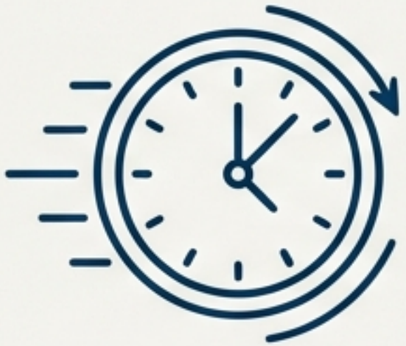
Architecting the Automated Scientific Auditor

A Multi-Modal Pipeline for Document Quality Assessment

Nougat OCR | Transformer GEC | CNN-BiLSTM Classification | T5 Fact-Checking

The Manual Review Process is Overwhelmed



Arduous & Slow

Manual review is a bottleneck in scientific publishing, prone to delays.

Inconsistent

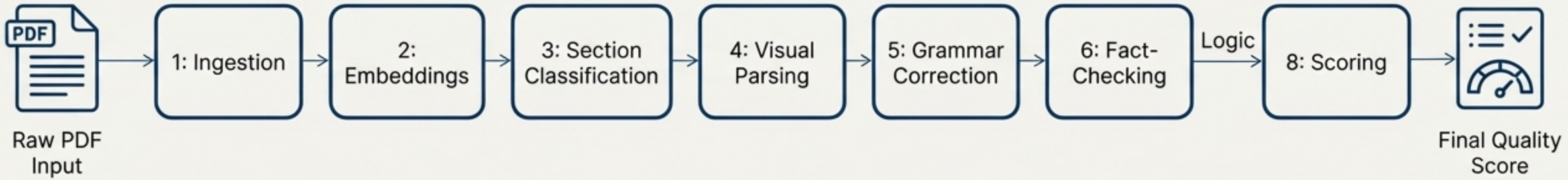
Human evaluation is subject to individual bias, leading to inconsistent quality standards.

Structurally Opaque

Complex visual layouts (LaTeX, tables, figures) are difficult for standard text parsers to understand, losing critical information.

"The rapid proliferation of scientific literature necessitates the development of automated tools for content verification, structural analysis, and linguistic auditing."

The Solution: A Modular, 8-Phase Document Auditor

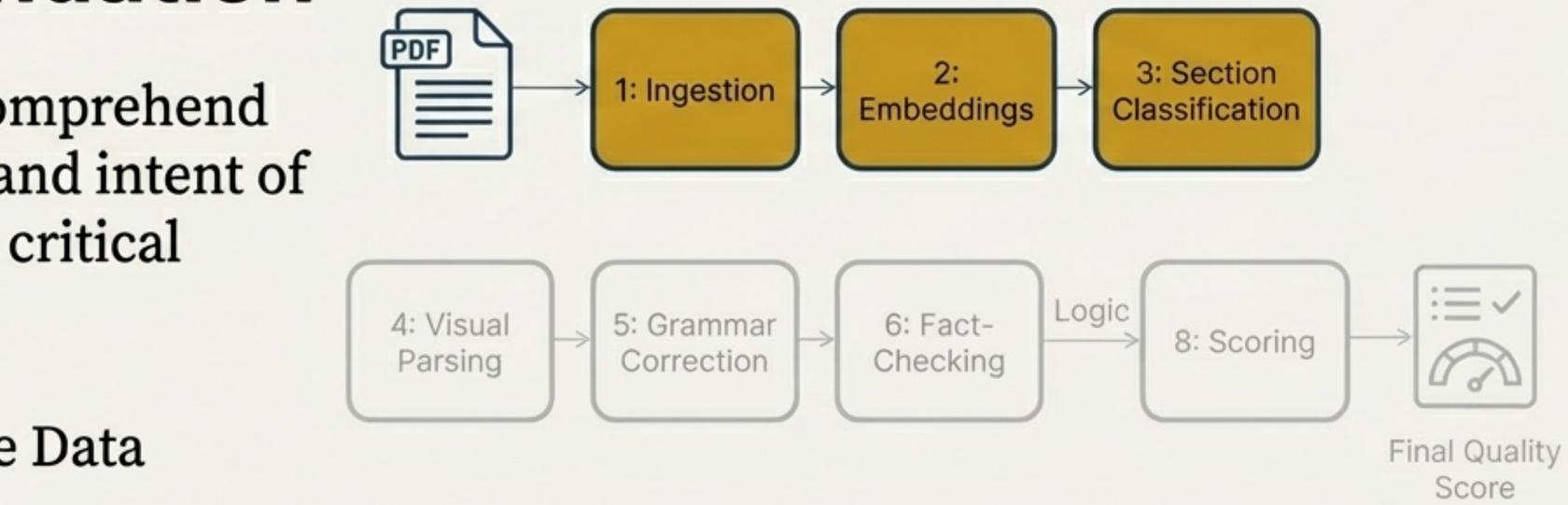


Built on the principle of **Modular Intelligence**, each phase is a specialized solution to a distinct challenge in document understanding.

