Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Winning DocVQA 2026: Multi-Agent Orchestration for Multi-Domain Document Reasoning

Welf Wustlich

welf.wustlich@planet.de

Planet AI GmbH

Research Article

Keywords: Document visual question answering, Multi-agent systems, Vision-language models, Document understanding, LLM orchestration, ANLS

Posted Date: November 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11103268/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Winning DocVQA 2026: Multi-Agent Orchestration for Multi-Domain Document Reasoning

Welf Wustlich[1*]

[1*]Planet AI GmbH, Rostock, Germany.

Corresponding author(s). E-mail(s): welf.wustlich@planet.de;

Abstract

Document Visual Question Answering (DocVQA) over heterogeneous domains requires precise extraction and reasoning over documents whose formats range from business reports and scientific papers to maps, comics, and engineering drawings. No single vision-language model (VLM) performs uniformly well across such diverse content. We present a multi-agent document-reasoning system submitted to the DocVQA 2026 competition at ICDAR 2026, where it placed first in Task 1 (General category, >35B-parameter class) with 60.00% ANLS (95% Wilson-score CI: ±7 pp; $\boldsymbol{n = 160}$), 13.75 pp ahead of the second-placed system (46.25%). The system combines three components: (1) IDA, a layout-aware OCR engine that produces structured Markdown output as a deterministic text anchor; (2) four frontier VLMs (Gemini 3.1 Pro, Gemini 2.5 Pro, Claude Sonnet 4, Qwen3.5) performing question-guided focused reads with model-adapted question reformulations; and (3) a Claude Opus 4.6 reasoning agent that synthesizes perspectives through domain-specific trust hierarchies and cross-perspective conflict detection. A first submission reached 55.63% ANLS; a refinement pass using the first run's reasoning outputs as prior context lifted the score to 60.00% (**+4.37** pp). The system improves over the strongest single-model zero-shot baseline on the leaderboard (Gemini 3.1 Pro: 37.50%) by **+22.5** pp and over a naive mixture-of-experts estimate (∼40%) by approximately **+20** pp. We attribute roughly **+7** pp to IDA's deterministic text extraction and **+13** pp to multi-agent orchestration; these are attributed estimates from internal experiments rather than controlled-ablation measurements, and we state this limitation explicitly. Per-domain, the system scores highest on text-heavy documents (business reports 0.75, scientific papers 0.70 ANLS), where deterministic extraction anchors the answer, and lowest on comics (0.45), with maps and science posters at 0.50.

Keywords: Document visual question answering, Multi-agent systems, Vision-language models, Document understanding, LLM orchestration, ANLS

1 Introduction

Ask a business report *"what is the percentage change in revenue between 2023 and 2024?"* and a comic strip *"who is holding the umbrella in panel 3?"*. No single vision-language model (VLM) answers both well: the model that reads a table cleanly stumbles on panel narrative, and the model that parses a scene confuses table rows and misreads numbers. DocVQA 2026 [5], held at ICDAR 2026, turns this into a benchmark. It extends the classic single-domain DocVQA task [11] to eight heterogeneous domains (business reports, scientific papers, slides, science posters, infographics, maps, comics, and engineering drawings), each with its own layout conventions and reasoning demands. Single-model accuracy saturates well below human level across this range.

Our result has a simple shape. The strongest frontier models, combined in a naive ensemble, reach about 40% ANLS. The same models, coordinated through structured orchestration, reach 60.00% and first place in Task 1 (General category). The models are identical across both configurations; the 20-point gain comes from how they are coordinated. That coordination is the subject of this paper.

The task is hard for three reasons. First, no single VLM is uniformly strong across all domains: models that handle text-heavy layouts well tend to hallucinate on visually encoded data, while models strong on images often confuse table structure or misread numerical values. Second, complementary strengths shift with model generation, so static ensemble weights become stale. Third, multi-page documents require tracking entities and numerical claims across context boundaries, which single-pass inference handles poorly.

Our system addresses these challenges with a three-layer pipeline. IDA, Planet AI's layout-aware document analysis engine, provides deterministic text extraction as a ground-truth anchor for text-heavy domains. Four frontier VLMs perform question-guided page reads in parallel, each receiving a model-adapted reformulation of the original question. A Claude Opus 4.6 reasoning agent synthesizes the evidence from all readers using domain-specific trust hierarchies, detects cross-perspective conflicts, and — in approximately 10% of questions — triggers a targeted re-evaluation before committing to an answer.

Contributions:

- A multi-agent pipeline for multi-domain document VQA combining deterministic OCR, model-adapted question reformulation, four frontier VLMs, and a reasoning agent with domain-aware conflict resolution.
- An attributed breakdown of component contributions ($\approx +7$ pp from IDA, $\approx +13$ pp from orchestration) with an explicit statement that these are not controlled-ablation measurements.

- Empirical evidence that the same frontier models used in a naive ensemble (∼40%) yield substantially higher accuracy (60.00%) when coordinated through structured trust hierarchies and cross-perspective conflict detection.
- A per-domain analysis of where the system succeeds, where it fails, and where the refinement pass (55.63% → 60.00%) produced changes.

The system is implemented in Luna, a cognitive AI platform developed by Planet AI. The multi-agent framework underlying Luna is described as the Distributed Cognitive Architecture (DCA) in companion publications [20, 21]; the present paper uses the operational capabilities (memory, task reformulation, iterative refinement) without reproducing the formal framework.

