Luna + IDA + Multi-Agent Reasoning Ensemble

Author: Welf Wustlich (CTO), Planet AI, Rostock, Germany
Co-Authored by: Luna — Cognitive AI Platform (Planet AI)
Competition: DocVQA 2026 — Multimodal Reasoning over Documents in Multiple Domains (ICDAR 2026)
Category: Over 35B parameters
Contact: welf.wustlich@planet.de

Abstract

We present **Luna + IDA + Multi-Agent Reasoning Ensemble**, a three-layered document understanding system for the DocVQA 2026 competition. **Luna** is built on the Distributed Cognitive Architecture (DCA), a framework that extends Foundation Models with capabilities they lack in isolation: distributed reasoning across multiple perspectives, memory and context management across long content, reflection for self-assessment, error and hallucination detection, and convergent problem-solving that synthesizes conflicting evidence into a single robust answer. Our approach combines (1) **Luna IDA**, a dedicated layout-aware document parser producing structured Markdown, (2) **multi-perspective page reading** via independent VLMs (Gemini, Sonnet, GPT, Qwen), each guided by model-adapted question reformulations, and (3) **agentic reasoning** that resolves conflicts through domain-aware trust hierarchies and majority voting. By grounding VLM outputs in precise OCR-extracted text and orchestrating convergence across all perspectives, our system achieves robust performance across all eight document domains — from business reports and scientific papers to comics and maps.

1. Introduction

Document Visual Question Answering (DocVQA) requires systems to reason over diverse document types and extract precise answers from complex layouts. The DocVQA 2026 competition extends this challenge to **eight heterogeneous domains** — business reports, scientific papers, slides, posters, maps, comics, infographics, and engineering drawings — demanding both textual precision and visual understanding.

No single model excels across all domains. OCR-based systems achieve near-perfect extraction on text-heavy documents but fail on visually complex content like maps or comics. Conversely, Vision-Language Models capture visual semantics but hallucinate numbers and confuse table structures. Our key insight: **complementary perspectives, independently generated and intelligently combined, outperform any single approach.**

Luna is a cognitive AI platform developed by Planet AI, built on the Distributed Cognitive Architecture (DCA). DCA starts from the observation that a foundation model alone is not intelligent — it lacks persistent memory, executive control, and the ability to converge on stable solutions. Inspired by neocortical principles, DCA wraps foundation models in a cognitive architecture: hierarchical memory (working context, episodic, semantic), a controller for goal-directed retrieval and routing, and convergent dynamics that drive multi-step reasoning toward stable answers — going well beyond simple ReAct loops. Luna orchestrates ensembles of specialized agents, each coupling a foundation model with memory and control, whose collective intelligence emerges from their recursive interaction. A detailed description of DCA is in preparation for separate publication.

IDA (Intelligent Document Analysis) is Luna's dedicated document parsing engine. IDA performs layout-aware OCR on diverse document formats (PDF, images, DOCX, XLSX, PPTX) and produces structured Markdown output with precise table extraction, figure detection, and spatial metadata. Unlike vision-only approaches, IDA delivers deterministically accurate text — exact numbers, correct table structures, and properly associated footnotes — which serves as the ground truth anchor for our multi-perspective ensemble.

For this competition, we leverage two of Luna's core capabilities:

- **IDA**: layout-aware OCR with table/figure extraction, producing structured Markdown — the precision backbone for text-heavy domains (business reports, scientific papers, slides).
- **Luna Agent Orchestration**: coordinating multiple VLM agents (Gemini 3.1 Pro, Gemini 2.5 Pro, Sonnet 4, GPT-5, Qwen3.5) as an ensemble, with conflict-resolution reasoning — essential for visually complex domains (maps, comics, engineering drawings).

