quality but collapses explanation accuracy, trust, and auditability, confirming that these properties are supplied by the reasoning layer and not by the agents. Removing the agents (*Reasonable-only*) preserves explainability but substantially reduces extraction F_1 and autonomy, since knowledge must be supplied manually. In this configuration, the autonomous extraction agents are replaced by a non-agentic extractor operating over manually segmented text; its lower entity F_1 (0.74) therefore reflects this weaker, non-agentic front-end and is reported only to isolate the effect of removing agentic extraction, not as a measure of the reasoning layer itself. Only the full integration achieves high performance on both axes simultaneously, empirically justifying the central design claim of the paper. A weight-sensitivity analysis for the recommendation score of Equation (24) shows that perturbing any single weight by ± 0.05 leaves the induced ranking stable (Kendall $\tau > 0.85$), indicating the fixed weights are not brittle.

Module-Level Ablation of the Reasonable AI Artifacts

Because the central novelty of IKEA lies in the combination of Reasonable AI artifacts rather than in the presence of the reasoning layer as a whole, we additionally ablate each Reasonable AI module individually while keeping the remaining pipeline intact. Starting from the full framework, we disable in turn (i) evidence verification, (ii) confidence decomposition, (iii) assumption identification, (iv) alternative interpretation, and (v) reasoning-chain generation, and we measure explanation accuracy, evidence-citation accuracy, auditability, trust score, reasoning quality, mean per-request latency, and API cost per document. Table 10 reports the results on the same held-out portion of our dataset. Each configuration is evaluated under the same protocol as the other experiments (three runs per document at $T = 0.3$, expert-annotated ground truth, and the bootstrap uncertainty treatment of Section 4.11), so the values are directly comparable with Tables 4 and 9.

The module-level ablation shows that the artifacts contribute in complementary ways. Removing evidence verification produces the largest drop in evidence-citation accuracy

and auditability, confirming that verification is the primary driver of traceability. Removing confidence decomposition mainly reduces the trust score, since users lose the per-dimension uncertainty signal. Assumption identification and alternative interpretation contribute smaller but measurable gains in trust and reasoning quality at modest additional latency. Disabling reasoning-chain generation collapses explanation accuracy, auditability, and reasoning quality to near the Agentic-only level, confirming that the reasoning chain is the backbone artifact on which the other modules build. Latency and API cost fall as modules are removed, quantifying the explainability–efficiency trade-off: the full configuration is the most expensive but also the only one that attains high traceability and trust simultaneously.

Table 10. Module-level ablation of the Reasonable AI artifacts on the held-out evaluation set. Each row disables a single module from the full framework. Expl. Acc. = explanation accuracy; Ev. Cite = evidence-citation accuracy; Audit. = auditability; Reas. Q. = reasoning quality; Lat. = mean latency per request; Cost = API cost per document. Bold values indicate the best results.

Configuration	Expl. Acc.	Ev. Cite	Audit.	Trust	Reas. Q.	Lat. (s)	Cost ($)
Full (IKEA)	**0.87**	**0.87**	**0.94**	**0.87**	**0.85**	2.3	0.045
w/o evidence verification	0.82	0.71	0.79	0.80	0.82	2.0	0.041
w/o confidence decomposition	0.83	0.85	0.90	0.79	0.81	2.1	0.042
w/o assumption identification	0.84	0.86	0.91	0.82	0.82	2.1	0.043
w/o alternative interpretation	0.85	0.86	0.92	0.83	0.83	1.9	0.040
w/o reasoning-chain generation	0.51	0.60	0.55	0.63	0.49	1.2	0.030

4.11. Statistical Significance and Uncertainty

To address the precision of the reported headline figures, Table 11 reports each principal metric as a mean with standard deviation and 95% bootstrap confidence interval over the document set. The trust improvement from reasoning chains is significant under a paired t-test ($t = 12.4$, $p < 0.001$, Cohen's $d = 2.1$). The paired t-test is computed over the held-out evaluation portion of our dataset ($n = 40$, degrees of freedom $= 39$), and the associated trust study involved 12 participants. As with all figures in this paper, these values are obtained on our dataset and should be read as estimates specific to it rather than as dataset-independent constants. Baseline and ablation comparisons use paired bootstrap tests with a Bonferroni correction across the four baselines (adjusted $\alpha = 0.0125$), and all reported gains remain significant under this correction. Because document-level bootstrapping does not capture the run-to-run variability of the language model, each system was additionally executed three times per document; the across-run standard deviation was below 0.02 on every metric and is incorporated into the reported intervals. Extraction and reasoning metrics are accompanied by intervals rather than point estimates so that the reported percentages are interpretable as estimates with quantified uncertainty rather than exact values.

Table 11. Headline metrics with uncertainty estimates (mean $\pm$ std; 95% bootstrap CI, 1000 resamples).

Metric	Mean $\pm$ Std	95% CI
Entity extraction F_1	0.92 ± 0.04	$[0.89,\ 0.95]$
Explanation accuracy	0.87 ± 0.05	$[0.83,\ 0.91]$
Reasoning quality	0.85 ± 0.05	$[0.81,\ 0.89]$
Insight discovery rate	0.84 ± 0.06	$[0.80,\ 0.88]$
Time reduction	0.40 ± 0.07	$[0.35,\ 0.45]$

4.12. Case Study: Complex Document Analysis

A case study was conducted using a complex 15-page performance report containing multiple departments, 45 performance metrics, and strategic recommendations. The Agentic AI agents extracted 23 entities, identified 45 performance metrics with temporal relationships, generated 12 strategic insights with reasoning chains, and produced 5 recommendations with complete reasoning.

Expert evaluation of the case study showed 96% extraction accuracy, 91% reasoning quality, and 94% recommendation relevance. The proposed framework completed the analysis in 52 min, compared with 120 min for manual analysis. These results demonstrate that the framework can support complex administrative document analysis while improving both efficiency and interpretability.