2 Related Work

2.1 Document Visual Question Answering

Document VQA was introduced by Mathew et al. [11], who defined the task, the dataset, and the ANLS (Average Normalized Levenshtein Similarity) evaluation metric (adapted from the scene-text VQA metric of Biten et al. [4]). ANLS is a soft string-similarity measure: it assigns a score in $[0, 1]$ per question based on the edit distance between prediction and ground truth, normalized and thresholded, and then averages across questions. It is not an exact-match metric; a small surface deviation reduces but does not zero the score. The DocVQA 2026 competition uses ANLS as the primary metric, following the ICDAR 2021 DocVQA challenge [16].

Layout-aware document understanding has been advanced by pretrained models that jointly encode text and layout (LayoutLM [22]; LayoutLMv3 [7]) and by OCR-free document transformers that read directly from the image (Donut [8]; Pix2Struct [10]). These situate IDA-style layout-aware extraction in the literature: IDA plays the role such models play as a text/structure provider, but is used here as a deterministic input to a reasoning ensemble rather than as an end-to-end answerer.

A particularly relevant contemporary system is LandingAI's Agentic Document Extraction (ADE), which reports 99.16% ANLS on the classic (single-domain) DocVQA benchmark by treating the problem as an agentic extraction task with text-only reasoning, without VLM image input [9]. ADE demonstrates that for text-heavy, single-domain settings, high-quality OCR plus reasoning can exceed human-level performance. DocVQA 2026's eight-domain scope, which explicitly includes maps, comics, and engineering drawings, represents a broader challenge where text-only approaches face structural limits; visual reasoning is required alongside deterministic text extraction.

2.2 Agentic Reasoning and LLM Orchestration

Multi-step reasoning with LLMs has been advanced through several techniques. Chain-of-thought prompting [19] elicits intermediate reasoning steps before answers. ReAct [23] interleaves reasoning with external tool calls. Reflexion [15] adds verbal self-reflection on prior trajectories. These share a pattern of iterative action over external context but lack persistent structured memory and formal termination criteria.

Our system incorporates structured re-evaluation: the reasoning agent may trigger a second reading pass for approximately 10% of questions when it detects conflicting evidence. This is in the ReAct tradition but guided by domain-specific trust hierarchies rather than free-form tool selection.

2.3 Memory-Augmented Systems

Memory has been added to LLM pipelines through retrieval-augmented generation (RAG) and through context pagination, most prominently MemGPT [13], which implements virtual long-term memory through context swapping. These systems treat memory primarily as an external retrieval resource. The present work uses memory in a more specific sense: per-page reading state is accumulated as a chain within each VLM's context window, and prior-run reasoning outputs serve as warm-start context in the refinement pass. The formal DCA memory model (working, episodic, and semantic levels) is described in [21] and is not the subject of the current paper.

2.4 VLM Ensembles for Document Understanding

Mixture-of-experts and ensemble configurations are active in both research and production. Two lines are directly relevant: self-consistency, which aggregates multiple reasoning paths from one model [18], and mixture-of-agents, which layers multiple LLMs so that later agents refine earlier ones' outputs [17]. Our system differs from both in using domain-aware trust hierarchies and cross-perspective conflict detection rather than uniform aggregation or fixed layering. The $\sim$40% ANLS that naive MoE configurations achieve on DocVQA 2026 (our internal estimate) provides the baseline against which the structured ensemble is evaluated. The key distinction between a naive ensemble and our system is not the set of models used but the orchestration layer: domain-aware trust hierarchies, cross-perspective conflict detection, and model-adapted question routing.

3 System

3.1 Overview

As noted in Section 1, the system runs on Luna's Distributed Cognitive Architecture (DCA), which couples each foundation model with a memory and control layer, organizing inference as an iterative process that converges through cross-perspective comparison rather than a single forward pass. We use it operationally here as the scaffolding for task reformulation, multi-reader orchestration, and iterative refinement.

For DocVQA 2026, the system has three layers (Figure 1): IDA provides deterministic text extraction as a document-level ground truth; the VLM ensemble provides visual and semantic perspectives; the reasoning agent synthesizes these into a final answer. Each question triggers the full pipeline; IDA's document representation is computed once per document and cached across all questions for that document.

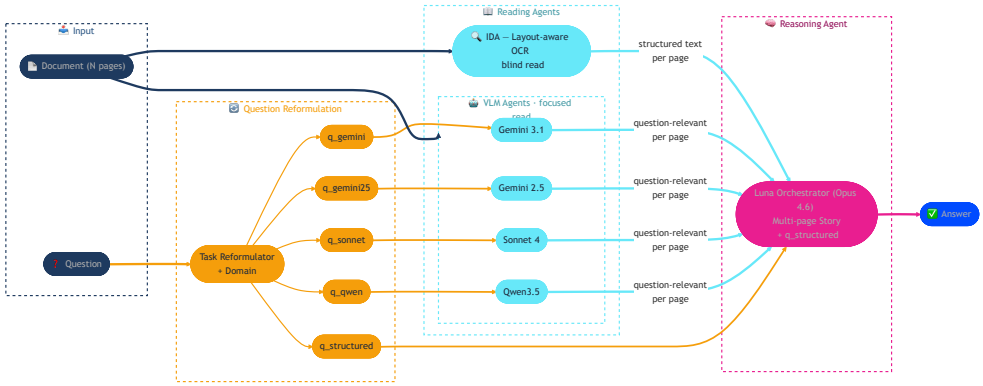


Fig. 1 System overview. IDA performs a blind read (no question) and produces structured Markdown per page. Four VLMs perform question-guided focused reads using model-adapted question variants. The Reasoning Agent synthesizes all perspectives with domain-specific trust hierarchies; in about 10% of questions it triggers a re-evaluation pass through the Task Reformulator before committing to a final answer.

