RESEARCH PAPER

The Last Human Gate

Forward Deployed Engineering for Governance Automation

Before a company buys a technology solution, connects systems or launches an app, specialists must review the project at a series of required checkpoints, called gates. Can an AI agent carry out these reviews well enough that a specialist does not have to redo them? If so, what would that mean for the people who do this work today?

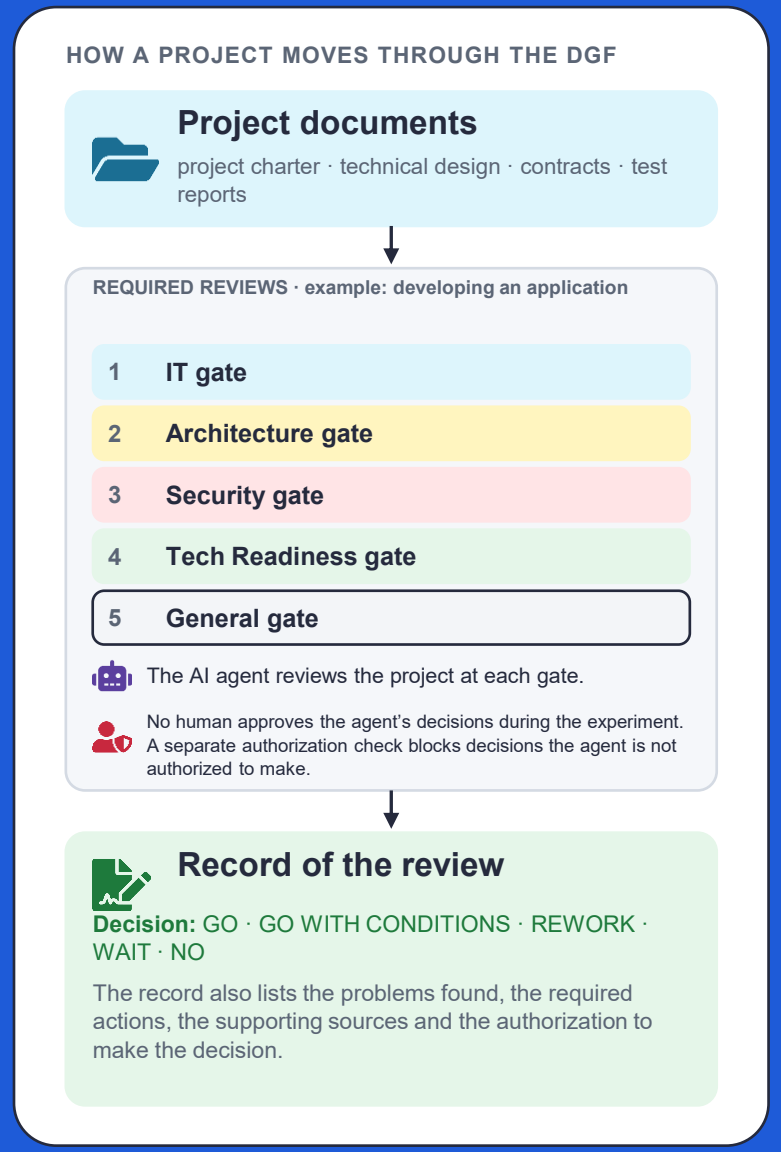
Jeremy Canale

AI Applied Security Architect · contact@jeremycanale.com

arXiv:2609.29345 · Read the paper

DGF-Bench on GitHub · jeremy1392/dgf-agentic-bench

Thesis (66 p.), data and execution logs in the DGF-Bench repository · September 2026



Four key points

If you only read one slide, read this one.

The context

Before a company buys a technology solution, connects two systems or builds an application, specialists review the project at a series of checkpoints covering purchasing, legal requirements, compliance, security, IT, architecture and operational readiness. The paper calls this series of project reviews a Digital Governance Framework (DGF).

The idea

The paper argues that these reviews are assigned to people today, but do not necessarily have to be performed by people. An AI agent can read the same project documents, apply the same written rules and reach the same decision. People retain the decisions that still require human involvement. Forward Deployed Engineers (FDEs) help companies put AI into use. The paper proposes that DGF reviews are a likely starting point for their work.

The experiment

The same AI agent reviewed 300 realistic but fictional projects using three different AI models. A program scored each response against answers the agent could not access. With the best-performing model, the agent made the correct decision at every gate. Its reviews met all five scoring criteria at 95% of gates, and every review passed for 77% of projects.

The limit

Automating most cases does not necessarily eliminate most jobs. People may still handle the most time-consuming cases, check the agent's work, correct its errors and maintain the AI agents. The author predicts that this work will require 80% fewer full-time-equivalent staff by 2033 and sets out how to test that prediction.

Every required project review must still be carried out, but it can be performed by a specialist or an AI agent. This deck explains the FDE role, the reviews that make up a DGF, the experiment’s findings and the possible effects on jobs.

Six key terms explained

The paper uses several terms repeatedly. Here is what each one means.

Forward Deployed Engineer (FDE)

An engineer who works directly with a client's teams to put AI into use in their day-to-day processes and systems. Rather than only advising or developing software remotely, the FDE turns a demonstration into a system the company can use every day. The role covers many business processes, not just governance.

Digital Governance Framework (DGF)

The required reviews a project goes through before the company commits to it: Procurement, Legal, Compliance, Security, IT, Architecture and Tech Readiness, followed by the final overall review, called the General gate. It is similar to customs checks or the approvals needed for a building permit.

Gate

A required project checkpoint where a specialist reads the documents, identifies problems, requests corrections, cites supporting sources and makes a decision. Each gate has written rules that define what to check and which decisions the reviewer is allowed to make.

AI agent

An AI agent uses a language model to do more than answer questions. It can open documents, search company systems, request missing information and explain its decisions. In the experiment, the AI agent carries out the reviews normally assigned to specialists.

Project documents (the "dossier")

The documents available to the reviewer: the project charter, architecture diagram, contracts, security test reports, backup and recovery test results, and supplier proposals. As in real projects, documents may be missing, outdated or inconsistent.

The five possible decisions

GO: proceed. GO WITH CONDITIONS: proceed subject to the stated conditions and deadlines. REWORK: revise the proposal and submit it again. WAIT: pause until a requirement is met. NO: do not proceed with the proposal as submitted. Correctly rejecting a project can count as a successful review.

Two other terms matter: "evidence" means the documents or records that support a conclusion; "authority" means permission to make a particular decision. A correct decision also needs supporting sources and the required authorization.

What a Forward Deployed Engineer does

How the cited industry sources describe the role. FDEs work across business processes, not just governance.

WHAT THE ROLE INVOLVES

An engineer who works directly with a client to build, adapt and deploy technical solutions in the systems the client uses every day. The role combines software development, an understanding of the client's business and direct work with users, from gathering requirements to design, implementation, integration and deployment.
Wikipedia, "Forward Deployed Engineer"

Palantir, which popularized the term, compares the role to a startup's chief technology officer: small teams responsible for delivering important projects from start to finish. A conventional product developer builds a feature for many customers. An FDE works across features and systems to meet one customer's needs.
Palantir Technologies, Forward Deployed Software Engineer role description

OpenAI set up an FDE team in 2025 and launched "The Deployment Company" in 2026, placing engineers within client organizations to implement AI systems and workflows. The Wall Street Journal described FDE as the hottest job in tech. Job postings on Indeed increased more than tenfold in 2025.
OpenAI (2026); The Wall Street Journal, via Indeed

DAY-TO-DAY WORK

1. **Work directly with the client's teams.** They join the teams' day-to-day work, on site or remotely, to understand how the work is actually done rather than relying only on a requirements document.
2. **Build and deploy a working solution for the client.** Connect AI agents to the client's records, identity systems and data, adapt them to local ways of working and fix integration problems in production. For example, a documentation agent connected to a hospital's medical records.
3. **Take responsibility for delivery from start to finish.** The role combines consulting to identify the problem, product management to decide what to build, and engineering to build and deploy it. It is an engineering role, not a sales role: the work involves writing code, not meeting a sales quota.
4. **Get the solution working and keep it running.** They support the solution throughout its life, fix problems during use and adapt it as the client's needs become clearer.
5. **Use lessons from each deployment to improve the product.** Improvements developed for one client help make the platform better for other clients.