Pillar I: Building the Semantic Foundation

To teach the system to comprehend the language, structure, and intent of scientific text before any critical analysis can occur.

- Phase 1: Multi-Source Data Ingestion
- Phase 2: Tiered Semantic Embeddings
- Phase 3: Deep Sequential Section Classification



From Raw Text to Structural Understanding

Learning from the “Gold Standard” (Phases 1 & 2)

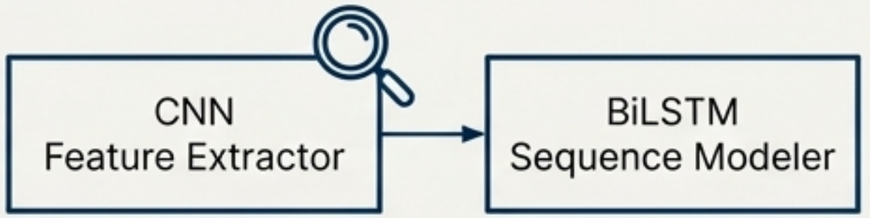
Corpus
Ingested ASAP-Review and PeerRead datasets.

Embedding Strategy: A tiered approach for comprehensive understanding.

- 👉 **FastText**: Handles out-of-vocabulary (OOV) terms common in science. (Key params: `sg=0 (CBOW)`, `vector_size=100`).
- 👉 **Sentence-BERT**: Generates 384-dimensional contextual embeddings (`all-MiniLM-L6-v2`) to grasp paragraph-level intent.

Classifying Scientific Discourse (Phase 3)

The Challenge: Distinguishing between sections with similar vocabulary, like “Introduction” and “Related Work.”



A Hybrid CNN-BiLSTM model.

Why a CNN? Acts as a feature extractor for key n-grams (e.g., “we propose,” “state-of-the-art”).

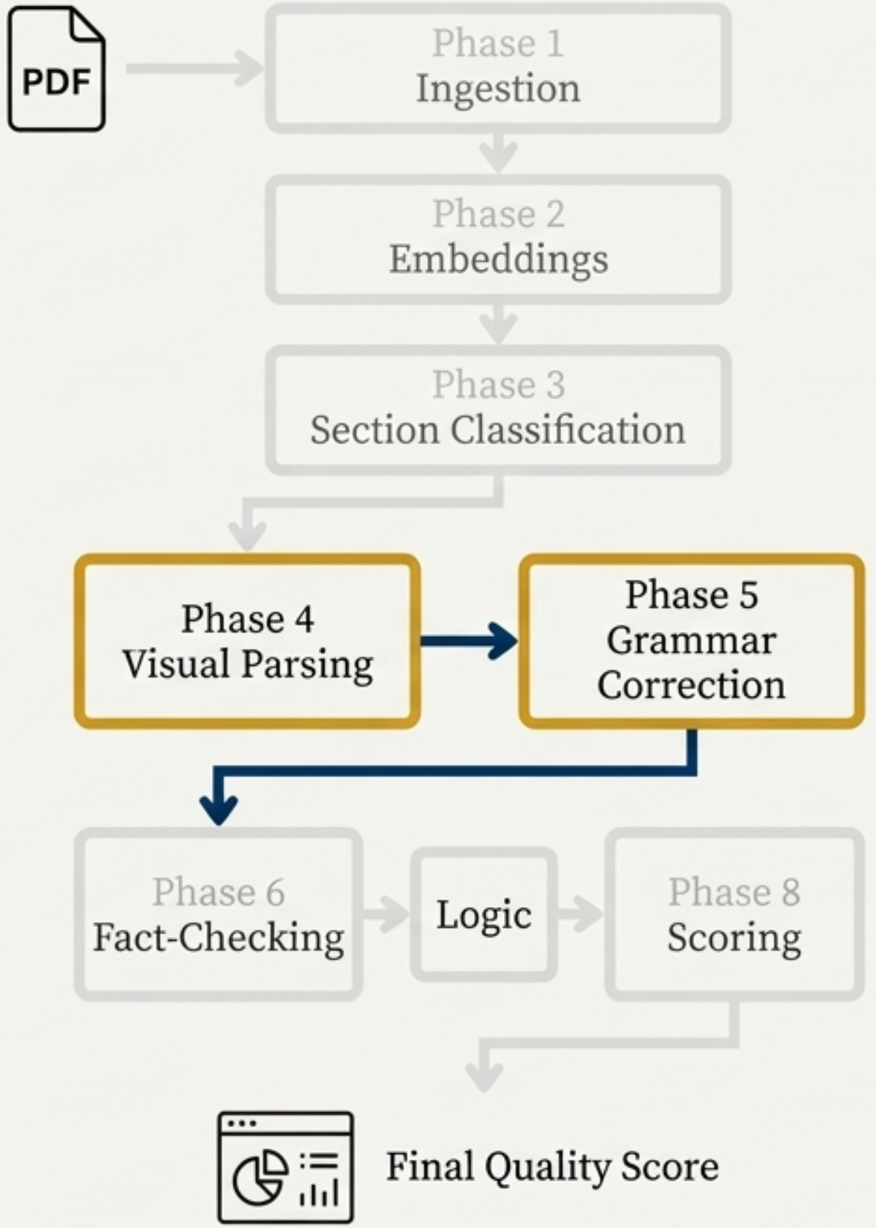
Why a BiLSTM? Models the sequential flow and context of the entire section.

