

RDB-1: Regulated Document Benchmark

Do you have to trade data security for quality? RDB-1 puts models that run inside an aviation company's own perimeter next to cloud models and a general agent tool, on one task: a technical report a domain expert would sign. It measures that on a consolidated data set from a completed regulatory assessment: the same files, the same brief, the same template for every run, scored against the sources.

USED DATA SIZE

96,180

EXCEL CELLS
12 WORKBOOKS

8,561

PDF PAGES
116 FILES

872

THREADED REVIEW COMMENTS
365 LOGGED CHANGES

1. The data set

One regulatory framework was screened requirement by requirement for applicability to a new aircraft design, reviewed line by line by domain experts, and consolidated with a full change log and threaded comments. The data set adds the technical baseline documents behind the assessment, the scope document that defines the assignment, and the regulation library needed to interpret it. The target goal of the task is the technical end report: what the assessment found, what is still open, what has to be decided, on a controlled document template under a fixed set of rules.

Nothing is pre-digested: the Excel workbooks are too large to read without a script, the decisive standard is 258 pages long, the planning documents contradict each other in places, and the brief deliberately contains one wrong statement about the assignment structure that the scope document corrects.

MATERIAL	FILES	SIZE	WHAT A DILIGENT AGENT HAS TO DO WITH IT
Assessment Excel workbooks, two versions	2 xlsx	68,219 cells	Reproduce the applicability statistics, diff the versions cell by cell (32 real changes among 365 logged events), read 872 threaded comments, aggregate a free-text column with 28 review items.
Domain compliance matrices	8 xlsx	17,029 cells	Understand a shared template, count what is and is not covered, notice rows never marked complete.
Framework comparison Excel workbooks	2 xlsx	10,932 cells	Count paragraph pairings between two regulatory frameworks without inventing a number.
Technical planning and design documents	4 pdf	106 pages	Extract the design baseline and its open points; spot the internal inconsistencies.
Scope document	1 docx	1,827 words	Find the real assignment structure, which contradicts the brief.
Regulation library	112 files	8,455 pages	Locate two decisive clauses in a 258-page standard, its scope clause, and a means-of-compliance issue that changes the safety argument.
Document template	1 docx	-	Use its styles, document-control front matter and fields correctly; leave no placeholder behind.

2. Rules and protocol

RULES IN THE BRIEF

- **No recommendation, no effort sizing.** State the gaps in plain language and leave the decision to the reader.
- **Authority frame.** A specific authority frame, named in the brief, must be applied and not merely mentioned.
- **House style.** Fixed typographic and document-control conventions. The report covers the whole assignment, not one work package.
- **Process.** Plan the structure first and discuss it before writing.

PROTOCOL

- One run per configuration on the unchanged data set. A run found to have used benchmark material or an earlier report is excluded.
- The brief is given verbatim. Planning questions are answered minimally and consistently. Harness, model, reasoning setting, date and tool failures are recorded. The .docx the run produces is the only artefact scored.
- The reviewer is an AI system with full source access, working under a human domain expert who challenges every finding. The reviewer's own first draft is part of the field as a reference, scored with the same rubric.

3. Scoring

Each output file goes through an automated pass and a manual read. The checker validates the file against the Office Open XML schema, counts hyperlinks, heading styles, template placeholders, over-long paragraphs and spreadsheet cell addresses in prose, and pattern-matches seventeen stated facts against a ground-truth file computed from the sources. The manual read verifies every further number and claim against the workbooks and PDFs and scores the judgement dimensions. A stated value that contradicts the sources counts against the run; a value the report does not state does not.

What 100 means. Seven dimensions are scored 0 to 5 and weighted to 70 points, then rescaled to 100. The 100 mark is the benchmark reference: the same task completed by an AI teammate and then corrected through three review iterations with a human domain expert. A run at 100 would need no correction.

Retrieved-data fidelity weight 3

Do the figures the report states match the source files? 5: no contradiction. 3: two miscounts. 0: five or more invented or misread values.

Knowledge accuracy weight 2

Are expansions, scope statements and regulatory context right? 5: none wrong. One point off per error.

Analysis depth weight 3

Per-domain differences, quantitative safety framing, the decisive clauses of the standard, the scope question of the standard, the organisational approval question, a decisions register, an authority-agreement list, the full scope of the assignment. 5: all. 3: about half. 0: workbook summary only.

Rule compliance weight 2

All brief rules held, correct document number, the misleading brief line caught, no instruction text leaked into the output document. 5: all. 3: one lapse. 0: recommendation or effort sizing present.

Readability weight 2

Real heading styles, a table of contents, paragraphs under about 100 words, no spreadsheet coordinates or operator notes in prose, no template residue. 5: all. 3: readable with residue. 0: unusable structure.

File integrity weight 1

Schema-valid, opens without repair. 5: zero errors. 3: under twenty. 0: repair prompt likely.

Independence weight 1

Sentence overlap with any earlier report, after excluding text quoted from the sources. 5: under five percent. 0: a rewrite.

Hallucination taxonomy

Fabricated or misread data: a stated figure the sources contradict. **Knowledge error:** a false statement from model memory, such as a wrong acronym expansion. **Omission:** relevant source content not used. The first two reduce the data and knowledge scores; omissions reduce depth.

4. Results

Four local models and one cloud model on Wingman360 Teammate, plus one reference on a general-purpose agent tool, all first outputs before any human correction, scored 6 September 2026. The 100 mark is the reference after three review iterations.