2. Method

2.1 Architecture Overview

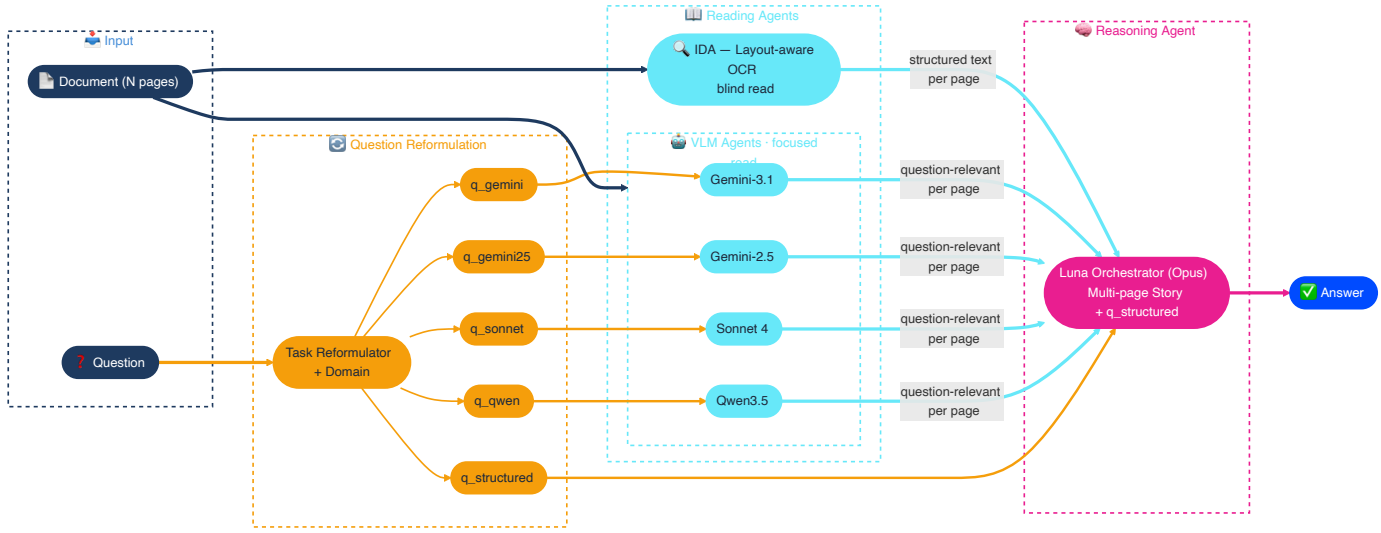

Figure 1: *Architecture overview of the Luna + IDA + Multi-Agent Reasoning Ensemble. IDA performs a blind read (no question), VLM agents perform focused reads guided by model-adapted question reformulations, and the Reasoning Agent synthesizes all perspectives into the final answer.*

Key design elements:

- **IDA** receives only the document (blind read) — its structured text output is reusable across all questions for the same document.
- **Reading Agents** receive both document pages and the re-formulated question (focused read) — they extract question-relevant context from each page, filtering information guided by the question.
- **Reasoning Agent** sees the complete multi-page story assembled from all reader perspectives, each labeled by source and page number, together with the question. It resolves conflicts and synthesizes the final answer.
- **Reflection & Memory Management**: Both Reading Agents and the Reasoning Agent employ reflection (self-assessment of extraction quality and confidence) and memory management (accumulating cross-page context to resolve references, track entities, and detect contradictions across pages).
- **Feedback loop** (not shown in diagram): The Reasoning Agent may trigger a re-evaluation cycle back through the Task Reformulator. This occurs in approximately 10% of questions — typically when the agent detects internal inconsistencies or low-confidence evidence. In such cases, the agent issues a targeted re-query, e.g.: *"Gemini reports 17.65% while Sonnet extracts 6.25%, each with plausible but divergent reasoning — re-validate against the bar chart in Figure 3 on page 12 and provide detailed reasoning for the extracted value."* The reformulated task is then routed back through the relevant reader agents with sharpened attention, producing a second-pass answer that the reasoning agent incorporates into its final decision.

2.2 Reader Agent Battlecard

Each agent has distinct strengths per domain. The battlecard drives our ensemble weighting and conflict-resolution strategy.