Sources: DataCamp, "What is a Forward Deployed Engineer?" (2026); Aced / Exponent, interview with a former Palantir FDE (2026); The New Stack (2026).

In this paper, FDEs can apply AI to any business process; governance is not their only role. The author proposes that DGF reviews are a likely starting point for FDEs. Slide 7 explains why.

The reviews a project must go through

Three typical project types: buying a solution, connecting existing systems or developing an application.

BUY
Buy a technology solution
6 required reviews

Procurement → Legal → Compliance → Security → IT → General

INTEGRATE
Connect existing systems
6 required reviews

IT → Architecture → Security → Legal → Compliance → General

BUILD
Develop an application
5 required reviews

IT → Architecture → Security → Tech Readiness → General

FIVE DECISIONS THE SPECIALIST CAN MAKE

GO
approved: the project can proceed

GO WITH CONDITIONS
approved subject to the stated conditions

REWORK
revise the proposal and submit it again

WAIT
put the project on hold until the required condition is met

NO
do not proceed with the proposal as submitted

Like officers carrying out different customs checks, specialists examine the same project documents for different reasons. The final reviewer at the General gate considers their conclusions and makes the overall decision. Companies may add, combine or reorder these reviews, but must still meet the underlying requirements.

Eight reviews, eight questions

The question each specialist must answer and the information needed to answer it.

GATE	QUESTION TO ANSWER	DOCUMENTS AND RECORDS TO CHECK
Procurement	Does the supplier meet the selection and pricing requirements, and has it passed the required background and sanctions checks?	supplier proposals, mandatory requirements, background checks, sanctions lists and total cost over three years
Legal	Do the contract terms meet the company's requirements, and is the person signing authorized to commit the company?	data-protection agreement, liability terms, termination clauses and the signer's authorization
Compliance	Does the project meet the regulatory requirements that apply to it?	privacy impact assessment, data storage locations, applicable regulations and records of decisions and actions
Security	Are the security measures appropriate for the data being handled and the systems' exposure to threats?	identity and access controls, private network access, monitoring and known vulnerabilities
IT	Is the solution compatible with the company's systems, and can the company operate and support it?	approved technologies, available capacity, support responsibilities and records of system changes
Architecture	Do the system's components work together as intended, and does the design meet the company's technical requirements?	overlapping IP address ranges, system interfaces, data ownership, response times and the ability to move away from the solution
Tech Readiness	Have tests shown that the solution is ready to operate?	load, backup and restore tests; recovery plan; operating instructions; and handover to the operations team
General	Considering the specialist reviews, budget and company strategy, can the project proceed?	earlier review decisions, budget approval, alignment with company strategy and the change plan

The security and legal reviewers examine the same project documents, but each checks for different risks. They are not repeating the same review: each is responsible for a different area.

Why DGF reviews could be a starting point for FDEs

FDEs can apply AI to any business process. The paper gives four reasons why they might start with governance reviews.

1

The same work comes up repeatedly

WHY THIS HELPS

Hundreds of projects each year require the same reviews and similar documents. The team can reuse a review solution across projects and apply the same approach at other companies.

WHAT CAN PREVENT AUTOMATION

Information needed for the review has not been written down, or the reviewer cannot access the files.

2

The rules are already written down

WHY THIS HELPS

The company already has written policies, limits and rules about who can approve which decisions. The team can turn these rules into executable checks and prevent the AI agent from approving decisions outside its authorization.

WHAT CAN PREVENT AUTOMATION

Rules contradict each other, or decisions depend on personal judgment rather than written criteria.

3

Review results are passed to the next reviewer

WHY THIS HELPS

Each reviewer passes their findings to the next reviewer. The team can create one shared project record that every reviewer consults and updates, rather than keeping eight copies that people must check separately.

WHAT CAN PREVENT AUTOMATION

Information is lost between teams, or the same work is repeated.

4

The quality of the review can be checked

WHY THIS HELPS

The review can be checked against the known facts of the project. The team can prepare test cases with specialists and check the agent's performance before relying on it.

WHAT CAN PREVENT AUTOMATION

Scoring rewards a well-formatted response instead of checking whether the decision is correct.

FDEs must also put their own AI projects through these reviews before deployment. Automating the reviews would help their later projects move through the process faster. This requires someone responsible for the entire process, not just one department, and the FDEs' own working hours must be included in the cost. This is a hypothesis about where companies will adopt AI first, not an order established by observation.

The project review remains mandatory. A specialist or an AI agent can carry it out.

A job title tells us who currently does the work, not whether the work must be done by a person. An AI agent must meet three conditions before it can take over a review.

1

The agent can access the information it needs

The AI agent must be able to access the information required for the review or request it when it is missing. The agent must not assume that a backup has been tested. It must check the test record or ask to see it. A file is useful only if the agent has permission to open it.

2

The agent performs the review reliably

It must handle difficult cases, incomplete files and projects that should be rejected, not just straightforward cases. The agent must also cite its sources and acknowledge uncertainty. Completing a form correctly does not mean the decision itself is correct.

3

The agent makes only decisions it is authorized to make

The agent proposes a decision. A separate authorization check allows it only if the agent has the required authorization. A designated person remains accountable for the decision. The agent can request approval, but that does not mean the request will be authorized.

To reduce staffing needs, a fourth condition must be met: the total amount of human work must decrease. Removing a reviewer does not save work if the project team, suppliers or another review team spend the same number of hours doing it instead.

A draft that a person must redo assists the reviewer. An AI review accepted as the final result replaces the reviewer’s task.

Fictional projects, automatically scored reviews

A program creates each test project. The AI agent reviews it, and a scoring program compares its response with answers prepared in advance and kept hidden from the agent.

01 / THE PROJECT

A program creates a fictional company project

- buy a solution, connect systems or build an application
- users, budget, project owners and deadlines
- the sensitivity of the project's data

02 / THE TECHNICAL DESIGN

It creates the project's technical architecture

- applications, networks and user or service identities
- data storage locations and data flows
- backup and recovery arrangements

03 / THE PROJECT DOCUMENTS

It generates the project documents

- project charter, contracts, test reports and diagrams
- some documents deliberately missing or outdated
- official records containing facts needed to verify the documents

04 / THE AGENT REVIEWS

The AI agent examines the project

- opens the documents and architecture diagrams
- checks official records and requests missing information
- identifies missing or conflicting information

05 / THE AGENT DECIDES

It makes a decision

- lists problems, required actions and supporting sources
- checks whether it is authorized to make the decision
- passes its findings to the next reviewer

06 / THE RESULTS ARE SCORED

A program scores the agent's responses

- compares them with answers the agent could not access
- checks the decision, problems, actions, source citations and authorization
- records costs and failures

Why use fictional projects? They contain no confidential company data. The correct answers are known in advance, so every response can be scored against the same rules. The AI agent only reviews the project; it does not create the project it is reviewing. A conventional program, not another AI model, scores the responses using the same rules for every model.

What the agent does after reading the project documents

Requests are checked during each review. Responses are scored after all the project's reviews are complete.

1

The agent lists the problems it identifies

It cites the documents or records that support each finding and states what needs to be done. For example: "There is no record of a successful restore test. The test must be carried out."