Key Detail: Class imbalance was addressed via oversampling to a target size of 3000 per label, preventing bias towards common sections like “Methodology.”

Pillar II: Achieving Visual and Textual Fidelity

To accurately translate complex visual documents (PDFs with LaTeX, tables, equations) into pristine, grammatically correct text suitable for analysis.

- Phase 4: Visual-to-Semantic Parsing Parsing
- Phase 5: Generative Grammar Error Correction (GEC)



Translating Pixels and Prose

Beyond Standard OCR (Phase 4)

The Tool: Nougat (`facebook/nougat-base`)

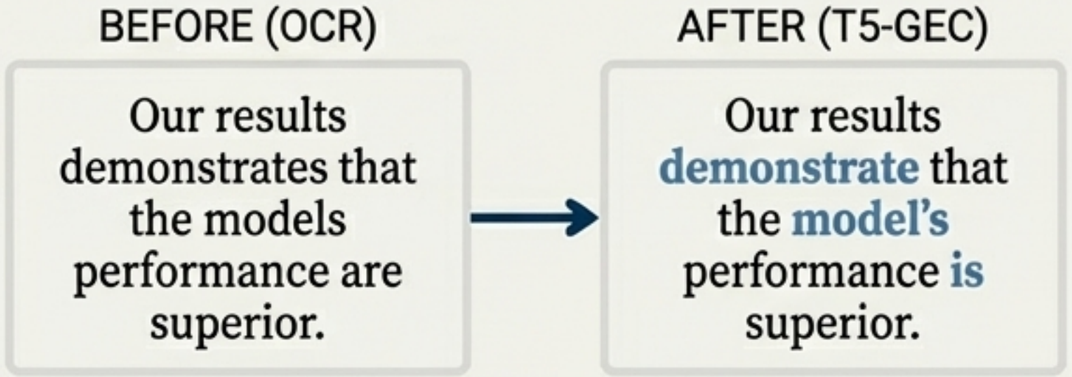
The Advantage: Unlike parsers like PyPDF2 that fail on complex layouts, Nougat treats the document as an image, decoding its structure into Markdown. This preserves tables, equations, and multi-column formats.

Implementation Note: `VisionEncoderDecoderModel` with a greedy decoding strategy was used for stable and reliable output.

Refining the Extraction (Phase 5)

The Problem: Raw OCR output contains artifacts, hallucinations, and non-native linguistic patterns.

The Solution: A T5-base model fine-tuned on the JFLEG dataset for Grammatical Error Correction.