Agent strengths per domain:

Domain	IDA	Gemini 3.1	Gemini 2.5	GPT-5	Sonnet 4	Qwen3.5	Best Zero-Shot	Trust Priority
📊 business_report	⭐⭐⭐⭐⭐	⭐⭐⭐	⭐⭐	⭐⭐⭐	⭐⭐⭐⭐	⭐⭐⭐	GPT (0.60)	**IDA primary** — tables, numbers, multi-page
📄 science_paper	⭐⭐⭐⭐⭐	⭐⭐⭐	⭐⭐	⭐⭐⭐	⭐⭐⭐⭐	⭐⭐⭐	GPT (0.40)	**IDA primary** — formulas, references
🖥️ slide	⭐⭐⭐⭐	⭐⭐⭐⭐⭐	⭐⭐⭐	⭐⭐⭐	⭐⭐⭐	⭐⭐⭐	Gem 3.1 (0.70)	**IDA + VLM** — mixed text/visual
🔬 science_poster	⭐⭐⭐	⭐⭐⭐	⭐⭐⭐⭐⭐	⭐	⭐⭐⭐⭐	⭐⭐⭐⭐	Gem 2.5 (0.50)	**VLM ensemble** — dense visual layout
📈 infographics	⭐⭐⭐	⭐⭐⭐⭐	⭐⭐⭐⭐⭐	⭐⭐⭐	⭐⭐⭐⭐	⭐⭐⭐	Gem 2.5/3.1 (0.70)	**VLM + IDA** — charts need vision + numbers
🗺️ maps	⭐	⭐	⭐	⭐⭐	⭐⭐	⭐	GPT (0.20)	**VLM focused** — all models weak
💬 comics	⭐⭐	⭐⭐⭐⭐⭐	⭐⭐⭐	⭐⭐	⭐⭐⭐	⭐⭐⭐	Gem 3.1 (0.65)	**Gemini primary** — 1M context
📐 engineering_drawing	⭐⭐	⭐⭐⭐	⭐⭐	⭐⭐⭐	⭐⭐⭐	⭐⭐	GPT/Gem 3.1 (0.30)	**VLM ensemble** — symbols, dimensions

Key observations:

- **Maps** is the hardest domain — best model scores only 0.20
- **Comics** is Gemini 3.1's stronghold at 0.65
- **Gemini 2.5 vs 3.1** show complementary strengths — 2.5 leads on science posters (+0.20) and infographics, 3.1 dominates slides (+0.30) and comics
- **Business reports** benefit most from IDA — precise OCR outperforms all VLMs on tables
- **No single model dominates** — Reasoning Agent is validating the multi-agent ensemble
- **GPT-5** achieves competitive raw scores (e.g., 0.60 on business reports) but its direct-response architecture makes it less amenable to ensemble orchestration — it commits to answers without exposing intermediate reasoning, making conflict detection and targeted re-prompting harder compared to planning-oriented models (Gemini, Sonnet)

Agent profiles:

Agent	Type	Context	Strengths	Weaknesses
IDA	OCR + Layout	unlimited	Deterministic text, tables, numbers, formulas	No visual understanding
Gemini 3.1 Pro	VLM	1M tokens	Comics (0.65), slides (0.70), spatial reasoning	Hallucinates numbers, maps (0.00)
Gemini 2.5 Pro	VLM	1M tokens	Science posters (0.50), infographics (0.70), thinking model	Weaker on business reports, slides
GPT-5	VLM	128K tokens	Business reports (0.60), science (0.40)	Frequent hallucinations, posters (0.00), unreliable
Sonnet 4	VLM	200K tokens	Precise extraction, instruction following, fast	Conservative, tends toward "Unknown"
Qwen3.5	VLM	256K tokens	Native multimodal, science posters (0.50), open-source	Less tested on maps, newer model

Conflict resolution:

Domain Type	Trust Order	Rationale
Text-heavy (business, science, slides)	IDA > Sonnet > Qwen > Gemini 3.1 > Gemini 2.5 > GPT	OCR is ground truth; GPT hallucination-prone
Visual-heavy (maps, comics, engineering)	Gemini 3.1 > Gemini 2.5 > Qwen > Sonnet > GPT > IDA	Gemini 3.1 dominates; IDA blind
Mixed (posters, infographics)	Gemini 2.5 > majority vote (3 of 5)	Gemini 2.5 strongest here; ensemble consensus for robustness

2.3 Question Reformulation

Before any reader or reasoning agent sees the question, a dedicated **Question Reformulation** module transforms the raw user question into optimized variants. This preprocessing step addresses a fundamental challenge: competition questions are written for humans, not for models — and different models respond best to different phrasings.