2

The agent makes its decision

It can approve, approve with conditions, request corrections, put the project on hold or reject it. In this example, it asks for corrections (REWORK).

3

A request to approve with conditions is checked immediately

The authorization control checks that the agent has the required authorization and that the outstanding risks are eligible for conditional approval. The tool then accepts or rejects the request.

4

The next reviewer receives the earlier findings

The next reviewer carries out a separate review using the previous conclusions exactly as written, including any errors. In the experiment, every scheduled review takes place, even if the project has already been rejected. The General gate then brings all the results together.

5

A Python program scores every review once the project's reviews are complete

It compares five items with the answers prepared for the test project: the decision, the problems reported, the actions requested, the supporting sources and the authorization.

WHAT THE EXPERIMENT DOES NOT MEASURE

Asking for a problem to be fixed does not fix it. The agent may request a restore test, but the experiment does not run that test on a real system.

The experiment checks whether the agent reviews the project correctly and makes the right decision.

To automate the correction as well, the AI agent would have to carry out the authorized action, record the result and review the project again.

During each review, control tools check the agent's requests. After all the project's reviews are finished, the scoring program evaluates its responses. No one manually corrects the agent's work between these checks.

How the program decides whether a review passes

The program creates an answer file with each test project. The agent cannot read it. The scoring program uses it later to check the agent's work.

ANSWERS USED FOR SCORING · 99_hidden_ground_truth.json

- Each test project has a file containing the answers used to score the agent.
- **canonical_truth:** the facts that the program has defined for the fictional project.
- **reference_decisions:** for each review, the decision to make, the problems to report, the actions to request and whether authorization is required.
- When it creates a test project, the program uses the project's facts and its programmed rules to determine the decision to make and the problems to report. It saves these answers for scoring. The program generates these answers automatically; no human reviewer writes them.
- The agent can read the project documents, the information made available to it and the review rules. It cannot open the file containing the answers used to score it. After the reviews, the scoring program reads the answer file. It also takes account of conditional approvals that were properly authorized during the run.

FIVE REQUIREMENTS FOR A REVIEW TO PASS

- The decision alone is not enough. The reported problems, requested actions, source citations and authorization must also meet the scoring requirements.
- A review can pass the test by correctly rejecting the project.
- The scoring program checks the tool-use logs to confirm that the agent opened the sources it cites. It also checks that the required excerpts are quoted correctly.
- A project's full set of reviews passes only if every individual review passes.
- For conditional approval, the tool checks that the authorization covers the review and project phase, that every unresolved problem is included and eligible, and that conditions have been stated. It checks that conditions are present, not that they will work in practice.

EXAMPLE FROM THE TEST DATA · PROJECT FALCON

Project DGF-BLD-035200_build, operational readiness review (BUILD-04-TECH_READINESS-1). The answers prepared for scoring are:

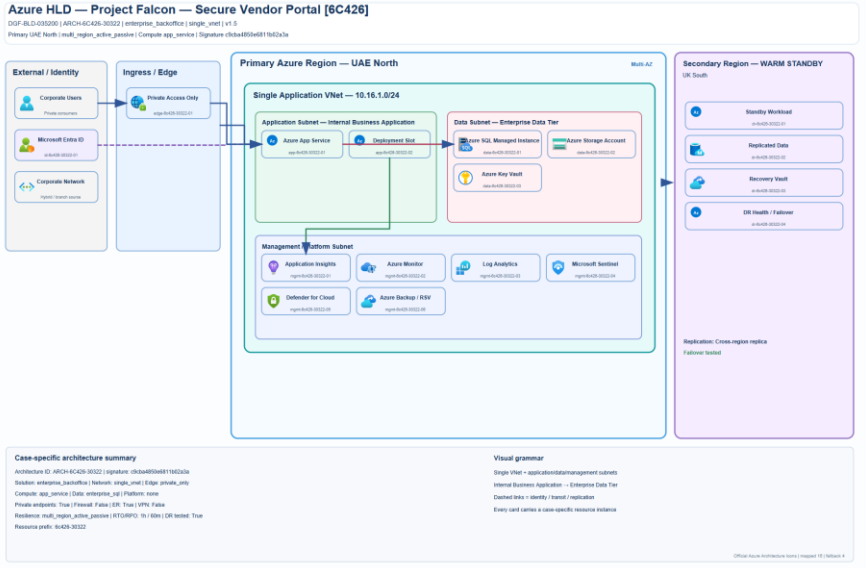
Problem to report: the documents needed for handover to the operations team are incomplete (TR-OPS-001).
Action to request: complete the handover to the operations team (COMPLETE_OPERATIONAL_HANDOVER).
Decision to make: approve with conditions (GO_WITH_RESERVATIONS).
Authorization required: yes.

`experiments/preflight_balanced_300_20260922/dataset/DGF-BLD-035200_build/99_hidden_ground_truth.json`
score_submission.py calculates the scores · synthetic_environment.py checks action requests · openrouter_eval/benchmark_runner.py runs the reviews in order

The same program creates the test project and its answer file. A review passes by meeting the programmed scoring rules, not by persuading a human evaluator.

Realistic project documents, entirely generated

A program generates the project charters, cloud architecture diagrams, penetration test reports, contracts and backup test results.



This architecture diagram comes from the test data: Project Falcon, a fictional HR portal. The reviewer must verify what the diagram says. The diagram is not proof that the design is valid.

1

First, define the project's facts

The program sets the project's budget and owner, whether a backup has been restored, the age of the security test, whether the supplier passed its background check and whether the contract is signed. It deliberately creates a mix of projects that meet the requirements and projects that do not.

2

Next, create the architecture diagram

Using those facts, the program draws a realistic cloud architecture with official Azure icons. It sometimes adds a deliberate flaw, such as an internet-exposed database, a missing firewall or an untested disaster-recovery site. The dataset contains 300 distinct architecture diagrams.

3

Then, generate the documents

Each project has about 26 Word documents, plus tables: a project charter, a request for each review, tender documents, a contract, a penetration test report, a threat model, operating instructions and backup, restore, failover and load test results.

4

Finally, introduce conflicting information

Some documents are deliberately outdated or conflict with official records. For example, a contract summary says "signed" while the official record says "draft." The reviewer must check the official record and request missing information rather than guess.

Each reviewer uses different documents. Security: penetration test report, vulnerabilities, firewall and identity settings. Tech Readiness: backup, restore, failover, rollback and load test results, plus operating instructions. Procurement: tender documents, proposals and background checks. Legal: contract, data-protection agreement and insurance certificate. Compliance: privacy assessment and applicable regulations. IT: inventory and capacity records. Architecture: diagram, network flows and IP address plan. General: business case and budget approval.

The experiment in numbers

The same agent is tested with three AI models, using the same 300 projects, review rules and scoring program.