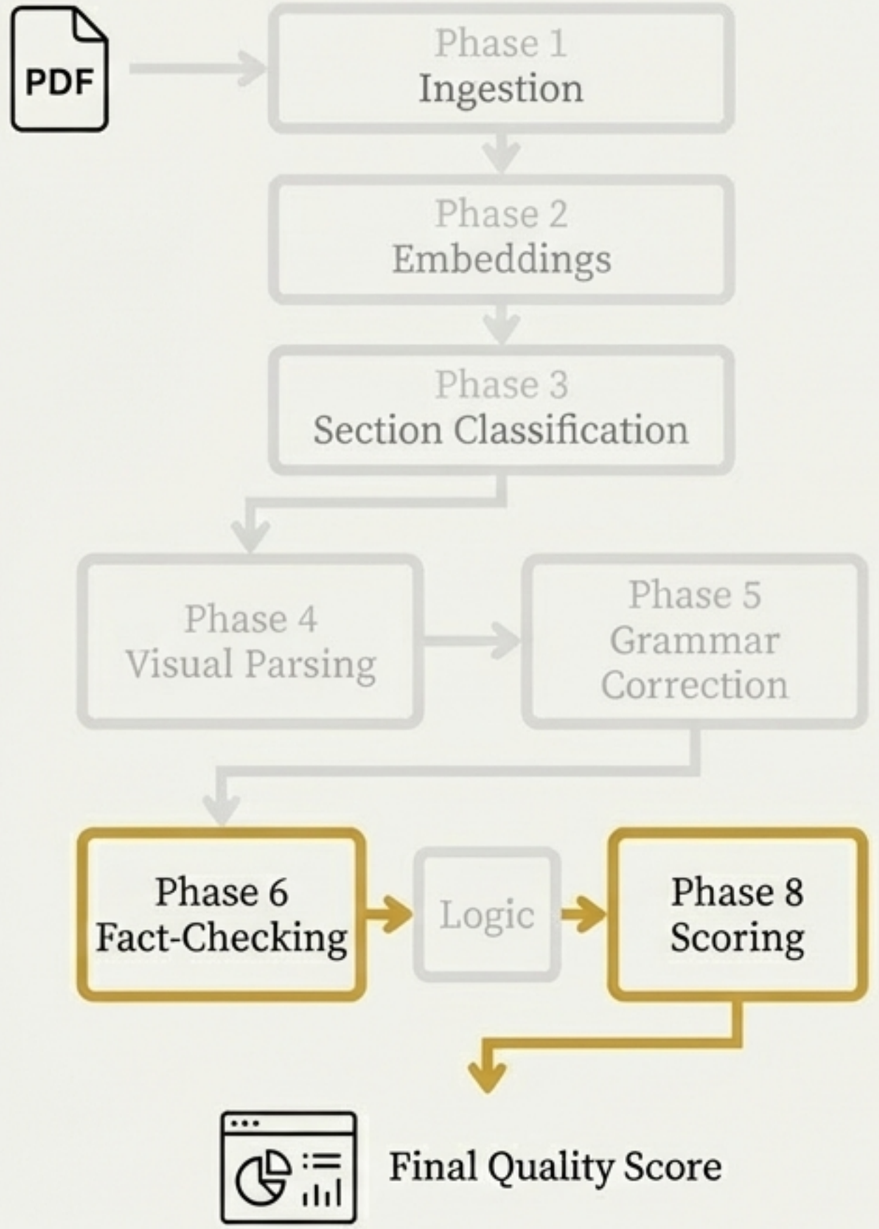


Advanced Technique: Employed "Multi-Reference Supervision," where the model learns from multiple valid human corrections for the same error, increasing its robustness and flexibility.