TOTAL SCORE OUT OF 100

GPT 6 Astra		74
WINGMAN360 TEAMMATE · CLOUD MODEL		
DeepSeek V4 Flash (local)		74
WINGMAN360 TEAMMATE		
GLM 5.3 Flash, high reasoning (local)		71
WINGMAN360 TEAMMATE		
Qwen 3.8 Flash (local)		66
WINGMAN360 TEAMMATE		
Claude Fable 5.1		66
REFERENCE · GENERAL AGENT TOOL		
GLM 5.3 Flash, low reasoning (local)		40
WINGMAN360 TEAMMATE		

WINGMAN360 TEAMMATE REFERENCE: GENERAL AGENT TOOL

SCORES BY DIMENSION, 0 TO 5

	DATA ×3	KNOWLEDGE ×2	DEPTH ×3	RULES ×2	READAB. ×2	FILE ×1	INDEPEND. ×1
GPT 6 Astra	5	5	4	4	1	0	5
74 · CLOUD							
DeepSeek V4 Flash (local)	4	5	2	5	3	3	5
74							
GLM 5.3 Flash, high (local)	4	4	2	5	3	3	5
71							
Qwen 3.8 Flash (local)	4	2	3	5	2	2	5
66							
Claude Fable 5.1, reference	4	5	2	2	3	3	5
66							
GLM 5.3 Flash, low (local)	1	2	1	3	2	3	5
40							

AUTOMATED SIGNALS AND VERIFIED ERROR COUNTS

RUN	WORDS		SCHEMA ERRORS	FACTS CORRECT	FACTS WRONG	KNOWLEDGE ERRORS	COVERAGE ITEMS	PARAGRAPHS OVER 100 WORDS
		TABLES						
GPT 6 Astra, cloud	7,340	14	336	12 / 17	0	0	8 / 13	1
DeepSeek V4 Flash (local)	4,611	14	8	9 / 17	1	0	6 / 13	5
GLM 5.3 Flash, high (local)	9,821	18	12	10 / 17	1	1	6 / 13	7
Qwen 3.8 Flash (local)	15,595	20	27	12 / 17	1	3	9 / 13	27
Claude Fable 5.1, reference	3,644	9	9	6 / 17	0	0	4 / 13	1
GLM 5.3 Flash, low (local)	2,322	9	3	7 / 17	4	3	3 / 13	4

"Facts correct" counts ground-truth values stated correctly; unstated values are not penalised. "Facts wrong" and "Knowledge errors" are the counts confirmed in the manual read. "Coverage items" is out of thirteen required content items. Schema errors are Office Open XML validation findings; one table-border defect from the document tool appears in every file built with it.

5. What separates the runs

- **Fabricated data is the discriminator.** Two runs stated no figure the sources contradict, three stated exactly one, and the weakest stated four: an assumption count that does not exist, transition counts never computed, line counts that contradict themselves within the same document. Raw counts copied from a sheet were right everywhere; interpreted figures were where a model went wrong.
- **Reasoning effort is worth thirty points.** The same model in the same harness moved from 40 to 71 when reasoning was turned up. The extra effort removed the invented figures and the missed workbook column; it did not add domain knowledge.
- **Knowledge errors are model signatures.** Wrong acronym and authority expansions came from model memory, not from the files, and each model had its own: one mis-expanded the same safety-level acronym in two independent runs; another mis-expanded two terms a domain reader meets on the first page. These are the errors an expert notices first.
- **Depth and polish rarely came together.** The cloud model produced the deepest analysis and a file with 336 schema errors and 142 spreadsheet coordinates in prose. The 15,600-word local run read the workbook most thoroughly and made a finding about a gap in the assessment that nobody else made, and produced 27 paragraphs over 100 words with 27 schema errors. The cleanest file came with a one-page analysis.
- **Reading the standard separated the field.** Three runs, two of them local, found the two decisive clauses in the 258-page standard and paraphrased them accurately. None found the means-of-compliance issue that sets the quantitative safety target, so every run left the largest technical difference unquantified.
- **The brief's trap caught one run.** Five of six read the scope document and corrected the wrong assignment structure planted in the brief; the lowest-scoring run copied it into its output.
- **Security did not cost quality.** A frontier model on a general agent tool scored 66. On Wingman360 Teammate the best local model, running inside the perimeter, scored 74, level with the best cloud model on the same harness. The 100 mark is the reference report after three review iterations with a domain expert.

Limits of this edition

RDB-1 v1.0 is a single-domain benchmark with one run per configuration. A score therefore reflects one output of one model on one day, not a distribution over repeated runs. The manual dimensions carry reviewer judgement even with the anchors given in section 3. The reviewer is an AI system working under a human domain expert, and the reference run is that system's own first output on a general agent tool; both facts are disclosed and the reference is scored with the same rubric as every other run. The brief contains one wrong statement by design. Reasoning settings and token budgets were left at each configuration's default unless stated. Scores will move as configurations are added; the version and date are on every page.

Lessons learned from this study

Local models with the right harness perform at the level of frontier cloud models on this task. The best local run on Wingman360 Teammate scored 74, level with the cloud model on the same harness and above a frontier model on a general agent tool, with no data leaving the company's perimeter. Data security and sovereignty did not cost quality.

The harness carries much of that result. Built-in aviation skills and an aviation-optimised data ingestion and retrieval strategy give a local model the regulatory context, the workbook structure and the document rules it cannot hold on its own. That is what lets a model running inside the perimeter match a frontier cloud model on a task that no configuration completed without correction.

Want a model or agent scored on RDB-1? Write to info@lavionic.com. Runs are performed on the unchanged data set and published with the same rubric.