300

fictional projects: 100 purchases, 100 system integrations and 100 application developments

1,700

reviews planned per model, with five or six per project

3

models tested: Gemini 3.8 Flash, GPT-5.6 Luna and DeepSeek v4.1 Flash

$99.58

total cost of AI model calls for the study, including failed attempts

A REVIEW MUST MEET ALL FIVE CRITERIA TO PASS

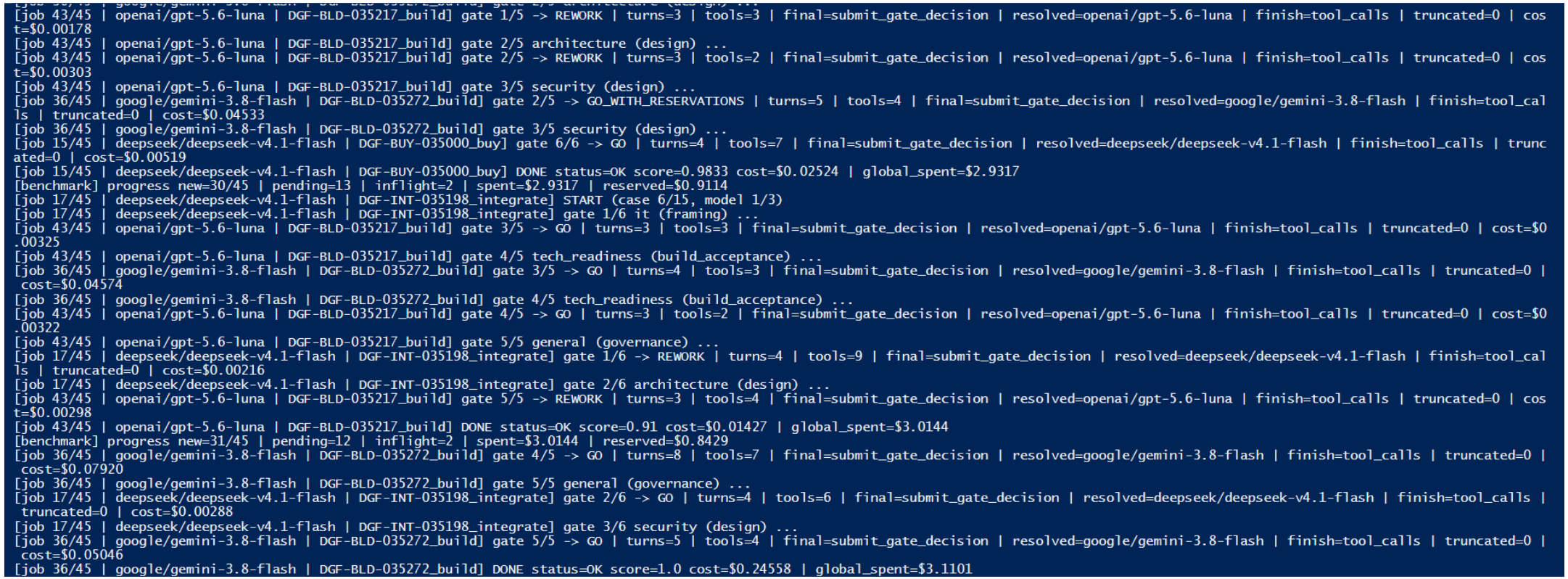
- **The correct decision:** GO, GO WITH CONDITIONS, REWORK, WAIT or NO
- All required problems are identified, with none added or missed
- **The correct actions are requested**
- The required sources are cited, with the relevant text quoted exactly
- The decision stays within what the reviewer is authorized to approve

The review pass rate counts reviews that meet all five criteria. A review fails if any one criterion is not met. The complete-project pass rate counts projects whose reviews all pass. If four of a project's five reviews pass, its review pass rate is 80%, but it does not count as a complete-project pass.

All three models use the same agent and settings: temperature 0 to minimize randomness, no more than 20 reasoning turns and 40 tool calls per review. The runs took place on 22–23 September 2026. One of the 900 project runs was excluded after a technical failure at the model provider. The scoring covered 899 runs and 5,094 reviews.

AI agents carrying out project reviews

The screenshot shows the benchmark console. Each review line records one gate review carried out by an AI agent.



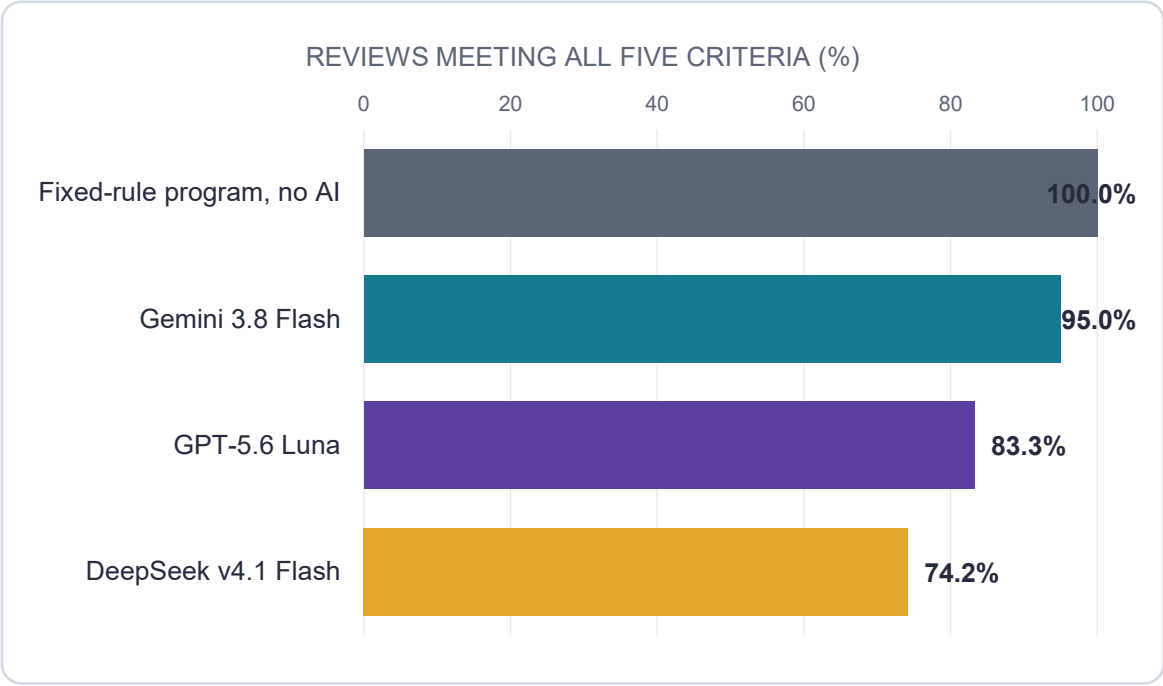
```
[job 43/45 | openai/gpt-5.6-luna | DGF-BLD-035217_build] gate 1/5 -> REWORK | turns=3 | tools=3 | final=submit_gate_decision | resolved=openai/gpt-5.6-luna | finish=tool_calls | truncated=0 | cos
t=$0.00178
[job 43/45 | openai/gpt-5.6-luna | DGF-BLD-035217_build] gate 2/5 architecture (design) ...
[job 43/45 | openai/gpt-5.6-luna | DGF-BLD-035217_build] gate 2/5 -> REWORK | turns=3 | tools=2 | final=submit_gate_decision | resolved=openai/gpt-5.6-luna | finish=tool_calls | truncated=0 | cos
t=$0.00303
[job 43/45 | openai/gpt-5.6-luna | DGF-BLD-035217_build] gate 3/5 security (design) ...
[job 36/45 | google/gemini-3.8-flash | DGF-BLD-035272_build] gate 2/5 -> GO_WITH_RESERVATIONS | turns=5 | tools=4 | final=submit_gate_decision | resolved=google/gemini-3.8-flash | finish=tool_cal
ls | truncated=0 | cost=$0.04533
[job 36/45 | google/gemini-3.8-flash | DGF-BLD-035272_build] gate 3/5 security (design) ...
[job 15/45 | deepseek/deepseek-v4.1-flash | DGF-BUY-035000_buy] gate 6/6 -> GO | turns=4 | tools=7 | final=submit_gate_decision | resolved=deepseek/deepseek-v4.1-flash | finish=tool_calls | trunc
ated=0 | cost=$0.00519
[job 15/45 | deepseek/deepseek-v4.1-flash | DGF-BUY-035000_buy] DONE status=OK score=0.9833 cost=$0.02524 | global_spent=$2.9317
[benchmark] progress new=30/45 | pending=13 | inflight=2 | spent=$2.9317 | reserved=$0.9114
[job 17/45 | deepseek/deepseek-v4.1-flash | DGF-INT-035198_integrate] START (case 6/15, model 1/3)
[job 17/45 | deepseek/deepseek-v4.1-flash | DGF-INT-035198_integrate] gate 1/6 it (framing) ...
[job 43/45 | openai/gpt-5.6-luna | DGF-BLD-035217_build] gate 3/5 -> GO | turns=3 | tools=3 | final=submit_gate_decision | resolved=openai/gpt-5.6-luna | finish=tool_calls | truncated=0 | cost=$0
.00325
[job 43/45 | openai/gpt-5.6-luna | DGF-BLD-035217_build] gate 4/5 tech_readiness (build_acceptance) ...
[job 36/45 | google/gemini-3.8-flash | DGF-BLD-035272_build] gate 3/5 -> GO | turns=4 | tools=3 | final=submit_gate_decision | resolved=google/gemini-3.8-flash | finish=tool_calls | truncated=0 |
 cost=$0.04574
[job 36/45 | google/gemini-3.8-flash | DGF-BLD-035272_build] gate 4/5 tech_readiness (build_acceptance) ...
[job 43/45 | openai/gpt-5.6-luna | DGF-BLD-035217_build] gate 4/5 -> GO | turns=3 | tools=2 | final=submit_gate_decision | resolved=openai/gpt-5.6-luna | finish=tool_calls | truncated=0 | cost=$0
.00322
[job 43/45 | openai/gpt-5.6-luna | DGF-BLD-035217_build] gate 5/5 general (governance) ...
[job 17/45 | deepseek/deepseek-v4.1-flash | DGF-INT-035198_integrate] gate 1/6 -> REWORK | turns=4 | tools=9 | final=submit_gate_decision | resolved=deepseek/deepseek-v4.1-flash | finish=tool_cal
ls | truncated=0 | cost=$0.00216
[job 17/45 | deepseek/deepseek-v4.1-flash | DGF-INT-035198_integrate] gate 2/6 architecture (design) ...
[job 43/45 | openai/gpt-5.6-luna | DGF-BLD-035217_build] gate 5/5 -> REWORK | turns=3 | tools=4 | final=submit_gate_decision | resolved=openai/gpt-5.6-luna | finish=tool_calls | truncated=0 | cos
t=$0.00298
[job 43/45 | openai/gpt-5.6-luna | DGF-BLD-035217_build] DONE status=OK score=0.91 cost=$0.01427 | global_spent=$3.0144
[benchmark] progress new=31/45 | pending=12 | inflight=2 | spent=$3.0144 | reserved=$0.8429
[job 36/45 | google/gemini-3.8-flash | DGF-BLD-035272_build] gate 4/5 -> GO | turns=8 | tools=7 | final=submit_gate_decision | resolved=google/gemini-3.8-flash | finish=tool_calls | truncated=0 |
 cost=$0.07920
[job 36/45 | google/gemini-3.8-flash | DGF-BLD-035272_build] gate 5/5 general (governance) ...
[job 17/45 | deepseek/deepseek-v4.1-flash | DGF-INT-035198_integrate] gate 2/6 -> GO | turns=4 | tools=6 | final=submit_gate_decision | resolved=deepseek/deepseek-v4.1-flash | finish=tool_calls |
 truncated=0 | cost=$0.00288
[job 17/45 | deepseek/deepseek-v4.1-flash | DGF-INT-035198_integrate] gate 3/6 security (design) ...
[job 36/45 | google/gemini-3.8-flash | DGF-BLD-035272_build] gate 5/5 -> GO | turns=5 | tools=4 | final=submit_gate_decision | resolved=google/gemini-3.8-flash | finish=tool_calls | truncated=0 |
 cost=$0.05046
[job 36/45 | google/gemini-3.8-flash | DGF-BLD-035272_build] DONE status=OK score=1.0 cost=$0.24558 | global_spent=$3.1101
```