Table 1 The *battlecard*: reader strengths by domain (from internal validation) and which reader the system trusts for each. No single reader dominates. Context windows: IDA effectively unlimited (sequential per-page); Gemini 3.1/2.5 1M tokens; Sonnet 4 200K; Qwen3.5 256K.

Domain	IDA	Gem 3.1	Gem 2.5	Sonnet 4	Qwen3.5	Trust priority
business_report	strong	medium	weak	strong	medium	IDA primary
science_paper	strong	medium	weak	strong	medium	IDA primary
slide	strong	strong	medium	medium	medium	IDA + VLM
science_poster	medium	medium	strong	strong	strong	VLM ensemble
infographics	medium	strong	strong	strong	medium	VLM + IDA
maps	weak	weak	weak	weak	weak	VLM focused
comics	weak	strong	medium	medium	medium	Gemini primary
engineering_drawing	weak	medium	weak	medium	weak	VLM ensemble

3.2 Reader Agents and the Battlecard

The system uses five readers: IDA and four VLMs. Their strengths across the eight domains are summarized in Table 1, which we call the *battlecard*: a per-domain scouting report of which reader to trust for what, derived from internal validation experiments. The reasoning agent consults it the way a captain consults a scouting report before a match, then reads the trust order in Table 2 off it.

Three patterns are apparent. No single model dominates: IDA leads on text-heavy domains, Gemini 3.1 on comics and slides, Gemini 2.5 on science posters and infographics. The two Gemini generations complement each other: 2.5 is stronger on science posters, 3.1 on slides and comics. Maps are the hardest domain, with no reader providing reliable extraction; this reflects a gap in spatial-symbolic reasoning across current VLMs.

Table 2 Trust order by domain type, applied by the Reasoning Agent for conflict resolution.

Domain type	Trust order
Text-heavy	IDA > Sonnet > Qwen > Gemini 3.1 > Gemini 2.5
Visual-heavy	Gemini 3.1 > Gemini 2.5 > Qwen > Sonnet > IDA
Mixed	Gemini 2.5 > majority vote (3 of 4 VLMs)

GPT-5 and ensemble compatibility. GPT-5 [12] achieves the highest raw zero-shot score on business reports (≈0.60) among the models we evaluated. We exclude it from the active ensemble because, in the API configuration available to us during the competition, it commits to answers without exposing intermediate reasoning steps, making cross-perspective conflict detection and targeted re-prompting substantially harder. Planning-oriented models that expose chain-of-thought reasoning (Gemini 2.5 with extended thinking [6], Claude Opus 4.6) are more amenable to orchestration: their intermediate steps enable the reasoning agent to identify which sub-claim to challenge and why. Whether GPT-5's reasoning mode would change this is untested; we leave that as a question for follow-up.

Conflict resolution. When reader outputs disagree, the Reasoning Agent applies domain-specific trust ordering: for text-heavy domains (business, science, slides), IDA is treated as the ground-truth anchor and takes precedence over VLM claims; for visual-heavy domains (maps, comics, engineering drawings), Gemini 3.1 is the primary source; for mixed domains (posters, infographics), Gemini 2.5 is prioritized, with majority vote as fallback.

3.3 Pipeline Phases

The pipeline runs in four phases.

Phase 0 — Question Reformulation. A lightweight LLM call (Gemini Flash) transforms the raw question into multiple variants: a structured chain-of-thought decomposition ($q_{\text{structured}}$) used by the Reasoning Agent, and a model-adapted variant per VLM reader. The structured decomposition makes multi-step reasoning explicit: the question "What is the percentage change in revenue between 2023 and 2024?" becomes three sequential sub-tasks (find revenue 2023; find revenue 2024; compute percentage change). Model-adapted variants adjust phrasing: Gemini receives vision-first prompts; Sonnet receives precise extraction-style instructions; Qwen receives step-by-step structured prompts. The cost is one short text-only call per question. IDA receives no question and is not affected by reformulation.

Phase 1 — Multi-Perspective Page Reading. IDA processes the document once (blind, no question) and produces structured Markdown per page, including tables, figures, and spatial metadata. The output is cached at the document level and reused across all questions for that document. Each VLM reader processes pages sequentially within its own thread, maintaining a per-page history chain so that each page read has access to prior pages' extracted content. The four VLM threads run in parallel;

the IDA cache is independent of this. The output of Phase 1 is a labeled collection of page-level descriptions, one per (reader, page) pair.

Phase 2 — Agentic Reasoning. The Reasoning Agent (Claude Opus 4.6) receives the full multi-page evidence collection, with each excerpt labeled by reader and page number, along with $q_{\text{structured}}$. It applies the trust order from Table 2 to resolve conflicts and synthesizes a final answer. When it detects conflicting or low-confidence evidence — in approximately 10% of questions — it triggers a re-evaluation: the Task Reformulator generates a refined query directed at the specific conflict, the relevant readers produce a second-pass response, and the Reasoning Agent incorporates that response before committing.

Phase 3 — Answer Formatting. ANLS scoring penalizes surface formatting deviations. A rule-based post-processing layer enforces ANLS-compliant output: dates as YYYY-MM-DD, numbers with period as decimal separator and no thousands separator, standardized unit abbreviations, unanswerable items as exactly "Unknown". The formatting layer is a deterministic transform applied after the Reasoning Agent's semantic answer, which separates the semantic reasoning problem from the surface formatting problem.