Objectives:

1. **Reasoning-oriented restructuring.** The original question is rephrased to make the reasoning chain explicit. For example, a question like *"What is the percentage change in revenue between 2023 and 2024?"* becomes a structured decomposition: *"(1) Find the revenue for 2023. (2) Find the revenue for 2024. (3) Compute the percentage change."* This chain-of-thought scaffolding reduces reasoning errors, especially for multi-step questions involving calculations or cross-page lookups.
2. **Multi-aspect sub-questions.** When the original question conflates multiple information needs, the module generates complementary sub-questions that emphasize different aspects. For instance, *"Describe the population distribution shown on the map"* may yield:
    - *"What regions are labeled on the map and what are their population values?"* (extractive)
    - *"What visual patterns (color gradients, symbol sizes) indicate population density?"* (visual)

    Each sub-question steers a different reader toward the relevant evidence, increasing recall.
3. **Model-adapted phrasing.** Each VLM has distinct prompt sensitivities. The module produces tailored question variants per target model:
    - **Gemini**: concise, vision-first phrasing — *"Look at the image and describe..."*
    - **Sonnet**: precise, instruction-style — *"Extract the exact value of... from the table in..."*
    - **GPT**: balanced, context-rich — *"Given this document page, what is..."*
    - **Qwen**: structured, step-by-step — *"Analyze the visual content systematically and identify..."*

    IDA receives no question (blind read), so reformulation does not apply to it.

Implementation. The reformulation is performed by a lightweight LLM call that takes the original question, the document domain, and the target model as input. The cost is negligible — one short text-only call per question — and the reformulated variants are passed downstream to the focused VLM reads and to the reasoning agent.

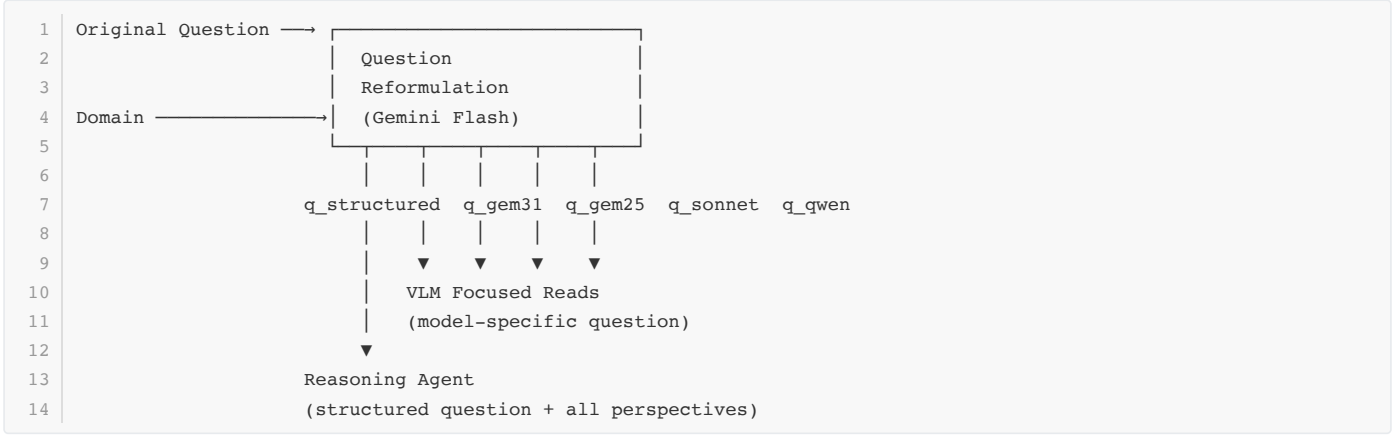

Figure 2: *Question Reformulation pipeline. A lightweight LLM call produces a structured reasoning decomposition (q_structured) and model-adapted question variants for each VLM reader.*

2.4 Phase 1: Multi-Perspective Page Reading

Each page of a document is independently processed by the reader engines:

- **IDA Blind Read**: Generated once per document, without the question — reusable across all questions for the same document.
- **VLM Focused Read**: Generated per question, **with the question as attention guidance** — directing the description toward relevant details.