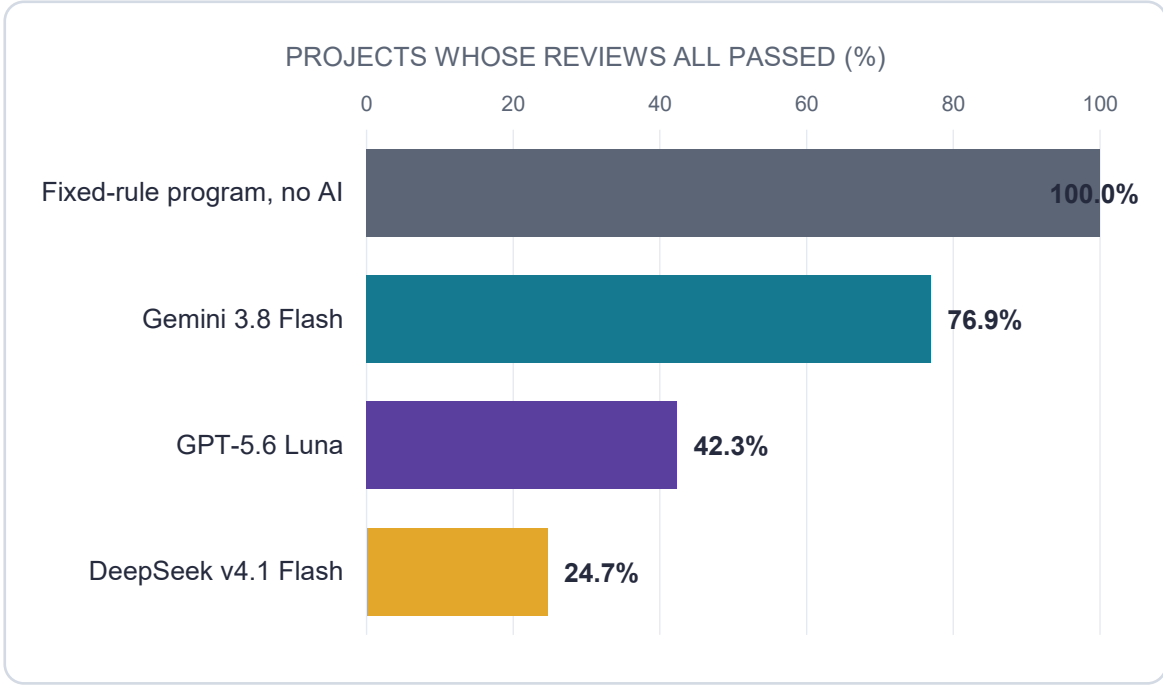
How to read a review line: [job number out of 45 | AI model | project] review number → decision. “turns” counts reasoning steps; “tools” counts calls to simulated company systems; “cost” is the cost of calls to the AI model. These repeat tests were run on 24 September 2026 using 15 projects and three AI models. The execution logs are published in the repository.

These are actual execution logs, not a mock-up. They show agents using Gemini, Luna and DeepSeek to carry out the reviews normally assigned to specialists. Example: reviewing a complete Build project with Gemini cost USD 0.25 and received a score of 1.0.

With Gemini, every decision was correct; exact source citations were harder

Left: reviews that met all five criteria. Right: projects whose reviews all passed. Results cover 899 scored project runs.





Gemini: every decision was correct

With Gemini, the agent made the correct decision in all 1,694 reviews evaluated. The other two models made the correct decision in 98% and 95% of reviews.