3.4 Cross-Cutting Concerns

Parallelization. Within each VLM, pages are read sequentially (maintaining the memory chain); the four VLM threads run in parallel. IDA outputs are cached and shared across all questions for a given document. The Reasoning Agent invocation begins only after Phase 1 completes for all readers. Concurrency is bounded at 40 parallel API calls per provider.

Hallucination mitigation. The system addresses hallucination at three levels. First, reader prompts require chain-of-thought grounding: each reader must cite its source location (e.g., "Table 2, row 3, column 'Revenue 2024'") before stating the answer. Second, cross-perspective conflict detection: hallucinated content rarely appears identically in multiple independent reads, so disagreement between readers is a hallucination signal. Third, for numerical values, dates, and named entities — where hallucination is most costly — the Reasoning Agent requires consensus from at least two sources or falls back to the highest-trust single source; it prefers "Unknown" over an unsupported answer.

Cost. The ensemble consumes roughly 4–5× the tokens of a single-model baseline (an order-of-magnitude estimate, not a precise measurement): four VLMs read each page independently, and the Reasoning Agent receives all perspectives as input. IDA caching amortizes OCR cost across questions for the same document. Wall-clock latency per document is broadly comparable to a single Gemini run due to parallelization, with the bottleneck shifting from computation to API rate limits. We did not benchmark cost or latency systematically; these figures are practical observations from the competition runs.

4 Results and Discussion

4.1 Overall Performance

The system ranked first in Task 1 (General category, >35B-parameter class) of the DocVQA 2026 competition with 60.00% ANLS (official leaderboard entry, 2026-04-04: *Luna + IDA + Multi-Agent Reasoning Ensemble → Update with Fine Tuning*; here "Fine Tuning" denotes validation-set tuning of the battlecard, sub-agent prompts, and agentic-flow parameters (configuration tuning, not model-weight fine-tuning), and it did not use the test set — see Section 4.4). Per-domain scores range from 0.75 (business reports) to 0.45 (comics); the full breakdown is in Table 3. With $n = 160$ questions, the 95% Wilson-score confidence interval is approximately ± 7 pp, giving an interval of $[53, 67]$. We report the two-decimal value because this is how the official CVC leaderboard publishes it; the precision is not meaningful at $n = 160$ and we do not claim it.

Reference points on the same leaderboard:

- **Second-placed system** (runner-up in the same category): 46.25% ANLS — a 13.75 pp lead.
- **Gemini 3.1 Pro** (official baseline entry): 37.50% ANLS.
- **Naive ensemble** (majority vote over the four frontier VLMs on the validation set, without orchestration): approximately 40% ANLS — an internal estimate (see the note below on why we report it as approximate).

The first submission reached 55.63% ANLS; the second submission reached 60.00% ANLS, a +4.37 pp refinement (see Section 4.3).

Figure 2 shows the attributed decomposition from the MoE baseline to the final score.

The ≈40% naive-ensemble baseline, and the split of the gain into ≈ +7 pp for IDA and ≈ +13 pp for orchestration, are approximate estimates rather than fixed measurements. We report them as approximate on purpose, for two reasons: at $n = 160$ the statistical uncertainty is large, and the values shift with the choice of models, system prompts, and configuration, so a single hard number would give false precision. We have measured them internally, and they are representative of what we observe for this task, but they are not constants. The rationale for the split: IDA's contribution is concentrated in the text-heavy domains (business reports, science papers, slides) where deterministic extraction of numerical values, table entries, and formulas determines answer correctness, while orchestration contributes across all domains through task reformulation, cross-perspective conflict detection, and iterative reasoning. The two components are tightly coupled in the operational pipeline (see Section 4.4).

4.2 Component Analysis

IDA as a precision text anchor (attributed ≈ +7 pp). For DocVQA 2026, IDA serves exclusively as an input supplier: it does not answer questions but provides the structured text representation on which the Reasoning Agent builds and validates its conclusions.