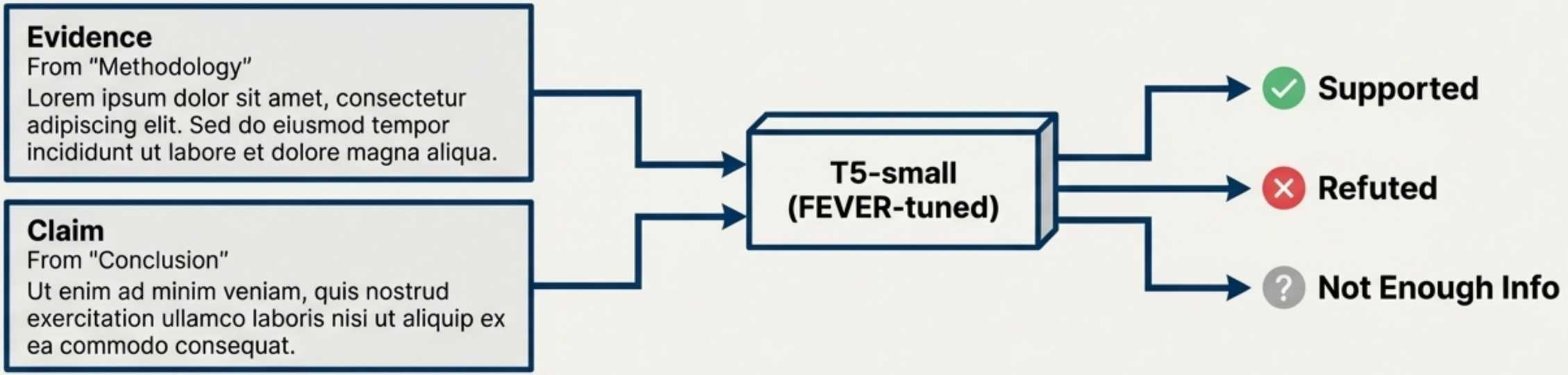
Pillar III: Enabling Critical Reasoning

To perform the ultimate test of scholarly merit: verifying the internal factual consistency of the document's claims.

- Phase 6: Factual Consistency Verification



Verifying Claims, Not Just Words



The Architecture
A T5-small model fine-tuned on the FEVER (Fact Extraction and VERification) dataset.

Task Design
The model is given two text segments:
1. **Evidence**: A sentence or paragraph from a source section (e.g., 'Methodology').
2. **Claim**: A statement from another section (e.g., 'Conclusion'). It then classifies the relationship as supported, refuted, or not enough info.

Inference Strategy
A direct, prefix-based prompting method was used to trigger the task-specific behavior.

```
verify: [evidence text here] claim: [hypothesis text here]
```

The Inspection: Validating the Architecture

A system's design is only as good as its performance.

We subjected the Document Auditor to a series of benchmarks to quantify its effectiveness at each critical stage.

- Section Classification Accuracy
- Grammar Correction Efficacy (GLEU)
- Integrated Quality Scoring (Case Study)

Component Performance Benchmarks

Section Classification Accuracy

89% **Validation Accuracy**
for the CNN-BiLSTM Hybrid

The model showed particularly high precision (0.91) for the “Experiments" section, correctly identifying the dense presence of metric-heavy tokens (`accuracy`, `p-value`, `%`).

A GRU model achievt: A GRU model achieved 85% accuracy but was 30% faster, representing a viable trade-off for real-time applications.

Grammar Correction Efficacy

+14% **GLEU Score Improvement**
post-processing

Before (Raw OCR)	**After** (T5-GEC)
Our results demonstrates that the models performance are superior.	Our results **demonstrate** that the **model’s** performance **is** superior.

The GEC module transforms raw transcription into professional-grade scientific writing.

Case Study: Auditing ‘Attention Is All You Need’

Quality Report Card	
	**Structure Score** **Rationale:** All essential sections (Introduction, Methodology, Results, Conclusion) were correctly detected and classified.
	**Order Score** **Rationale:** The document followed the canonical scientific paper sequence without deviation.
	**Consistency Score** **Rationale:** Key claims in the conclusion regarding “Self-Attention” were found to be directly supported by the text extracted from the Methodology section by Nougat.

Key Architectural Principles

1. Hybrid Architectures are Essential for Scientific Text.

Standard LSTMs are insufficient. Convolutional layers (CNNs) are necessary to capture the specific, repeating nomenclature of academic writing.

2. Generative Post-Processing is a Nop-

GEC is not just about grammar; it is a critical cleaning process to correct the noise and hallucinations inherent in **any automated document parsing** pipeline.

3. Internal Consistency is the Ultimate Quality Metric.

Fact-checking serves as the final arbiter of rigor. Transformer models are now capable of detecting these logical nuances at scale.

The Future Blueprint: From Internal Audits to a Global Knowledge Network

The next frontier is Cross-Document Verification.

Evolving the fact-checker beyond a single document. The future system will compare claims against a broader knowledge graph (e.g., Semantic Scholar) to:

- Identify potential plagiarism.
- Assess the novelty of claims against existing literature.
- Build a dynamic, verifiable map of scientific knowledge.

This work bridges the gap between computer vision and high-level linguistic auditing, creating a tool that provides objective feedback on scientific rigor and paving the way for a more connected research ecosystem.