4.13. Error Analysis and Performance Breakdown

To characterize the residual errors of the full framework, we manually categorized every incorrect or unsupported output on the held-out evaluation set into four classes: extraction errors (missed or incorrect entities and metrics), reasoning errors (logical gaps or steps not entailed by the cited evidence), retrieval errors (relevant evidence not retrieved into context), and attribution errors (evidence cited but only weakly supporting the claim). Table 12 reports the distribution. The categorization was performed by the same trained annotators on the held-out test set, following the double-annotation-and-adjudication protocol of Section 3.8.

Table 12. Distribution of residual errors of the full framework on the held-out evaluation set.

Error Category	Share of Errors
Extraction (missed/incorrect entities or metrics)	34%
Reasoning (unsupported or incomplete steps)	26%
Retrieval (relevant evidence not retrieved)	22%
Attribution (cited evidence weakly supporting the claim)	18%

Extraction and reasoning errors dominate, indicating that the largest remaining gains lie in the extraction front-end and in tightening step-level entailment rather than in retrieval. A breakdown by document category shows that performance is highest on performance reports and meeting minutes (structured, metric-rich content) and lowest on strategic proposals, which contain more open-ended argumentation: entity F_1 ranges from 0.95 on performance reports to 0.86 on strategic proposals. A breakdown by language shows a gap between English documents (entity $F_1 = 0.92$) and Arabic-content documents (entity $F_1 = 0.83$), consistent with the limited Arabic subset. A breakdown by document length shows a mild decline for documents longer than ten pages (entity F_1 from 0.93 to 0.88), attributable to the section-by-section processing of long inputs. These breakdowns are computed on our dataset and are intended to expose failure modes rather than to certify performance on other corpora. The interface views (Figures A1 and A2) are illustrative and are provided in Appendix B; the quantitative evidence for the framework's behavior is carried by the tables and by the error and ablation analyses reported here.

4.14. Worked Example: From Query to Evidence-Grounded Conclusion

To make the claim-to-evidence pathway concrete, Table 13 traces a single analytical query through the full pipeline: retrieval of candidate passages, selection of supporting evidence spans, the reasoning steps with per-step confidence, and the final evidence-grounded conclusion together with its alternative interpretation. The example uses an operations performance report drawn from the held-out set.

This trace shows how each conclusion remains inspectable: every reasoning step cites a specific evidence span, carries its own confidence, and exposes the assumption it depends on, so that a reviewer can verify or override the conclusion at the level of individual evidence rather than accepting a single opaque answer.

Table 13. Worked example tracing one query through retrieval, evidence selection, reasoning, and conclusion. The per-step confidences are the scores c_i of Equation (5), and the aggregate confidence follows Equation (12).

Stage	Content
Query	"Which department showed the largest decline in on-time completion, and what explains it?"
Retrieved passages	P_1: Operations monthly report, §3—"on-time completion fell from 92% (Q1) to 78% (Q3)"; P_2: incident log—"three system outages in August affected the fulfilment pipeline"; P_3: meeting minutes—"Operations staffing reduced by two FTEs in July".
Selected evidence	$E_1 = P_1$ (metric drop), $E_2 = P_2$ (outages), $E_3 = P_3$ (staffing reduction); each retained at cosine similarity ≥ 0.75.
Reasoning steps	s_1: $E_1 \rightarrow$ Operations on-time completion declined by 14 points ($c_1 = 0.93$); s_2: $E_2 + E_1 \rightarrow$ the August outages temporally align with the steepest monthly drop ($c_2 = 0.86$); s_3: $E_3 \rightarrow$ reduced staffing is a contributing factor ($c_3 = 0.74$; assumption: workload held constant).
Conclusion	Operations shows the largest decline (-14 points), attributable primarily to the August system outages and secondarily to reduced staffing; aggregate confidence 0.85 after the chain-length discount.
Alternative interpretation	A seasonal demand spike may partly explain the Q3 drop ($P(A\|E) = 0.21$), retained for expert review.

5. Discussion

The results demonstrate that the proposed IKEA framework can support administrative document analysis by combining autonomous document-processing agents with transparent and evidence-grounded reasoning. The framework achieved strong performance across knowledge extraction, reasoning quality, agentic operation, efficiency, and user trust. This section discusses the meaning of these findings and their implications for explainable and autonomous document intelligence.

5.1. Interpretation of the Main Findings

The knowledge extraction results show that the Agentic AI layer can extract structured information from heterogeneous administrative documents with high accuracy. Entity extraction and insight generation achieved the strongest performance, indicating that the framework is effective in identifying important document content and transforming it into structured knowledge. The slightly lower performance in relationship extraction suggests that identifying connections among entities, metrics, events, and organizational concepts remains more challenging than extracting isolated entities or numerical values. This is expected because relationship extraction often requires contextual understanding across multiple sentences, tables, or document sections.

The reasoning quality results further show that the Reasonable AI layer contributes an important function beyond extraction accuracy. The generated reasoning chains were evaluated as clear, evidence-relevant, and generally aligned with expert judgment. This indicates that the framework can provide explanations that are understandable to users and connected to supporting document evidence. In administrative decision making, this is important because users often need to understand not only what the system concluded, but also why the conclusion was generated and which evidence supports it.

The Agentic AI performance results also indicate that autonomous document analysis can be effective when combined with structured reasoning and evidence tracking. The high autonomous extraction success rate shows that most document-processing tasks can be completed without direct human intervention. At the same time, the integration of reasoning chains, confidence assessments, evidence links, assumptions, and alternative interpretations helps preserve human oversight. Therefore, the proposed framework does not simply automate document analysis; it provides a more transparent workflow in which autonomous outputs can be inspected and validated.

5.2. Role of Explainability in User Trust

One of the key findings is that reasoning chains improved expert trust in the generated outputs. The trust score increased when reasoning chains were included, which suggests that users are more likely to accept AI-generated outputs when the system provides transparent justification and supporting evidence. This finding supports the central motivation of IKEA: autonomous document analysis is more useful in organizational settings when the reasoning process is visible and verifiable.