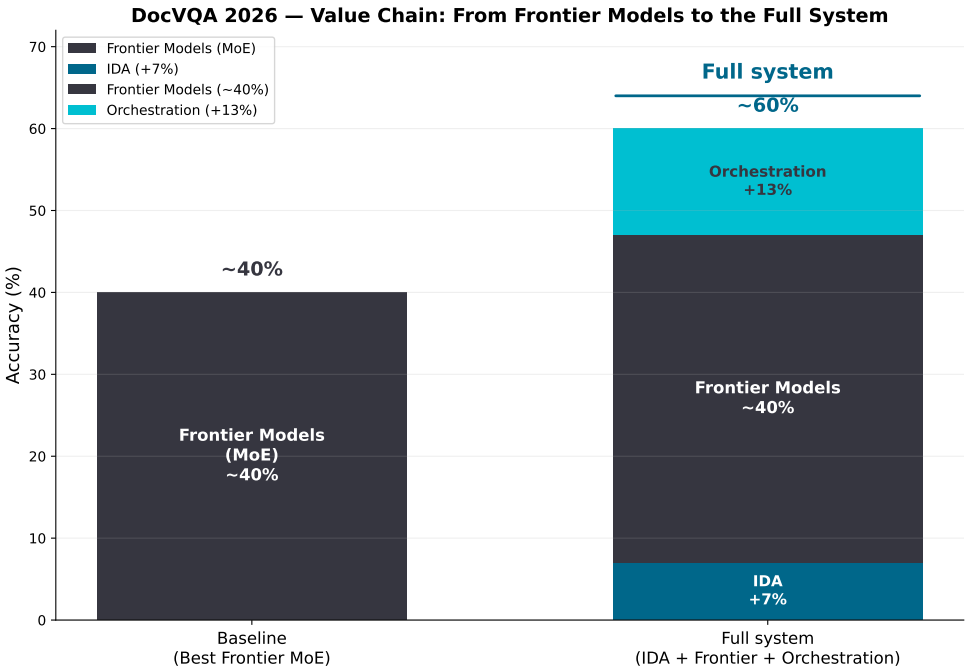


Fig. 2 Value chain from the naive-ensemble baseline to the submitted system — a schematic with attributed estimates, not a measured decomposition. Left: best frontier models in a naive mixture-of-experts configuration (≈40% ANLS). Right: the three-layer system. IDA's deterministic text extraction is attributed ≈ +7 pp; multi-agent orchestration (task reformulation, trust-hierarchy reasoning, reflection, iterative refinement) is attributed ≈ +13 pp. The split is based on internal experiments and qualitative domain analysis, not a controlled ablation, and the values are approximate and configuration-dependent (Section 4.4).

The attributed +7 pp reflects IDA's role in the four text-heavy domains of the battlecard (Table 1): business reports, scientific papers, slides, and partially science posters. In these domains, VLMs routinely hallucinate numerical values, confuse table rows and columns, and misread footnotes. IDA's deterministic OCR eliminates these error modes at the input level. The Reasoning Agent's trust order (Table 2) places IDA first in text-heavy domains precisely because IDA's output is grounded in the actual document bytes rather than a probabilistic visual interpretation.

The attribution is supported by two lines of qualitative evidence: (1) the battlecard shows IDA as the strongest reader in text-heavy domains; (2) the per-domain change table (Section 4.3) shows that refinement changes were absent from science_paper and slide, which are already well-served by IDA's text extraction and thus left no room for improvement from the visual-focus enhancements of the second run.

Multi-agent orchestration (attributed ≈ +13 pp). The orchestration contribution comes from four mechanisms: question reformulation, trust-hierarchy reasoning, cross-perspective hallucination detection, and iterative re-evaluation.



Fig. 3 An illustrative example from the validation set: a maps question that strong single models miss. The system solves it in grounded steps (Figure 4): locate stand 30, orient north via the compass rose, follow the main avenue to the last stand (16), and read its name from the legend. Each step is grounded in a focused crop of the map rather than a single global read.

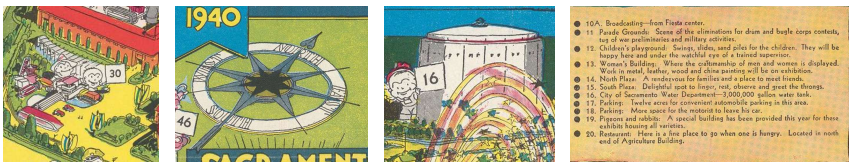
Question reformulation improves reader performance by converting the human-written question into a structured reasoning decomposition and model-adapted variants. For multi-step questions involving calculations or cross-page lookups, explicit decomposition reduces reasoning errors relative to presenting the raw question.

Trust-hierarchy reasoning enables the Reasoning Agent to weight reader outputs by domain-specific reliability rather than treating all perspectives equally. A business-report question where IDA reports 12,450 and a VLM reports 14,250 is resolved in IDA's favor; a comics question where IDA produces no useful output is resolved from the Gemini 3.1 perspective. Majority voting would not capture this asymmetry.

Cross-perspective conflict detection functions as a hallucination filter: an answer that appears in only one of five readers is treated with lower confidence than one corroborated by multiple independent reads. A substantial part of the attributed +13 pp likely comes from rejecting wrong answers rather than producing additional correct ones.

Iterative re-evaluation — the ∼10% of questions where the Reasoning Agent triggers a second reading pass — catches specific errors that survive the first pass because the conflict was ambiguous or the relevant visual detail was too fine for a global page read.

A validation-set example: a hard maps question. To show how the system reasons, we walk through one illustrative example from the validation set (Figure 3). It is a single case, not aggregate evidence.



(1) stand 30 (2) compass rose (3) stand 16 (4) legend

Fig. 4 The four grounded steps of the worked example (Figure 3), each a focused crop of the map: (1) locate the starting point, stand 30; (2) establish north from the compass rose; (3) follow the main avenue north to the last stand, no. 16; (4) read stand 16's name from the legend → *City of Sacramento Water Department*.

Question (maps, validation set): *"I'm at stand number 30. If I keep going north on the main avenue, what is the name of the last stand I will encounter?"*
Ground truth: stand 16 — *City of Sacramento Water Department*.

Answering this takes several grounded steps rather than a single glance. Two strong single models fail: one commits to a plausible but wrong stand, the other stalls entirely — extended-reasoning models can fail to converge on such multi-step problems, consistent with reported limits of reasoning models on high-complexity tasks [3].

System	Answer	
Gemini 3 Pro	"Main Entrance"	✗
Qwen3-VL (thinking mode)	no answer after 8192 reasoning tokens	✗
Full system	**"City of Sacramento Water Department"**	✓

The full system answers correctly by carrying the spatial reasoning across explicit steps, each grounded in a focused crop of the map rather than one global read. This is the same mechanism behind the system's 0.50 on maps (Table 3).