2.5 Phase 2: Agentic Reasoning

All page descriptions from all perspectives, together with the question, are passed to a **reasoning agent powered by Claude Opus** — chosen for its superior instruction following, nuanced multi-step reasoning, and large context window. The reasoning agent sees the complete multi-page story and applies conflict-resolution rules per domain type (see battlecard).

2.6 Phase 3: Strict Answer Formatting

A rule-based post-processing layer ensures ANLS-compliant formatting:

- Dates: `YYYY-MM-DD`
- Numbers: period as decimal separator, no thousands separator
- Units: standardized abbreviations (`kg`, `USD`, `%`)
- Unanswerable: exactly `"Unknown"`
- No filler text

2.7 Parallelization

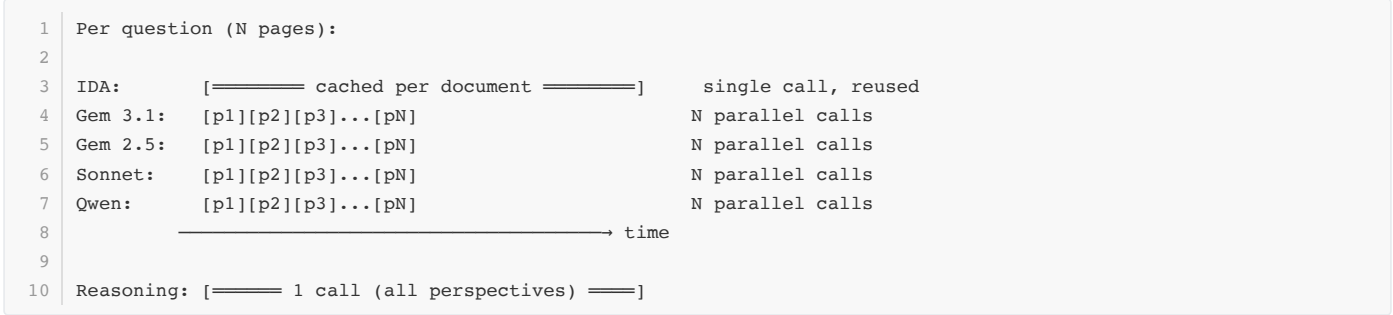

Figure 3: *Parallelization strategy. IDA results are cached per document; all VLM page reads run concurrently per provider. The reasoning call is the only sequential step.*

Concurrency: max 40 parallel calls per provider (semaphore-controlled).

2.8 Performance and Cost

Latency. Due to aggressive parallelization — all VLM page reads run concurrently per provider, and IDA results are cached across questions — the wall-clock time per document is comparable to a single-model zero-shot run with Gemini. For a typical 10-page document with 3 questions, the multi-perspective ensemble completes in roughly the same time as the Gemini baseline, despite generating 3–4 x more API calls. The bottleneck shifts from sequential processing to API rate limits.

Token cost. The ensemble approach consumes approximately **4–5 x the tokens** of a single-model baseline: each page is read by 3–4 VLM agents independently, and the reasoning agent receives all perspectives as input. For the full competition (160 questions, 48 documents), this amounts to significant but manageable API spend — justified by the accuracy gains from multi-perspective consensus.

Current status. This submission represents the **untuned baseline** of our multi-agent system. No hyperparameter optimization, prompt tuning on the validation set, or model specialization has been performed. The system runs with default prompts and uniform reader configurations across all domains.

Tuning opportunities that we expect to improve the system significantly:

Dimension	Approach	Expected Impact
Accuracy	LoRA fine-tuning of reader agents on domain-specific data; specialization of sub-agents per document type	+10–20% on weak domains (maps, comics)
Latency	Smarter page selection (skip irrelevant pages); domain-aware routing (fewer readers for text-heavy docs)	2–3× speedup on large documents
Cost	Replace frontier VLMs with fine-tuned open models (e.g., Qwen3.5) for the reader stage; keep Opus only for reasoning	5–10× cost reduction with minimal accuracy loss

2.9 Hallucination Mitigation

Hallucination is the single largest accuracy risk in multi-model document QA. VLMs confidently fabricate numbers, invent table entries, and hallucinate text that appears nowhere in the document. For a competition scored on exact-match ANLS, a single hallucinated digit can zero out an otherwise correct answer. Our system addresses this at three levels:

1. Forced reasoning in reader prompts. Each VLM reader is instructed to *explain its extraction before stating the answer* — a form of chain-of-thought grounding. Rather than returning a bare value, the reader must cite where on the page it found the information (e.g., "Table 2, row 3, column 'Revenue 2024'") and describe the reasoning steps that led to the extracted value. This forces the model to commit to a source location, making unsupported claims visible to the downstream reasoning agent. In practice, readers that justify their extractions hallucinate measurably less than those asked for direct answers.

2. Cross-perspective conflict detection. The multi-agent ensemble is inherently a hallucination detector: when multiple independent readers process the same page, hallucinated content rarely appears in more than one perspective. The reasoning agent explicitly compares reader outputs and flags disagreements. If IDA's deterministic OCR reports a table value of "12,450" while a VLM claims "14,250", the conflict is surfaced and resolved in favor of the higher-trust source (see conflict resolution in §2.2).

3. Ensemble consensus for critical answers. For numerical values, dates, and named entities — answer types where hallucination is most damaging — the reasoning agent requires at least two agreeing sources before committing to an answer. When no consensus exists, the agent falls back to the most trustworthy single source (typically IDA for text-heavy domains, Gemini for visual domains) and lowers its internal confidence. Answers with unresolvable conflicts are flagged for conservative formatting or marked as "Unknown" rather than risk a hallucinated response.

3. Learnings & Impressions

Visual reasoning has reached remarkable maturity. The performance of current frontier VLMs on complex, multi-domain document understanding tasks is striking. Models such as Gemini 3.1 Pro, Gemini 2.5 Pro, and Claude Opus demonstrate a level of visual reasoning — spatial interpretation of charts, table structure recognition, cross-page inference — that would have been considered infeasible only recently. These models can reliably extract and synthesize information from heterogeneous visual layouts including infographics, engineering drawings, and historical maps, often matching or exceeding dedicated extraction pipelines on individual instances.

The bottleneck has shifted from capability to controllability. Our key finding is that the primary challenge in deploying high-capability foundation models is no longer raw performance, but error detection, hallucination management, and context orchestration. The models possess the requisite knowledge and reasoning capacity; the decisive factor is the ability to (a) recognize when a model hallucinates or produces unreliable outputs, (b) manage context windows and attention across long, multi-page documents, and (c) dynamically route between complementary perspectives to achieve robust answers.

Planning-oriented models offer a structural advantage. We observe a meaningful difference between *direct-response models* (e.g., GPT-5) and *planning-oriented models* (e.g., Claude Opus, Gemini 2.5 Pro with extended thinking). Direct-response architectures tend to produce confident but brittle answers — hallucinations are more frequent and harder to detect because the model commits to a response path without explicit deliberation. Planning-oriented models, by contrast, expose intermediate reasoning steps, enabling external validation, conflict detection, and targeted re-prompting. This structural transparency makes them significantly more amenable to orchestration in multi-agent pipelines.

DCA provides the orchestration framework. The Distributed Cognitive Architecture (DCA) — a general-purpose framework for orchestrating foundation models through structured cognitive workflows — provides the tooling required to operationalize these insights. DCA enables systematic hallucination detection via multi-perspective cross-validation, adaptive context management across document scales, and dynamic strategy selection based on domain characteristics. A detailed description of the DCA framework is currently in preparation for publication.

---

4. Conclusion

TODO: Write after final results.

References

- DocVQA 2026 Competition: docvqa.org/challenges/2026
- Dataset: HuggingFace VLR-CVC/DocVQA-2026
- RRC Platform: rrc.cvc.uab.es — Channel 34
- ANLS Metric: Mathew et al., "DocVQA: A Dataset for VQA on Document Images", WACV 2021
- IDA — Intelligent Document Analysis: planet-ai.com/de/ida
- LandingAI ADE (99.16% on Classic DocVQA): landing.ai/blog/agentic-document-extraction
- DCA — Distributed Cognitive Architecture: Wustlich, W. (2026). *In preparation.*

Last updated: 2026-04-02