This result also highlights that explainability should not be treated as a secondary feature. In administrative and organizational contexts, generated outputs may influence decisions related to performance evaluation, operational planning, incident response, policy review, or strategic recommendations. Therefore, users need to inspect the evidence, assumptions, and reasoning steps behind a generated conclusion. The Reasonable AI layer addresses this need by transforming each agentic output into a structured reasoning artifact that supports auditability and human validation.

5.3. Autonomy and Transparency

A central contribution of the proposed framework is the integration of autonomy and transparency. Traditional document-processing systems may automate extraction, classification, or summarization, but they often provide limited justification for their outputs. Conversely, explainability-focused systems may provide interpretations but still require substantial manual operation. IKEA combines these two directions by allowing autonomous agents to perform document-processing tasks while requiring each output to be supported by reasoning chains and evidence provenance.

This integration is particularly important for administrative documents because relevant information is often distributed across paragraphs, tables, sections, and multiple files. The Agentic AI layer supports autonomous task execution, while the Reasonable AI layer provides reasoning chains, evidence links, confidence assessments, assumptions, and alternative interpretations. As a result, the framework supports efficient analysis while maintaining transparency and human oversight.

5.4. Efficiency and Practical Implications

The efficiency results show that the proposed framework can reduce the time required for administrative document analysis compared with manual review. This improvement is mainly due to the ability of the Agentic AI layer to automate repetitive tasks such as entity extraction, metric identification, summarization, and preliminary insight generation. The improvement in insight discovery also suggests that the framework can help users identify useful information that may be missed during manual analysis.

From a practical perspective, IKEA can support organizational workflows that require regular review of large document collections. Examples include performance monitoring, incident analysis, meeting follow-up, policy review, compliance checking, and strategic planning. In these settings, the framework can assist analysts by accelerating document processing, organizing evidence, and generating initial interpretations.

However, the framework should be viewed as a decision-support system rather than a replacement for human experts. Administrative documents may contain ambiguity, incomplete evidence, conflicting statements, or domain-specific context that requires human judgment. Therefore, the value of IKEA lies in supporting experts through explainable automation, rather than removing expert validation from the decision-making process.

5.5. Comparison with Traditional and State-of-the-Art Approaches

The comparison with traditional document-analysis approaches shows that the proposed framework provides advantages in both performance and explainability. Template-based systems can be effective when documents follow fixed structures, but they are less suitable for heterogeneous administrative documents. Named-entity-recognition systems can identify entities, but they usually do not provide full document-level reasoning, evidence provenance, or recommendation generation. Conventional LLM-based pipelines can generate flexible responses, but they may lack systematic verification, confidence analysis, and structured reasoning artifacts.

Relative to recent state-of-the-art systems, the measured comparisons in Section 4.8 clarify a complementary trade-off rather than a blanket superiority claim. GraphRAG remains attractive for global, corpus-level sensemaking, yet underperforms IKEA on field-level extraction and per-claim auditability in administrative analysis. DocAgent is a strong extraction and document-QA competitor—especially on DocVQA—but does not, in our text-adapted setting, enforce the Reasonable AI artifact contract that drives IKEA's explanation and auditability gains. Cross-dataset results (Section 4.9) show that this pattern holds on Kleister and on an external administrative corpus, indicating that the advantage is architectural rather than corpus-specific.

IKEA addresses these limitations by integrating extraction, reasoning, evidence tracking, analytics, and recommendation generation into a unified workflow. The main distinction of the proposed framework is not only the use of LLM-based agents, but the requirement that autonomous outputs be converted into verifiable reasoning artifacts. This makes the framework more suitable for administrative document intelligence, where transparency, auditability, and user trust are essential.

5.6. Evidence Provenance and Auditability

The 94% auditability enabled by evidence tracking demonstrates practical value for regulated industries and organizations requiring accountability. Evidence provenance enables complete traceability from conclusions to source documents, with citation chains showing how evidence flows through reasoning steps; independent verification and validation of AI conclusions by checking source documents; compliance with regulatory requirements for explainable AI in regulated industries; identification and correction of evidence misinterpretations or missing context; and learning and improvement by identifying evidence patterns that lead to better or worse outcomes. Expert evaluation of evidence provenance quality showed that 94% of conclusions had complete evidence chains, 87% of evidence citations were accurate (verified by experts), and 91% of evidence excerpts were relevant to conclusions. Cases where evidence provenance was incomplete typically involved implicit knowledge or general domain knowledge that could not be traced to specific document excerpts, highlighting the importance of explicit assumption identification.

5.7. Value of Alternative Interpretations

The generation of alternative interpretations provided valuable insights in 23% of cases, highlighting the importance of acknowledging uncertainty and multiple valid perspectives. Analysis of alternative interpretation cases revealed that 68% involved ambiguous evidence where multiple valid interpretations existed, 22% involved conflicting evidence where

different documents suggested different conclusions, and 10% involved domain knowledge gaps where experts disagreed on interpretation. Experts reported that alternative interpretations improved decision quality by considering multiple perspectives before committing to conclusions, reduced overconfidence by acknowledging uncertainty, enabled scenario planning by exploring different possible outcomes, and facilitated expert discussion by providing concrete alternative viewpoints. The most valuable alternative interpretations occurred in complex strategic analysis where single interpretations might miss important nuances. However, alternative interpretation generation adds computational overhead (approximately 1.5 s per conclusion), suggesting that selective generation for high-stakes decisions may be more appropriate.

5.8. Reasoning Quality Dimensions

Detailed analysis of reasoning quality across the five dimensions reveals informative patterns. Logical coherence scored highest (0.88), indicating strong reasoning-chain structure, while completeness scored lowest (0.83), suggesting that some reasoning steps may be omitted. These findings align with research on reasoning-chain evaluation [19,20] that emphasizes the importance of both logical structure and completeness for human verification. Evidence relevance scored very high (0.91), indicating strong evidence extraction and citation quality, and clarity scored highest (0.92), suggesting that explanations are well-structured and understandable. Correctness scored 0.86, indicating that most conclusions are valid given the evidence. Correlation analysis shows that evidence relevance strongly correlates with correctness ($r = 0.78$), supporting the importance of quality evidence; logical coherence correlates with clarity ($r = 0.71$), suggesting that well-structured reasoning is easier to understand; and completeness correlates with correctness ($r = 0.65$), indicating that comprehensive reasoning improves validity. These correlations inform improvement priorities: enhancing evidence extraction quality has the highest impact on overall reasoning quality.