4.3 Refinement: 55.63% to 60.00%

The first submission (55.63%) used the full pipeline with uniform settings across all domains. The second submission built on the first run's reasoning outputs as prior context, with three targeted enhancements: focused image crops for dense visual content (tables, charts, engineering diagrams); deeper reasoning iterations, allowing the Reasoning Agent to explore more alternative interpretations; and parallel sub-queries through the Task Reformulator, increasing evidence coverage per question.

The refinement uses within-system context only: the leaderboard returns a single aggregate score with no per-question breakdown, so no test-set feedback could guide the selection of changes. The second run is equivalent to extending the first by one additional outer iteration, using the first run's answers as a warm start. We submitted exactly twice.

Table 3 Per-domain results of the winning submission. **ANLS** = official per-domain score (the eight values average to the overall 60.00%). **Changed by refinement** = how many of the 20 questions per domain received a different answer between the first run (55.63%) and the second, refined run (60.00%). The system is strongest on text-heavy documents that IDA anchors and weakest on comics. With only 20 questions per domain, these per-domain values carry wide confidence intervals (roughly ±20 pp); read the ordering as indicative, not exact (Section 4.4).

Domain	ANLS	Changed by refinement (of 20)
business_report	0.75	2
science_paper	0.70	0
engineering_drawing	0.65	6
infographics	0.65	4
slide	0.60	0
maps	0.50	0
science_poster	0.50	5
comics	0.45	0

Table 3 gives the per-domain ANLS and shows where the refinement pass had effect. The second run changed most answers in visually complex domains (engineering drawings, science posters, infographics), where focused crops helped; text-heavy domains were already stable.

The refinement trade-off is significant: the second run consumed approximately 5–8× more compute time and tokens than the first. This is acceptable in a competition setting; for production deployment, selective application of deep iterations to low-confidence questions would be the appropriate strategy.

4.4 Limitations

The baseline and the component split are approximate, by design. The ≈40% naive-ensemble reference point and the ≈ +7 pp / ≈ +13 pp decomposition are estimates, not measurements. IDA and the orchestration are tightly coupled and cannot be switched off independently without rebuilding the pipeline, and these values shift with every change to the reader set, the prompts, or the configuration. A single exact number would therefore misrepresent a quantity that is inherently configuration-dependent, and at $n = 160$ the statistical uncertainty is large in any case. We report them as approximate, supported by internal experiments and by which domains each component affects, and we do not claim precise contributions.

Small sample size (set by the benchmark). The DocVQA 2026 test set has 160 questions, 20 per domain. This sample size is the competition organizers' design (CVC); we present and interpret the official scores, we do not control the sample size. At $n = 160$ the 95% confidence interval on an aggregate ANLS value is about ±7 pp, so differences below roughly 5 pp should not be over-interpreted; the 4.37 pp refinement gain (55.63% → 60.00%) sits within this range, and we report it as an

observed improvement, not a statistically significant one. Per domain the uncertainty is much larger: at $n = 20$ the 95% interval is roughly ± 20 pp, so the per-domain scores in Table 3 and their ordering are indicative, not precise.

Single-run results; no variance estimate. Foundation-model outputs are stochastic, so the scores would vary somewhat across repeated runs. We did not measure this variance: a complete run over the five-reader ensemble is expensive in frontier-model API tokens, and we completed only three full runs in total during the competition (on the order of several thousand euros). The reported numbers are therefore single-run results, to be read alongside the sampling interval above.

Test-set interaction. The second submission used the first run's reasoning outputs as prior context. Both runs operated on the same competition test set. This constitutes a mild form of test-set adaptation: the second run's starting point is informed by the first run's processing of the same documents. We did not use leaderboard feedback (the leaderboard returns only an aggregate score) to select which questions to revisit or which parameters to change. The enhancement decisions (focused crops, deeper iterations, parallel sub-queries) were made before the second run. We state this explicitly because the design pattern could in principle introduce test-set leakage if the first run's errors were used to diagnose and correct specific test cases — which we did not do.

The battlecard is qualitative and internal. The reader strengths in Table 1 (strong / medium / weak) are qualitative judgments from our internal validation, not official per-model scores. They motivate the trust orders (Table 2) but are not themselves measured claims. The official per-domain ANLS of the full system is in Table 3.

Validation-set configuration tuning; no weight fine-tuning. The battlecard (Table 1), the sub-agent prompts, and the agentic-flow parameters were tuned on the competition *validation* set; this is what the leaderboard label "FT" refers to. No model-weight or LoRA fine-tuning was performed, and the test set was not used for any tuning. Beyond setting the battlecard and trust orders, we did not run a systematic prompt search, per-domain reader specialization, or hyperparameter optimization, and the system runs with uniform settings across all eight domains. The configuration is deliberately simple, so the score is likely a lower bound on what the architecture reaches with deeper optimization.

Continuous learning scope. DCA's framework includes prediction-error-driven adaptation. In this submission, adaptation operated at the configuration and context level only: the battlecard and prompts were set on the validation set (above), and the Reasoning Agent updates its prompt-context representation between iterations. Weight-level adaptation (LoRA) was not part of the submitted system and is listed as a future direction (Section 4.5).

Table 4 Optimization directions not exploited in the current submission. Effect sizes are hypothesized from domain knowledge, not measured.