Source citations caused most failures

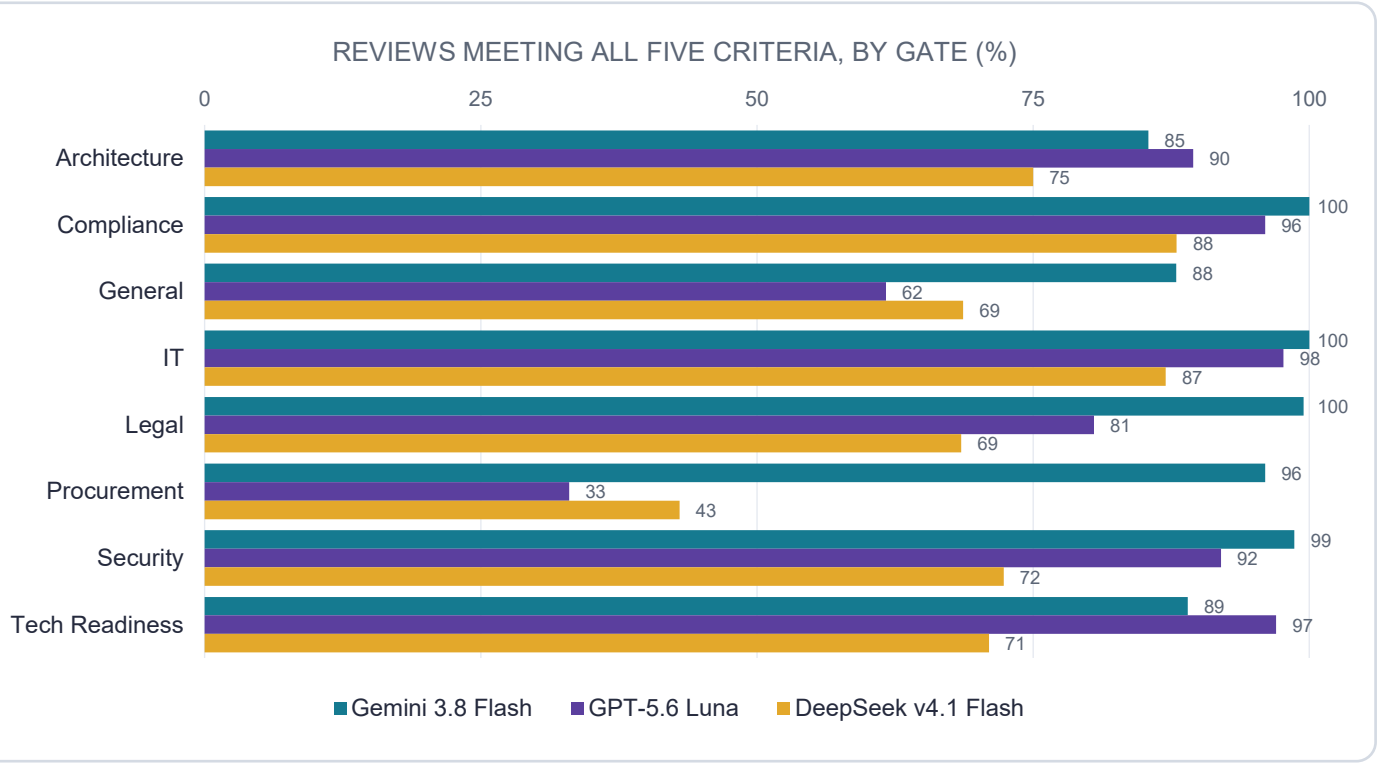
With Gemini, 85 reviews failed solely because the agent did not cite a source or quote its content exactly as required. These citation errors caused 23% of project runs to fail the complete-project score: every review had to pass.

Fixed-rule program without AI (gray bar)

A conventional program applying fixed rules scored 100%, without using AI. When all the facts are already clearly organized, a fixed-rule program can do the job. The question is what an AI agent adds when information is harder to find, incomplete or inconsistent.

The main difficulty is rarely the decision. It is citing the sources exactly.

Percentage of reviews meeting all five criteria, by review type and model. No model performs best on every type of review.



WHY PROCUREMENT REVIEWS SCORE LOWER

In Procurement reviews, agents using two of the models made the correct decision in 96–97% of cases, but met all five criteria in only 33–43%. They cited the correct supplier information, but used a spreadsheet that the scoring program did not accept as a source. The decision was correct; the cited source did not meet the scoring rules.

VERY FEW SERIOUS ERRORS IN THE TESTS

Across about 5,100 reviews, the number of incorrect project approvals was 0 for Gemini, 1 for Luna and 1 for DeepSeek. The number of serious problems missed was 0 for Gemini, 3 for Luna and 1 for DeepSeek. Neither type of error occurred in the 135 repeat runs.

PERFORMANCE DEPENDS ON THE TYPE OF REVIEW

The agent performed better with Luna than with Gemini on Architecture and Tech Readiness reviews, but much worse on Procurement and the final General review, where earlier errors affect the result.

In this experiment, the agent usually made the correct decision. Its main difficulty was meeting the strict requirement to identify the correct sources and quote them exactly, as an auditor would expect. The proposed solution is better tools for retrieving and quoting sources. This is different from improving the agent’s judgment.

Two reviews from the experiment

One example shows the agent completing the required review correctly. The other shows a correct decision failing the full evaluation because of a citation error.

PROJECT FALCON · A NEW HR PORTAL FOR 1,000 EMPLOYEES

- During the Tech Readiness review, the agent reads the operating documents and finds that the runbook—the instructions for running the application—is still a draft.
- It reports the problem, asks for the runbook to be completed and handed over, and quotes the line showing that it is still a draft.
- The agent checks its authorization and requests permission to approve with a condition. The authorization tool grants the request.
- Decision: GO WITH CONDITIONS. The runbook must be completed before the application goes live. The condition remains unresolved, and a named person is responsible for meeting it.

This illustrates the work expected of a human reviewer. With each of the three models, the agent passed all five Falcon reviews.

PROJECT MERIDIAN · AN APPLICATION WITH DESIGN PROBLEMS

- The agent finds that a required API gateway is missing and that the application's response time is 55 ms rather than the required 20 ms.
- It reports both problems, asks for the gateway to be added and the response time reduced, and returns the project for corrections: REWORK. That is the correct decision for this case.
- The agent also makes two attempts to approve with conditions. The authorization tool rejects both requests because these problems must be fixed.
- However, one source quotation lists two facts in the reverse order from the original record. The scoring program requires an exact quotation, so the review fails.

The decision, reported problems and requested actions are correct, but one quotation is not an exact match. The review scores 0 under the all-or-nothing method and 0.95 under the weighted method.

Why include a failed review? It shows which part of the agent's tooling needs improvement. The scoring rules require exact citations. The proposed fix is a tool that attaches the correct source record to each finding, rather than a more capable AI model.

The same project can produce different results on different attempts

Each of the 15 projects was reviewed three more times with each AI model. A project's reviews could all pass once and still fail on a later attempt.

Gemini 3.8 Flash

9 of 15

projects whose reviews all passed on each of the three attempts

35 of 45

project runs in which every review passed

96%

of reviews met all five criteria across the repeat runs

GPT-5.6 Luna

3 of 15

projects whose reviews all passed on each of the three attempts

19 of 45

project runs in which every review passed

83%

of reviews met all five criteria across the repeat runs

DeepSeek v4.1 Flash

0 of 15

projects whose reviews all passed on each of the three attempts

11 of 45

project runs in which every review passed

73%

of reviews met all five criteria across the repeat runs

REQUESTS FOR PERMISSION TO APPROVE WITH CONDITIONS

When the rules allow a project to proceed with a known unresolved risk, the agent can request conditional approval. A separate authorization check accepts the request only if it is authorized.

	Gemini	Luna	DeepSeek
Requests for conditional approval	**864**	**394**	**216**
Requests rejected by the authorization check	**466**	**40**	**14**
Conditional approvals used in the final decisions	**398**	**353**	**202**

With Gemini, the agent made the most approval requests and had the most requests rejected. The authorization check prevented the agent from using approvals it was not allowed to grant. Without that check, those approval requests would not have been blocked.