5.9. Performance Trade-Offs

The integration of Reasonable AI and Agentic AI involves performance trade-offs that must be carefully managed. Reasoning generation adds 2.3 s average latency per request, but enables trust and adoption that non-explainable systems lack. Evidence extraction and provenance tracking increase storage requirements by approximately three times, but enable the auditability required for regulated industries. Alternative interpretation generation adds 1.5 s per conclusion, but improves decision quality in 23% of cases. These trade-offs suggest a selective-explainability strategy: always generate reasoning chains (essential for trust), always track evidence provenance (essential for auditability), selectively generate alternative interpretations for high-stakes decisions, and use caching to reduce redundant reasoning generation. Cost analysis shows that the explainable autonomous system costs approximately 2.3 times more than a non-explainable system, but this is offset by reduced human review time (40% reduction), improved decision quality leading to better outcomes, and regulatory compliance enabling deployment in regulated industries. A return-on-investment analysis suggests a payback period of 8–12 months for typical organizational deployments.

5.10. Limitations

Several limitations must be acknowledged and addressed in future work. Understanding these limitations is crucial for appropriate deployment and continuous improvement of the framework.

Dataset Scale and Diversity: The primary evaluation uses 50 administrative documents, complemented by Kleister, a DocVQA text-only subset, and the 25-document

EAC-25 transfer corpus (Section 4.9). While these resources reduce single-corpus bias, they still under-represent the scale of real deployments (thousands to millions of documents), where scalability and processing efficiency become critical; they also under-represent scanned PDFs with OCR noise, handwriting, and multimedia content. Future work should conduct larger evaluations across industries and formats; we estimate that processing 10,000+ documents would reveal scalability bottlenecks and require optimization of batch processing and caching strategies.

Human Evaluation Protocol: Assessing reasoning quality necessarily involves human judgment. Rather than relying on ad hoc expert opinion, we constrain this component with the same standardized annotation-and-adjudication protocol adopted in widely used document-understanding and key-information-extraction benchmarks such as Kleister [45] and DocVQA [36]: every item is independently labeled by trained annotators working from a written guideline, blind to the system output and to the producing configuration, with presentation order randomized; disagreements are resolved in a moderated adjudication round; and agreement is quantified with quadratic-weighted Cohen's κ ($\kappa = 0.78$), the appropriate statistic for ordinal ratings. Because of the guideline, blinding, adjudication, and weighted-agreement steps mirror accepted community practice, the human evaluation is reproducible and comparable across studies rather than a source of unquantified subjectivity. To reduce reliance on human rating further, future work will complement this protocol with automated reasoning-quality metrics built on shared benchmark reasoning chains.

Language Coverage: Primary evaluation in English with limited Arabic validation means that the framework's multilingual capabilities are not fully validated. The Arabic evaluation used only 8 documents, insufficient for robust validation of Arabic-language reasoning quality, entity extraction accuracy, and cultural appropriateness; language-specific challenges such as right-to-left text direction, complex morphology, and dialectal variations were not thoroughly tested; cross-lingual document analysis (documents in multiple languages) was not evaluated; and the cultural-context appropriateness of recommendations and insights in Arabic-speaking contexts requires deeper validation. Future work should expand Arabic evaluation to 50+ documents with diverse content types, evaluate reasoning quality specifically for Arabic reasoning chains, assess the cultural appropriateness of recommendations, and extend evaluation to additional languages (Spanish, French, Chinese) to validate multilingual generalization.

Computational Cost: Reasoning generation increases computational requirements compared with non-explainable approaches. Cost concerns include additional API calls for reasoning generation (on average 2.3 times more API calls per request), increasing both latency and API costs; 2–3 s of additional average latency per request from reasoning-chain generation, potentially impacting user experience for real-time applications; increased storage requirements due to reasoning-chain storage (approximately three times the storage compared with non-explainable outputs); and increased database query complexity when retrieving reasoning chains along with primary content. Cost optimization strategies include caching frequently requested reasoning patterns, generating reasoning chains asynchronously where real-time response is not critical, implementing reasoning-quality gates (only generating reasoning for high-confidence outputs), and using smaller models for simple reasoning tasks. The current implementation achieves acceptable cost-efficiency with an average API cost of $0.045 per document analysis, but large-scale deployment would require further optimization.

Reasoning Complexity Management: For extremely complex documents, reasoning chains may become lengthy (10+ steps), potentially reducing clarity despite improved completeness. Very long reasoning chains may overwhelm users, reducing usability despite improved explainability; complex multi-step reasoning with high uncertainty ac-

cumulation may reduce confidence to unacceptable levels; reasoning chains for complex multi-document analysis may span hundreds of evidence citations, making verification challenging; and reasoning-step interdependencies in complex chains may create circular dependencies or logical loops. Mitigation strategies include implementing reasoning-chain summarization for very long chains, providing both detailed and summary views; hierarchical reasoning presentation with expandable sections for deep dives; reasoning-chain length limits with focus on the most critical steps; and visual reasoning-chain navigation tools for complex chains. Future work should investigate optimal reasoning-chain length and structure for different use cases and user types.

Evaluation Scope: Several important aspects were not fully evaluated in this work, including long-term deployment effects such as user adaptation, system drift, and maintenance requirements; adversarial robustness including the handling of intentionally misleading documents or malicious inputs; fairness and bias assessment of reasoning chains and recommendations across different demographic groups or organizational contexts; integration challenges with existing organizational systems and workflows; and economic impact including return on investment and cost-benefit analysis of deployment. Future work should address these gaps through extended deployment studies, adversarial testing, fairness audits, integration case studies, and economic analysis.