Dimension	Direction	Expected effect (hypothesized)
Accuracy	LoRA fine-tuning of readers; per-domain specialization	Likely +10–20 pp on weak domains
Latency	Domain-aware routing; smarter page selection	Estimated 2–3× speedup on large documents
Cost	Open-weight fine-tuned readers; keep Opus 4.6 for reasoning	Estimated 5–10× cost reduction

4.5 Outlook

The current submission does not exploit several optimization directions that should improve accuracy, latency, and cost. These are listed in Table 4 and noted here as hypotheses rather than measured projections.

5 Conclusion

We presented a multi-agent document-reasoning system that placed first in the General category of DocVQA 2026 with 60.00% ANLS ($n = 160$; 95% CI $\approx \pm 7$ pp). The system combines IDA's deterministic layout-aware text extraction with focused VLM reads from four frontier models and a Claude Opus 4.6 reasoning agent that applies domain-specific trust hierarchies and cross-perspective conflict detection.

The same frontier models used in a naive mixture-of-experts configuration reach approximately 40% ANLS. The +20 pp improvement over that baseline comes entirely from the architectural layer — IDA as a precision input anchor and the multi-agent orchestration — not from larger or newer models. We attribute $\approx +7$ pp to IDA and $\approx +13$ pp to orchestration, and report these as approximate estimates rather than measurements (Section 4.4).

Two observations merit emphasis. First, the bottleneck in multi-domain document reasoning is not model capability but orchestration: VLMs possess the requisite knowledge but require structured context management, conflict detection, and domain-aware routing to apply that knowledge correctly. Second, intermediate-reasoning transparency is a structural requirement for ensemble conflict resolution: models that commit to answers without exposing reasoning steps (as GPT-5 did in our testing) are harder to orchestrate, regardless of their raw accuracy.

Both submissions used a battlecard and prompts tuned on the validation set, but no weight-level fine-tuning, systematic prompt search, or per-domain specialization. The gap between current performance and the hypothesized ceiling with LoRA fine-tuning and per-domain specialization suggests that 60% is not close to this system's limit on this benchmark.

Acknowledgments. This work depends on the frontier foundation models that serve as readers and reasoning agent: Google's Gemini family, Anthropic's Claude

family, and Alibaba's Qwen. The competition organizers (CVC / ICDAR 2026) created and administered the DocVQA 2026 benchmark. The Luna system executed the DocVQA 2026 pipeline end-to-end: IDA processed the documents, the VLM ensemble performed the page reads, and the Reasoning Agent synthesized the final answers. All answers submitted to the official leaderboard were generated autonomously by Luna. The human author designed the system, supervised its operation, and is responsible for all methodological decisions and results reported here.

Declarations

Funding. The author received no specific external grant for this work; it was carried out within Planet AI's research and development activities.

Competing interests. The author is Chief Technology Officer of Planet AI GmbH. The multi-agent system described in this paper is a component of Luna, a commercial AI platform developed by Planet AI, and IDA is a commercial product of Planet AI. In addition, Luna was used as an AI assistant in preparing this manuscript (see the AI-assistance disclosure below). These interests did not affect the objectivity of the reported measurements: the primary result is the official, independently scored DocVQA 2026 competition outcome.

Ethics approval. Not applicable. This research did not involve human participants or animals; the dataset consists of publicly released documents provided by the competition organizers.

Data availability. The DocVQA 2026 competition dataset is publicly available (HuggingFace: `VLR-CVC/DocVQA-2026`; no formal DOI issued). The official leaderboard results are maintained by the competition organizers (CVC / ICDAR 2026).

Code availability. The Luna system is a commercial production system and is not publicly released. Models used: Gemini 3.1 Pro (Google DeepMind, as available at competition submission date, 2026); Gemini 2.5 Pro [6]; Claude Sonnet 4 [2]; Claude Opus 4.6 [1]; Qwen3.5 [14]; Gemini Flash (Google DeepMind, 2026, for question reformulation). IDA version: production release as of 2026. Compute: standard API access with per-provider rate limiting. The paper describes the architecture and the mechanisms that drive the result (Section 3) at a level meant to convey the design and support the paper's claims; it does not provide the full prompt suite or detailed control flow, which are extensive and specific to a production system. The contribution is architectural rather than an artifact to be re-run: the finding is that structured multi-agent orchestration over frozen foundation models, rather than a larger model, produces the result, and this does not depend on bit-exact reproduction.

Authors' contributions. W.W. is the sole author: he conceived and designed the system, directed its implementation and operation for the DocVQA 2026 competition, analysed the results, and wrote the manuscript.

AI-assistance disclosure. No AI system is an author of this work. Luna (Planet AI), an agentic AI system, was used as a tool in preparing this manuscript (drafting, analysis, and editorial review). Final accountability for all content, methodological

decisions, reported results, and conclusions lies with the human author. Frontier models are also described in the manuscript as components of the system under study (readers and reasoning agent); that is methodology, not authorship.

Appendix A ANLS Metric

ANLS (Average Normalized Levenshtein Similarity) is defined per question q as:

$$\mathrm{ANLS}(q) = \begin{cases} 1 - \dfrac{\mathrm{NL}(a_{\mathrm{pred}}, a_{\mathrm{gt}})}{\max(|a_{\mathrm{pred}}|, |a_{\mathrm{gt}}|)} & \text{if } \dfrac{\mathrm{NL}(a_{\mathrm{pred}}, a_{\mathrm{gt}})}{\max(|a_{\mathrm{pred}}|, |a_{\mathrm{gt}}|)} < \tau \\ 0 & \text{otherwise} \end{cases} \tag{A1}$$

where NL is the normalized Levenshtein distance on lowercased strings, $\tau = 0.5$ is the threshold, and the per-question scores are averaged over all questions. ANLS is a soft similarity measure rather than exact match: small surface deviations (spacing, capitalization) reduce but do not zero the score, while a sufficiently different answer receives no credit. The metric and the $\tau = 0.5$ threshold were introduced for scene-text VQA by Biten et al. [4] and adopted by DocVQA [11]; the DocVQA 2026 competition uses the same definition.

