A multi-agent system for technical research

Answers to engineering questions built only from findings that carry a verbatim quote from their source.

Jacek Szybicki

01

PURPOSE

Why the system exists

Language models answer technical questions fluently, but a fluent answer does not show where a number or a design rule comes from. This system is built the other way round: every finding it keeps carries a verbatim quote from a page that was actually read, and a link to that page. Reports, tables and design rules are then written only from such findings.

Scope: engineering questions that need evidence from scientific papers, manufacturers' technical documentation and the web — for example machine vision on embedded hardware, sheet-metal design rules or production documentation.

SCALE SO FAR

45
Technical studies

76,627
Findings accepted by the verifier

10,989
Unique source addresses

1,873 / 12,646
Hypotheses / information tests

02

COURSE OF A STUDY

How a study runs

A study starts from a written commission: the topic, the goal, the sections or questions, what must be covered, what to avoid, a glossary and a time budget. The plan splits the commission into cells, one per section.

01 Cell owner

Breaks the problem into sub-questions, sets out 3–6 competing hypotheses, including a null hypothesis, and plans information tests. At least a third of the tests look for evidence against the hypotheses. The owner closes the cell when its model stops changing.

02 Researchers

Run the tests: search, choose pages by expected information gain, read them and write claims, each with a verbatim quote. A free researcher takes the highest-priority test from any open cell.

03 Sceptic

Raises alternative explanations as challenge tests, so that a hypothesis is not confirmed only by the evidence that was looked for.

04 Synthesiser

Maintains the cell's model of the answer as findings arrive.

05 Verifier

Checks every claim against its quote in the context of the page (section 04).

06 Agent profiles

Each test gets a profile: temperature, depth of reasoning and a working strategy (primary sources, breadth, numbers, counter-evidence). After three tests per profile the system favours the profile with the best average gain after verification.

The time budget is split between research and closing (verification and final reviews). The number of parallel model calls equals the serving slots, with one eighth reserved for the steering roles, so that many readers do not hold up decisions. The direction of a running study can be corrected with written directives, without restarting it.

03

RULES ENFORCED IN CODE

Code guards the facts, not the strategy

- **Quote on the page:** a claim is accepted only if its quote stands on the page that was read, and the subject of the claim appears on that page.
- **Numbers need a source:** a number in a hypothesis must come from a finding.
- **Confidence is computed, not asserted:** the confidence of a hypothesis is calculated from independent domains and source types, and evidence against weighs more than evidence for.
- **Age of evidence:** evidence older than a set number of years (5 by default) does not confirm a hypothesis.
- **Images:** a claim read from an image has no text quote to compare, so a second, independent reading of the same image must confirm it.

Who said it: source types of accepted findings

SOURCE TYPE	FINDINGS
Manufacturers and vendors	33,457
Independent sources	31,807
Media	3,103
Aggregators	2,526
Regulators	1,961
Operators	1,061
Customers	721
Tenders	162
Not classified	1,829
All accepted findings	**76,627**

A quote found on the page shows that the claim is in the source, not that the source is right. That is why the source type is recorded for every finding and why vendor claims are kept apart from independent confirmation.

04

QUOTE VERIFICATION

A small fine-tuned verifier

The verifier receives the page title and address, the quote and the claim, and answers **Yes**, **Partly** (with the claim trimmed to what the quote supports) or **No**. It is Qwen3.5-4B fine-tuned with LoRA (r = 16) on 1,700 judgments made by Claude (Sonnet), and it replaced the general 27B model in this role.

VERIFIER, 300 TEST FINDINGS FROM PAGES NOT SEEN IN TRAINING	AGREEMENT WITH CLAUDE	PRECISION OF “YES”
Qwen3.8-27B (previous verifier)	0.674	0.640
Qwen3.5-4B without fine-tuning	0.861	0.819
Fine-tuned verifier v1	**0.929**	**0.918**

Agreement across all three classes is 0.917. Served with vLLM, the verifier handles 19 judgments per second on a single RTX 3090. The reference for these scores is Claude’s judgment, not a human expert panel.

05

FINDING AND READING SOURCES

Reading technical sources in full

Discovery

- **Web search** through a private SearXNG instance, with a separate pause per search engine; no paid search APIs.
- **Scholarly databases:** Europe PMC, arXiv, Crossref and OpenAlex lead to open-access full text (PDF or JATS XML), or to the abstract when the publisher blocks the page. Semantic Scholar is integrated and waits for an API key; without a key the shared pool returns rate-limit errors.
- **Site maps** of the sources named in the commission, and official APIs such as Wikipedia and Stack Exchange.

Reading

- **Long documents:** PDFs of up to 80 MB and 3,000 pages are indexed by paragraph with page markers, and a researcher receives the passages that best match the test, not the beginning of the file.
- **Technical drawings in PDF:** pages that look like drawings are analysed for dimensions, tolerances and fits, scale, title block, notes and tables such as bills of materials and bend tables.
- **CAD files:** STEP models are measured with OpenCascade and DXF drawings with ezdxf in an isolated sandbox without network access: wall thickness, bends with radius and angle, holes, hole-to-bend distances, cut length.
- **Safety:** the browser connects only through a local guard proxy that allows public addresses on ports 80 and 443. CAPTCHA and bot challenges are not bypassed.

06

INFRASTRUCTURE AND OUTPUT

Self-hosted model, rented GPUs

The agents run on the open Qwen3.8-27B model served with vLLM on GPUs rented by the hour. A watchdog removes the instance when no study has been running for 15 minutes, so the GPU is paid only while it works. A typical study costs 1–3 USD of GPU time; the main license plate recognition study ran for 3.7 hours on two RTX PRO 6000 cards for USD 13.08.

Output of a study: report sections with numbered citations, all findings with their verbatim quotes, the list of sources, a list of gaps the study could not close, and optional tables in which every value carries its quote and source.

07

STUDIES IN TECHNICAL DOMAINS

Selected studies

STUDY	FINDINGS	SOURCES
License plate recognition on one single-board computer, without a GPU	16,854	1,371
Design rules for manufacturable sheet-metal enclosures	7,249	1,558
Single-lane plate recognition module at a barrier column	3,345	331
Industrial design of enclosures for self-service machines	3,268	570
Design knowledge for a CAD agent working in FreeCAD	1,040	254
Modelling parts for 3D printing in FreeCAD	850	227
Corners and reliefs in parts bent from one sheet	777	107
Purchased components mounted in sheet-metal enclosures	597	127
What to hand over to a sheet-metal contractor	541	138
Press brake tooling compatible with Amada presses	441	75

Counts are findings accepted by the verifier and distinct source pages behind them. The license plate recognition study is described in a separate case study.