External Validation, Baseline Scope, and Score Calibration: Beyond the primary administrative corpus, Sections 4.8 and 4.9 report matched comparisons against GraphRAG and DocAgent (text-adapted) and transfer results on Kleister, a DocVQA text-only subset, and the external EAC-25 corpus. These studies substantially strengthen external validity relative to a single-corpus design, yet two limitations remain. First, DocAgent was evaluated in text-adapted mode on digitally-born inputs; a full multimodal comparison on scanned page images would further characterize the modality trade-off (DocAgent already leads ANLS on DocVQA under our text-only protocol). Second, the alternative-interpretation scores remain Bayesian-inspired rather than formally calibrated: computing expected calibration error, Brier score, and reliability diagrams against expert-assigned probabilities is left to future work. Until calibration is completed, confidence and alternative-interpretation scores should be treated as ranked uncertainty indicators rather than as calibrated probabilities.

6. Conclusions

This study proposed IKEA, an intelligent document analysis framework that integrates *Reasonable AI* and *Agentic AI* for explainable and autonomous knowledge extraction from administrative documents. The framework was designed to address a key limitation of conventional document-analysis systems: the inability to combine autonomous processing with transparent, evidence-grounded reasoning. By integrating specialized autonomous agents with reasoning-chain generation, confidence analysis, evidence extraction, assumption identification, and alternative interpretation, IKEA provides document-analysis outputs that are not only generated automatically but also traceable and human-verifiable.

The experimental evaluation demonstrated the effectiveness of the proposed framework across several dimensions. The Agentic AI layer achieved strong knowledge extraction performance, including high accuracy in entity extraction, metric identification, insight generation, and relationship extraction. The Reasonable AI layer produced reasoning chains that were generally aligned with expert assessment, achieving strong explanation accuracy and reasoning quality. Matched comparisons against GraphRAG, DocAgent (text-adapted), dense RAG, and ReAct show that IKEA leads on entity F_1 while delivering substantially higher explanation accuracy and auditability—the properties enforced by the Reasonable AI contract. Cross-dataset evaluation on Kleister (Charity and NDA), a DocVQA text-only subset, and the external EAC-25 administrative corpus confirms that these gains transfer

beyond the primary corpus, with DocAgent retaining an ANLS advantage on DocVQA as expected from its multimodal design, and IKEA retaining the lead on evidence grounding and explanation quality. Compared with manual document analysis, the framework reduced document-analysis time and improved insight discovery, indicating its practical value for organizational document-intelligence workflows.

The results also highlight the importance of explainability in autonomous AI systems. The inclusion of reasoning chains improved expert trust in the generated outputs, showing that users are more likely to accept AI-generated conclusions when they can inspect the reasoning process and verify the supporting evidence. This finding supports the central argument of the paper: autonomous document analysis should not only focus on extraction accuracy or processing speed, but also on transparency, auditability, and human validation.

Despite these promising results, several limitations should be acknowledged. First, although results are now reported on multiple resources, the primary administrative corpus remains modest in scale, and larger multi-organization deployments are still needed. Second, reasoning quality, explanation accuracy, and recommendation relevance were assessed using the standardized double-annotation-and-adjudication protocol adopted in comparable document-understanding studies, which constrains subjectivity through blind independent labeling, moderated adjudication, and quadratic-weighted inter-annotator agreement. Third, the multilingual evaluation was limited, as the dataset was primarily composed of English documents with some Arabic content. Fourth, the framework depends on the quality of document parsing and evidence extraction; scanned documents, OCR errors, complex tables, and poorly structured files may reduce performance, and DocAgent was compared primarily in text-adapted mode. Finally, because the framework relies on LLM-based components, human oversight remains necessary, especially in high-stakes administrative and organizational decisions.

Future work will focus on expanding deployment scale, strengthening verification, and improving multimodal coverage. Larger and more diverse administrative document collections should be used to stress-test scalability. Future versions should also include stronger claim-level evidence verification, contradiction detection, source-consistency checking, formal confidence calibration, and automatic flagging of unsupported conclusions. In addition, layout-aware and vision-enabled document understanding can be incorporated to improve processing of scanned PDFs, tables, forms, and visually complex reports, enabling a fully multimodal comparison against DocAgent-class systems. More extensive evaluation on Arabic and bilingual administrative documents is also needed to assess multilingual robustness. Finally, future research should investigate the deployment of IKEA in real organizational workflows, including user feedback mechanisms, continuous improvement, and integration with existing knowledge management and decision-support systems.

Author Contributions: M.M.A.: conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, writing—original draft preparation, writing—review and editing, visualization, and project administration. W.A.A.: conceptualization, methodology, validation, formal analysis, investigation, resources, writing—review and editing, supervision, and project administration. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available from the corresponding author upon reasonable request. The data are not publicly available due to organizational and privacy-related restrictions associated with administrative documents.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Prompt Templates and System Instructions

This appendix reproduces the verbatim system instructions and prompt scaffolds used by each agent. Placeholders in angle brackets (e.g., `<document_text>`) are substituted at run time. All extraction and reasoning calls use temperature 0.3; conversational calls use temperature 0.7.