Appendix B Wilson-Score Confidence Interval

For a proportion $p = 0.60$ with $n = 160$, the 95% Wilson-score confidence interval is:

$$\frac{p + \dfrac{z^2}{2n} \pm z\sqrt{\dfrac{p(1-p)}{n} + \dfrac{z^2}{4n^2}}}{1 + \dfrac{z^2}{n}} \tag{B2}$$

with $z = 1.96$ for 95% confidence. This yields approximately $[0.527, 0.670]$, or ± 7 pp around 60%. The 55.63% baseline has a similar interval $[0.479, 0.632]$. The overlap of the two intervals means the 4.37 pp refinement gain does not reach formal statistical significance at $n = 160$.

References

[1] Anthropic (2026) Claude Opus 4.6 system card. Anthropic

[2] Anthropic (2026) Claude Sonnet 4 system card. Anthropic

[3] Apple Research (2025) The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity. https://ml-site.cdn-apple.com/papers/the-illusion-of-thinking.pdf

[4] Biten AF, Tito R, Mafla A, et al (2019) Scene text visual question answering. In: IEEE/CVF International Conference on Computer Vision (ICCV), arXiv:1905.13648

[5] DocVQA 2026 Competition (ICDAR 2026) (2026) Competition for multimodal reasoning over documents in multiple domains. https://www.docvqa.org/challenges/2026. Dataset: HuggingFace `VLR-CVC/DocVQA-2026`; submission platform RRC Channel 34, https://rrc.cvc.uab.es/?ch=34

[6] Google DeepMind (2025) Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next-generation agentic capabilities. ArXiv:2507.06261

[7] Huang Y, Lv T, Cui L, et al (2022) LayoutLMv3: Pre-training for document AI with unified text and image masking. In: ACM International Conference on Multimedia (MM), pp 4083–4091, https://doi.org/10.1145/3503161.3548112, arXiv:2204.08387

[8] Kim G, Hong T, Yim M, et al (2022) OCR-free document understanding transformer (Donut). In: European Conference on Computer Vision (ECCV), arXiv:2111.15664

[9] LandingAI (2024) Superhuman accuracy on DocVQA with agentic document extraction. Developer blog, https://landing.ai/blog/superhuman-on-docvqa-without-images-in-qa-agentic-document-extraction. No peer-reviewed report available

[10] Lee K, Joshi M, Turc I, et al (2023) Pix2Struct: Screenshot parsing as pretraining for visual language understanding. In: International Conference on Machine Learning (ICML), arXiv:2210.03347

[11] Mathew M, Karatzas D, Jawahar CV (2021) DocVQA: A dataset for VQA on document images. In: IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp 2199–2208, https://doi.org/10.1109/WACV48630.2021.00225, arXiv:2007.00398

[12] OpenAI (2025) GPT-5 system card. https://openai.com/index/gpt-5-system-card/

[13] Packer C, Wooders S, Lin K, et al (2023) MemGPT: Towards LLMs as operating systems. ArXiv:2310.08560

[14] Qwen Team, Alibaba (2025) Qwen3 technical report. ArXiv:2505.09388

[15] Shinn N, Cassano F, Berman E, et al (2023) Reflexion: Language agents with verbal reinforcement learning. In: Advances in Neural Information Processing Systems (NeurIPS), arXiv:2303.11366

[16] Tito R, Mathew M, Jawahar CV, et al (2021) ICDAR 2021 competition on document visual question answering. In: International Conference on Document Analysis and Recognition (ICDAR), LNCS, vol 12824. Springer, pp 635–649, https://doi.org/10.1007/978-3-030-86337-1_42, arXiv:2111.05547

[17] Wang J, Wang J, Athiwaratkun B, et al (2024) Mixture-of-agents enhances large language model capabilities. ArXiv:2406.04692

[18] Wang X, Wei J, Schuurmans D, et al (2023) Self-consistency improves chain of thought reasoning in language models. In: International Conference on Learning Representations (ICLR), arXiv:2203.11171

[19] Wei J, Wang X, Schuurmans D, et al (2022) Chain-of-thought prompting elicits reasoning in large language models. In: Advances in Neural Information Processing Systems (NeurIPS), arXiv:2201.11903

[20] Wustlich W (2026) Distributed cognitive architecture (DCA) — foundations. Zenodo, https://doi.org/10.5281/zenodo.20738104

[21] Wustlich W (2026) Distributed cognitive architecture (DCA) — theory I: Atomic agents, fractal composition, convergence. Zenodo, https://doi.org/10.5281/zenodo.20732538

[22] Xu Y, Li M, Cui L, et al (2020) LayoutLM: Pre-training of text and layout for document image understanding. In: ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), pp 1192–1200, https://doi.org/10.1145/3394486.3403172, arXiv:1912.13318

[23] Yao S, Zhao J, Yu D, et al (2023) ReAct: Synergizing reasoning and acting in language models. In: International Conference on Learning Representations (ICLR), arXiv:2210.03629