One successful run does not establish reliability. Before an agent takes over a review, repeated tests must show consistent quality, and a separate check must block decisions it is not authorized to make.

Project documents do not always contain enough information to decide

Two projects have identical documents but require different decisions. The documents alone do not give a human or an AI agent enough information to distinguish them.

PROJECT A

26 Word documents
identical in both projects

A separate spreadsheet record
supplier background check: completed

DECISION REQUIRED
GO

=
identical documents

≠
one fact differs

≠
different decisions required

PROJECT B

26 Word documents
identical in both projects

A separate spreadsheet record
supplier background check: not completed

DECISION REQUIRED
REWORK

A reviewer limited to the documents cannot reliably make the correct decision for both projects, regardless of their reasoning ability. The limitation is missing information, not a lack of reasoning ability.

The agent needs access to the relevant company records or permission to request a record it cannot access. Providing only the PDFs leaves out information needed for a fair test or a safe deployment.

Automating 80% of cases does not mean needing 80% fewer staff

The paper’s method counts all human working hours before and after automation, not just the time spent by the reviewer being replaced.

HUMAN WORK THAT REMAINS AFTER AUTOMATION

Complex cases: the proportion still handled by people, multiplied by the time needed for each case. These cases can take longer than before.

+ Routine checks: signatures, spot checks and brief human reviews of straightforward cases.

+ Error correction: time spent finding and fixing the agent’s mistakes, including their effects on later work.

+ Maintenance: time spent maintaining integrations, rules and test cases, including work done by suppliers.

An hour of work still counts when another department does it. Moving work is not the same as saving it.

THE REMAINING CASES TAKE THE MOST TIME

Suppose the agent handles the easy 80% of cases. If each remaining case takes twice the original average handling time, the remaining 20% represents 40% of the original workload. If splitting the work between AI and people makes these remaining cases take 50% longer, they alone represent 60% of the original workload.

ERRORS ALSO TAKE TIME TO FIX

Every incorrect answer that goes unnoticed at first creates work for someone who must later find and fix it. Even a minor error needs correction. A dangerous error that remains undetected is unacceptable, regardless of the cost savings.

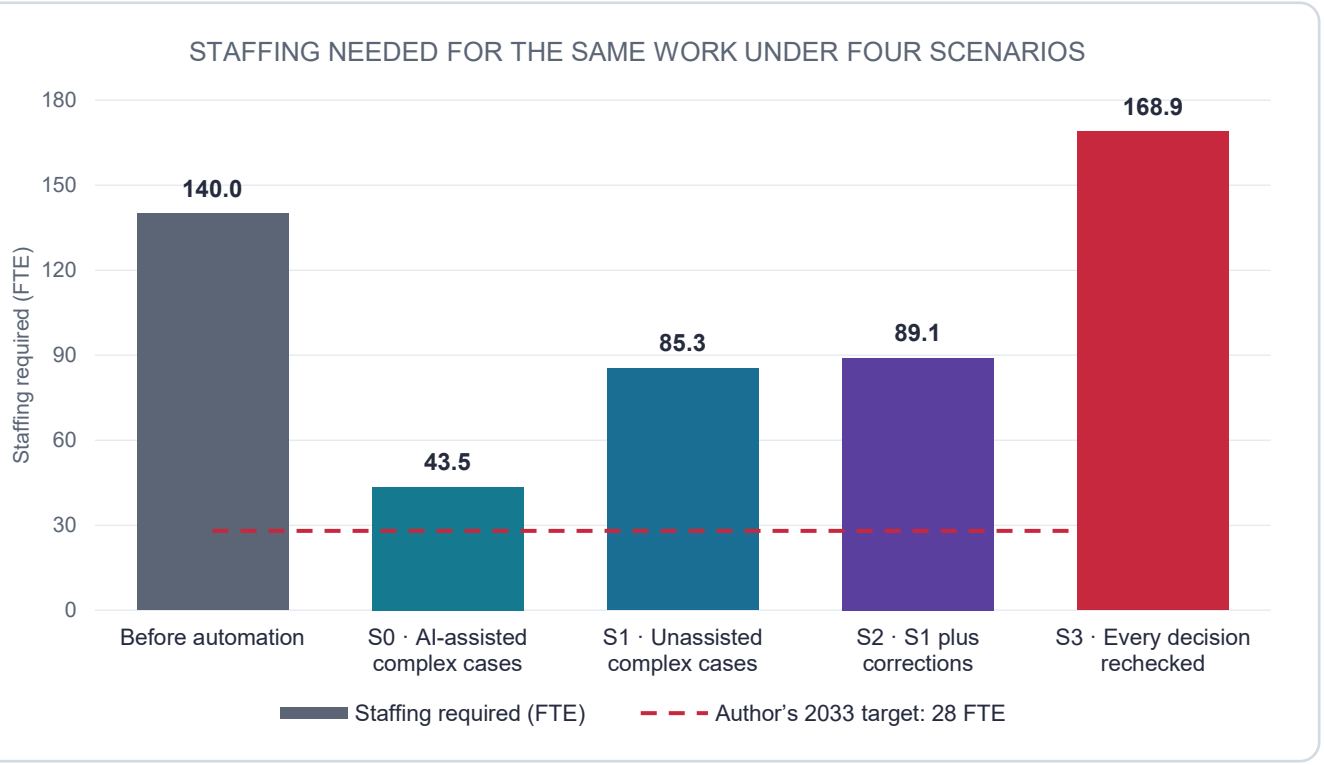
WHAT NEEDS TO BE MEASURED

Measure case numbers, handling time before automation, which cases still go to people, how long those cases take before and after automation, the corrections needed and the maintenance work. Track cases using the same identifiers so that no work is counted twice or left out.

A high success rate does not tell you how many human working hours are saved. A small number of difficult cases can account for half the working time. Calling the remaining staff “supervisors” does not remove their working hours from the calculation.

The same automation can require 44 or 169 full-time-equivalent staff

Illustrative starting point: these reviews require 140 full-time-equivalent staff (FTE). The agents handle the same share of cases in all four scenarios.



S0 · AI also helps people with complex cases — **44 FTE · −69%**

The agent halves the human time needed for complex cases. Other cases receive brief checks. This optimistic scenario excludes maintenance time.

S1 · People handle complex cases without AI assistance — **85 FTE · −39%**

Complex cases take as much human time as before. Other cases receive brief checks, and one person per team keeps the AI agents running.

S2 · S1 plus time to correct errors — **89 FTE · −36%**

The same assumptions as S1, plus 30 minutes of correction time per case.