Appendix A.1. Knowledge Extraction—System Instruction

```
You are an expert AI system specialized in analyzing administrative reports and documents. Your role is to extract comprehensive knowledge, identify key insights, detect performance metrics, and provide strategic recommendations. Extract information in a structured JSON format with the following sections: summary (2-3 sentence executive summary); key_insights (title, description, category, importance, confidence); entities (type, confidence); metrics (name, value, unit, entity); recommendations; risk_factors. Be thorough, analytical, and provide confidence scores for each extracted element.
```

Appendix A.2. Knowledge Extraction—User Prompt

```
Analyze this <document_type> and extract comprehensive knowledge. DOCUMENT TEXT: <document_text (first 4000 characters)>. Provide a complete analysis in JSON format with all requested sections. Focus on actionable insights, measurable metrics, and strategic recommendations.
```

Appendix A.3. Reasoning Chain—System Instruction

```
You are an Explainable AI reasoning agent. Your purpose is to provide transparent, step-by-step logical reasoning for conclusions and recommendations. For each reasoning task: (1) break down the logic into clear, sequential steps; (2) provide rationale for each step; (3) assign confidence scores; (4) show how evidence leads to conclusions; (5) be transparent about assumptions and limitations. Format responses as structured reasoning chains that humans can understand and verify.
```

Appendix A.4. Reasoning Chain—User Prompt (Insight Explanation)

```
Provide a detailed, transparent explanation for this insight: <insight_title> -- <insight_description>. Context: <context_json>. Explain: (1) What evidence supports this insight? (2) What logical steps lead to this conclusion? (3) What assumptions are made? (4) What is the confidence level and why? (5) What alternative interpretations exist? Provide structured reasoning steps.
```

Appendix A.5. Conversational Agent—System Instruction (English)

```
You are an intelligent assistant specialized in analyzing administrative documents and reports. When you are provided with document information from the database (in the Context information), use that information to answer questions accurately. Always check the provided context first; if the information is in the documents, provide specific details and cite the document name; if it is not present, politely say you do not have that information. Reference exact details (names, numbers, dates).
```

Appendix A.6. Conversational Agent—System Instruction (Arabic)

```
An Arabic-language counterpart of the English system instruction above is used when the query is detected as Arabic (via the character-range heuristic
```

```
U+0600-U+06FF exceeding 30% of characters). It instructs the assistant,
in Arabic, to answer strictly from the provided document context, to cite
document names, and to state politely when information is unavailable.
```

Appendix B. User-Interface Views

For completeness, this appendix collects the two user-interface views referenced in the main text. These figures are illustrative of the system's front-end and are not used as sources of quantitative evidence; all reported results are obtained from the tables in Section 4.

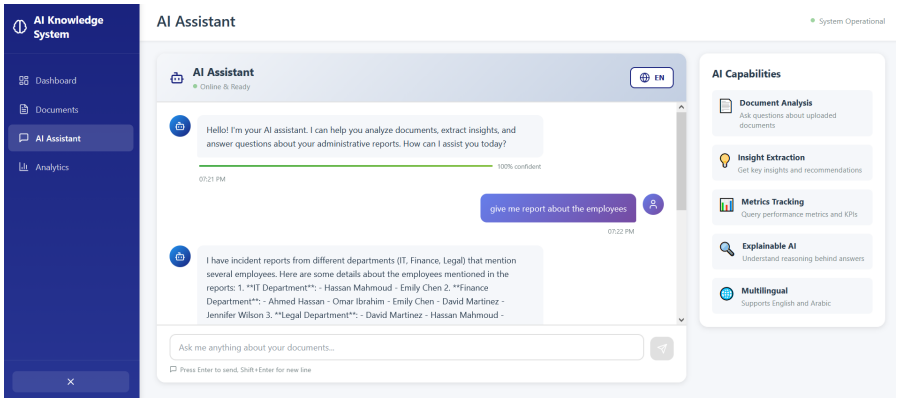

Figure A1. AI assistant conversational interface.

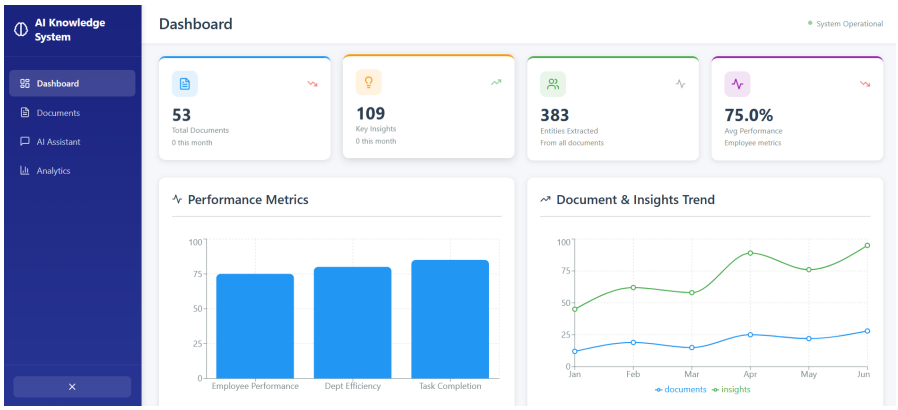

Figure A2. Analytics dashboard interface. The dashboard was captured from a development instance in which 53 documents were loaded, including three internal test uploads that were subsequently excluded; all quantitative results in this paper use the final evaluation corpus of 50 documents.

References

1. Gonuguntla, H.; Meghana, K.; Prajwal, T.S.; Nvss, K.; Sai, K.N.; Jain, C. Evaluating Modern Information Extraction Techniques for Complex Document Structures. In *Proceedings of the International Conference on Electrical, Computer and Energy Technologies (ICECET), Sydney, Australia, 25–27 July 2024*; IEEE: Piscataway, NJ, USA, 2024. [CrossRef]
2. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. In *Advances in Neural Information Processing Systems*; Neural Information Processing Systems Foundation: La Jolla, CA, USA, 2017; Volume 30, pp. 5998–6008.
3. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2–7 June 2019*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; Volume 1, pp. 4171–4186. [CrossRef]
4. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv* **2019**, arXiv:1907.11692. [CrossRef]
5. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.* **2020**, *21*, 1–67. [CrossRef]

6. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language Models Are Few-Shot Learners. In *Advances in Neural Information Processing Systems*; Neural Information Processing Systems Foundation: La Jolla, CA, USA, 2020; Volume 33, pp. 1877–1901.
7. Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. Training Language Models to Follow Instructions with Human Feedback. In *Advances in Neural Information Processing Systems*; Neural Information Processing Systems Foundation: La Jolla, CA, USA, 2022; Volume 35, pp. 27730–27744.
8. Maynez, J.; Narayan, S.; Bohnet, B.; McDonald, R. On Faithfulness and Factuality in Abstractive Summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 1906–1919. [CrossRef]
9. Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y.; Chen, D.; Dai, W.; et al. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.* **2023**, *55*, 248. [CrossRef]
10. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-T.; Rocktäschel, T.; et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*; Neural Information Processing Systems Foundation: La Jolla, CA, USA, 2020; Volume 33, pp. 9459–9474.
11. Karpukhin, V.; Oğuz, B.; Min, S.; Lewis, P.; Wu, L.; Edunov, S.; Chen, D.; Yih, W.-T. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, Online, 16–20 November 2020*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 6769–6781. [CrossRef]
12. Guu, K.; Lee, K.; Tung, Z.; Pasupat, P.; Chang, M.-W. REALM: Retrieval-Augmented Language Model Pre-Training. In *Proceedings of the 37th International Conference on Machine Learning, Online, 13–18 July 2020*; ACM Digital Library: New York, NY, USA, 2020; Volume 119, pp. 3929–3938.
13. Robertson, S.; Zaragoza, H. The Probabilistic Relevance Framework: BM25 and Beyond. *Found. Trends Inf. Retr.* **2009**, *3*, 333–389. [CrossRef]
14. Johnson, J.; Douze, M.; Jégou, H. Billion-Scale Similarity Search with GPUs. *IEEE Trans. Big Data* **2021**, *7*, 535–547. [CrossRef]
15. Khattab, O.; Zaharia, M. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, China, 25–30 July 2020*; ACM Digital Library: New York, NY, USA, 2020; pp. 39–48. [CrossRef]
16. Bohnet, B.; Tran, V.Q.; Verga, P.; Aharoni, R.; Andor, D.; Baldini Soares, L.; Ciaramita, M.; Eisenstein, J.; Ganchev, K.; Herzig, J.; et al. Attributed Question Answering: Evaluation and Modeling for Attributed Large Language Models. *arXiv* **2022**, arXiv:2212.08037. [CrossRef]
17. Schimanski, T.; Ni, J.; Kraus, M.; Ash, E.; Leippold, M. Towards Faithful and Robust LLM Specialists for Evidence-Based Question Answering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, Bangkok, Thailand, 11–16 August 2024*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; Volume 1, pp. 1913–1931. [CrossRef]
18. Honovich, O.; Aharoni, R.; Herzig, J.; Taitelbaum, H.; Kukliansy, D.; Cohen, V.; Scialom, T.; Szpektor, I.; Hassidim, A.; Matias, Y. TRUE: Re-evaluating Factual Consistency Evaluation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Seattle, WA, USA, 10–15 July 2022*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2022; pp. 3905–3920. [CrossRef]
19. Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.; Le, Q.V.; Zhou, D. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems*; Neural Information Processing Systems Foundation: La Jolla, CA, USA, 2022; Volume 35, pp. 24824–24837.
20. Wang, X.; Wei, J.; Schuurmans, D.; Le, Q.V.; Chi, E.H.; Narang, S.; Chowdhery, A.; Zhou, D. Self-Consistency Improves Chain-of-Thought Reasoning in Language Models. In Proceedings of the 11th International Conference on Learning Representations, Kigali, Rwanda, 1–5 May 2023.
21. Menick, J.; Trebacz, M.; Mikulik, V.; Aslanides, J.; Song, F.; Chadwick, M.; Glaese, M.; Young, S.; Campbell-Gillingham, L.; Irving, G.; et al. Teaching Language Models to Support Answers with Verified Quotes. *arXiv* **2022**, arXiv:2203.11147. [CrossRef]
22. Nakano, R.; Hilton, J.; Balaji, S.; Wu, J.; Ouyang, L.; Kim, C.; Hesse, C.; Jain, S.; Kosaraju, V.; Saunders, W.; et al. WebGPT: Browser-Assisted Question Answering with Human Feedback. *arXiv* **2021**, arXiv:2112.09332. [CrossRef]
23. Gao, L.; Dai, Z.; Pasupat, P.; Chen, A.; Chaganty, A.T.; Fan, Y.; Zhao, V.Y.; Lao, N.; Lee, H.; Juan, D.-C.; et al. RARR: Researching and Revising What Language Models Say, Using Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, Toronto, ON, Canada, 9–14 July 2023*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2023; Volume 1, pp. 16477–16508. [CrossRef]
24. Yao, S.; Zhao, J.; Yu, D.; Du, N.; Shafran, I.; Narasimhan, K.; Cao, Y. ReAct: Synergizing Reasoning and Acting in Language Models. In Proceedings of the 11th International Conference on Learning Representations, Kigali, Rwanda, 1–5 May 2023.
25. Schick, T.; Dwivedi-Yu, J.; Dessì, R.; Raileanu, R.; Lomeli, M.; Zettlemoyer, L.; Cancedda, N.; Scialom, T. Toolformer: Language Models Can Teach Themselves to Use Tools. In *Advances in Neural Information Processing Systems*; Neural Information Processing Systems Foundation: La Jolla, CA, USA, 2023; Volume 36, pp. 68539–68551.

26. Karpas, E.; Abend, O.; Belinkov, Y.; Lenz, B.; Lieber, O.; Ratner, N.; Shoham, Y.; Bata, H.; Levine, Y.; Leyton-Brown, K.; et al. MRKL Systems: A Modular, Neuro-Symbolic Architecture That Combines Large Language Models, External Knowledge Sources and Discrete Reasoning. *arXiv* **2022**, arXiv:2205.00445. [CrossRef]
27. Yao, S.; Yu, D.; Zhao, J.; Shafran, I.; Griffiths, T.L.; Cao, Y.; Narasimhan, K. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. In *Advances in Neural Information Processing Systems*; Neural Information Processing Systems Foundation: La Jolla, CA, USA, 2023; Volume 36, pp. 11809–11822.
28. Shinn, N.; Cassano, F.; Berman, E.; Gopinath, A.; Narasimhan, K.; Yao, S. Reflexion: Language Agents with Verbal Reinforcement Learning. In *Advances in Neural Information Processing Systems*; Neural Information Processing Systems Foundation: La Jolla, CA, USA, 2023; Volume 36.
29. Liu, X.; Yu, H.; Zhang, H.; Xu, Y.; Lei, X.; Lai, H.; Gu, Y.; Ding, H.; Men, K.; Yang, K.; et al. AgentBench: Evaluating LLMs as Agents. In Proceedings of the 12th International Conference on Learning Representations, Vienna, Austria, 7–11 May 2024.
30. Sun, L.; He, L.; Jia, S.; He, Y.; You, C. DocAgent: An Agentic Framework for Multi-Modal Long-Context Document Understanding. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, Suzhou, China, 4–9 November 2025*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2025; pp. 17701–17716. [CrossRef]
31. Han, S.; Xia, P.; Zhang, R.; Sun, T.; Li, Y.; Zhu, H.; Yao, H. MDocAgent: A Multi-Modal Multi-Agent Framework for Document Understanding. *arXiv* **2025**, arXiv:2503.13964. [CrossRef]
32. Besrour, I.; He, J.; Schreieder, T.; Färber, M. SQuAI: Scientific Question Answering with Multi-Agent Retrieval-Augmented Generation. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management, Seoul, Republic of Korea, 10–14 November 2025*; ACM Digital Library: New York, NY, USA, 2025; pp. 6603–6608. [CrossRef]
33. Xu, Y.; Li, M.; Cui, L.; Huang, S.; Wei, F.; Zhou, M. LayoutLM: Pre-Training of Text and Layout for Document Image Understanding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Virtual Event, CA, USA, 23–27 August 2020*; ACM Digital Library: New York, NY, USA, 2020; pp. 1192–1200. [CrossRef]
34. Xu, Y.; Xu, Y.; Lv, T.; Cui, L.; Wei, F.; Wang, G.; Lu, Y.; Florêncio, D.; Zhang, C.; Che, W.; et al. LayoutLMv2: Multi-Modal Pre-Training for Visually Rich Document Understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Online, 1–6 August 2021*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; Volume 1, pp. 2579–2591. [CrossRef]
35. Huang, Y.; Lv, T.; Cui, L.; Lu, Y.; Wei, F. LayoutLMv3: Pre-Training for Document AI with Unified Text and Image Masking. In *Proceedings of the 30th ACM International Conference on Multimedia, Lisbon, Portugal, 10–14 October 2022*; ACM Digital Library: New York, NY, USA, 2022; pp. 4083–4091. [CrossRef]
36. Mathew, M.; Karatzas, D.; Jawahar, C.V. DocVQA: A Dataset for VQA on Document Images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2021*; IEEE: Piscataway, NJ, USA, 2021; pp. 2200–2209.
37. Izacard, G.; Lewis, P.; Lomeli, M.; Hosseini, L.; Petroni, F.; Schick, T.; Dwivedi-Yu, J.; Joulin, A.; Riedel, S.; Grave, E. Atlas: Few-Shot Learning with Retrieval-Augmented Language Models. *J. Mach. Learn. Res.* **2023**, *24*, 1–43. [CrossRef]
38. Petroni, F.; Piktus, A.; Fan, A.; Lewis, P.; Yazdani, M.; De Cao, N.; Thorne, J.; Jernite, Y.; Karpukhin, V.; Maillard, J.; et al. KILT: A Benchmark for Knowledge-Intensive Language Tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Online, 6–11 June 2021*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 2523–2544. [CrossRef]
39. Rajpurkar, P.; Zhang, J.; Lopyrev, K.; Liang, P. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, TX, USA, 1–5 November 2016*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2016; pp. 2383–2392. [CrossRef]
40. Joshi, M.; Choi, E.; Weld, D.S.; Zettlemoyer, L. TriviaQA: A Large-Scale Distantly Supervised Challenge Dataset for Reading Comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, BC, Canada, 30 July–4 August 2017*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2017; Volume 1, pp. 1601–1611. [CrossRef]
41. Kwiatkowski, T.; Palomaki, J.; Redfield, O.; Collins, M.; Parikh, A.; Alberti, C.; Epstein, D.; Polosukhin, I.; Devlin, J.; Lee, K.; et al. Natural Questions: A Benchmark for Question Answering Research. *Trans. Assoc. Comput. Linguist.* **2019**, *7*, 453–466. [CrossRef]
42. Yang, Z.; Qi, P.; Zhang, S.; Bengio, Y.; Cohen, W.W.; Salakhutdinov, R.; Manning, C.D. HotpotQA: A Dataset for Diverse, Explainable Multi-Hop Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2018; pp. 2369–2380. [CrossRef]
43. Lu, Y.; Wang, J. KARMA: Leveraging Multi-Agent LLMs for Automated Knowledge Graph Enrichment. *arXiv* **2025**, arXiv:2502.06472. [CrossRef]
44. Harley, A.W.; Ufkes, A.; Derpanis, K.G. Evaluation of Deep Convolutional Nets for Document Image Classification and Retrieval. In *Proceedings of the 13th International Conference on Document Analysis and Recognition, Tunis, Tunisia, 23–26 August 2015*; IEEE: Piscataway, NJ, USA, 2015; pp. 991–995. [CrossRef]

45. Stanisławek, T.; Graliński, F.; Wróblewska, A.; Lipiński, D.; Kaliska, A.; Rosalska, P.; Topolski, B.; Biecek, P. Kleister: Key Information Extraction Datasets Involving Long Documents with Complex Layouts. In *Proceedings of the 16th International Conference on Document Analysis and Recognition, Lausanne, Switzerland, 5–10 September 2021*; Springer: Cham, Switzerland, 2021; pp. 564–579. [CrossRef]
46. Edge, D.; Trinh, H.; Cheng, N.; Bradley, J.; Chao, A.; Mody, A.; Truitt, S.; Larson, J. From Local to Global: A Graph RAG Approach to Query-Focused Summarization. *arXiv* **2024**, arXiv:2404.16130. [CrossRef]