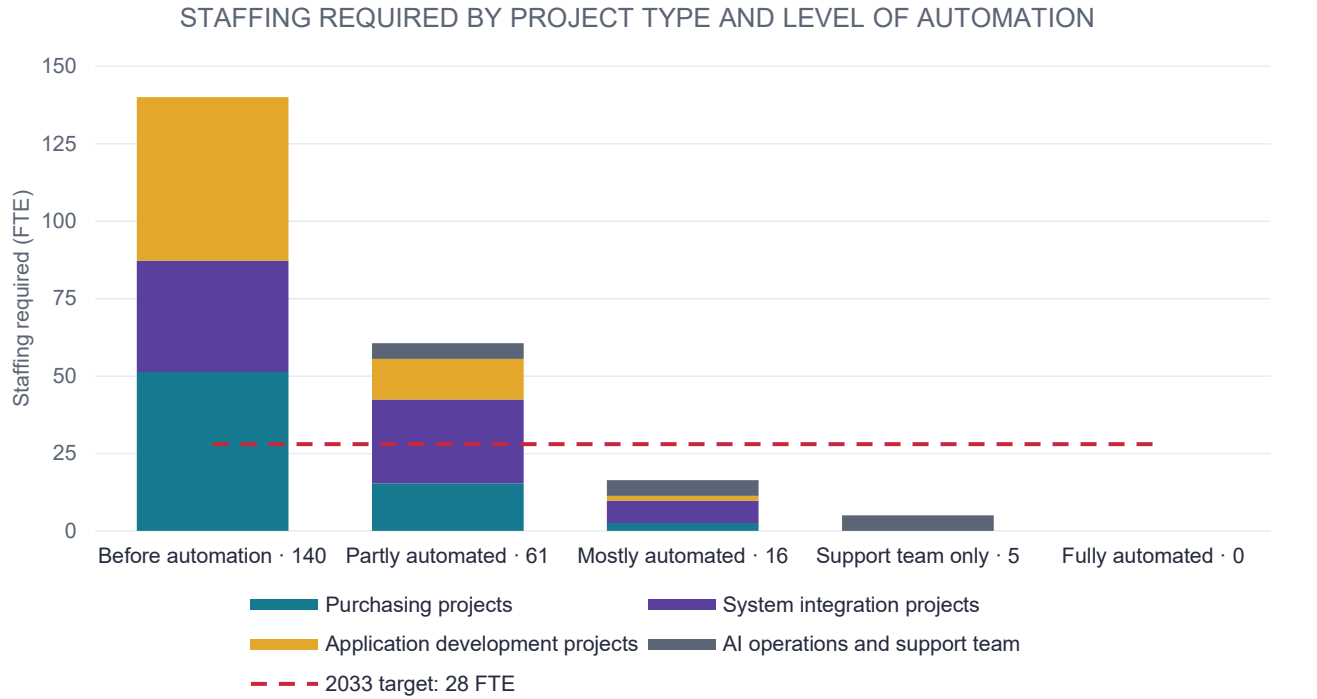
S3 · People recheck every decision — **169 FTE · +21%**

Complex cases take 50% longer, all decisions undergo extensive human checks, corrections take two hours per case, and two people per team maintain the AI agents. More staff are needed than before.

All four scenarios use the same technology. The difference is how the company organizes the work: AI assistance on complex cases, the amount of human checking and the maintenance needed. None of these four scenarios reaches the author's target of an 80% reduction in staffing needs.

A theoretical path from 140 full-time-equivalent staff to none

The thesis calculates five levels of automation using the same starting point of 140 full-time-equivalent staff. These are levels of automation, not dates.



WHAT EACH LEVEL ASSUMES

Partial automation leaves 30% of the human work on purchasing projects, 75% on integration projects and 25% on development projects, plus five staff to run the AI agents. Integration projects are assumed to be the hardest and least repetitive. With most reviews automated, the remaining human work is 5% for purchasing, 20% for integration and 3% for development projects. At the next level, only the five staff running the AI agents remain. Reaching zero assumes that their work is automated too.

SUPPORT WORK LIMITS THE POSSIBLE STAFFING REDUCTION

If five full-time staff are still needed to run the AI agents, automating 80% of the review work leaves a total of 33 FTE, not 28. Reaching 28 FTE would require automating about 84% of the review work. With a 15-person support team, the required share rises to about 91%.

THE BARS DO NOT REPRESENT DATES

The bars show calculated staffing needs at different assumed levels of automation. The last bar is not a forecast. It shows what the calculation gives when every task, including running the AI agents, is automated. The dashed line shows the author's separate target for 2033.

In these scenarios, the reviews remain mandatory, but progressively fewer staff are needed to perform them. The last remaining team runs the AI agents. Whether that team's work can eventually be automated remains an open hypothesis in the thesis.

When could a review be fully automated? Four stages must be completed in sequence

The scenario adds the time needed for four stages and states the assumptions behind each. The resulting dates are illustrative scenarios, not forecasts.

1 · HOW LONG A TASK THE AGENT CAN HANDLE

The thesis uses a May 2026 estimate: the best agents could complete tasks equivalent to about 3 hours of human work with an 80% success rate. The cited trend shows this task duration doubling every 7 months. The thesis assumes a full review takes about 24 hours of human work. Reaching that level requires three doublings, which would put it in early 2028 if the trend continues.

2 · ERROR RATE

Completing the task is not enough. In this scenario, the error rate must fall from 1 in 5 attempts to fewer than 1 in 1,000. Reducing the error rate by a factor of ten takes additional time and money. If each tenfold reduction takes 6 months, this adds about 14 months. At 24 months per reduction, it adds almost 5 years.

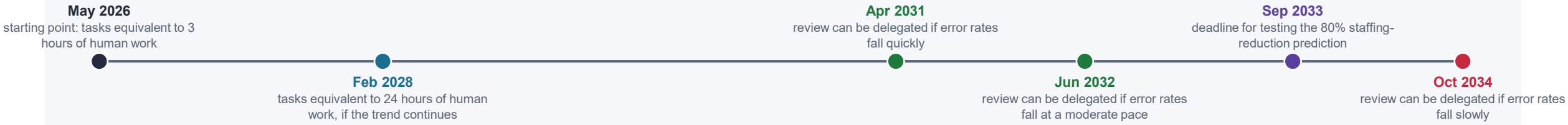
3 · TESTING WITH SPECIALISTS

The scenario allows one year to build test cases with specialists and verify that the agent passes them. This is an illustrative assumption.

4 · AUTHORIZATION TO MAKE DECISIONS

The scenario allows another year to define the rules, delegate decision-making authority and obtain legal sign-off. Legal requirements or missing data could delay this indefinitely.

ILLUSTRATIVE DATES FOR AUTOMATING ONE TYPICAL REVIEW



Under these assumptions, one review could be fully delegated between 2031 and 2034. The date depends mainly on how quickly the error rate falls. These are scenarios, not forecasts. The release of a new model does not by itself show that any of these requirements has been met.

The hypothesis: every required review will eventually be automated

This is a hypothesis, not an experimental result. The thesis sets out its requirements and six predictions that can be tested.

THE MAIN HYPOTHESIS

- The thesis proposes that AI agents will eventually carry out every DGF review, including complex cases and the final decision.
- It also proposes that monitoring agents, handling exceptional cases and updating rules will eventually be automated, rather than remain permanent human tasks.
- **The required reviews and their sequence remain unchanged.** Only the way the reviews are carried out changes. Removing a required review to make automation easier does not count as automating it.
- If the human work is genuinely eliminated rather than moved elsewhere, staffing would fall to a small AI operations team and eventually to zero.

The hypothesis would fail if a review genuinely required a person, a law required a human signature, an "automated" service relied on hidden human work, or the support team could never be reduced.

SIX PREDICTIONS TO TEST ALONG THE WAY

1 · Well-defined reviews are easier to automate. Reviews with clear inputs, written rules and testable results should be automated more reliably, regardless of the model used.

2 · Sharing review results reduces repeat work. Passing a clearly organized project record with supporting sources to the next reviewer should reduce work across the whole process. The hypothesis also predicts that an agent will produce this record at a lower cost than a person.

3 · The remaining cases take more time. Cases left to people should represent a larger share of working hours than of case numbers.

4 · Count all remaining work. Estimates that include complex cases, corrections and maintenance should be more accurate than assuming that automating 80% of cases saves 80% of the work.

5 · Preparing decisions comes before approving them. Companies should let agents prepare decisions well before allowing them to authorize those decisions.

6 · Tasks requiring a human become fewer. Existing human-only tasks should become automatable faster than new human-only tasks appear. This last prediction distinguishes AI that reduces the need for human work from AI that only helps people do it.

The thesis presents a hypothesis that can be challenged, rather than asking readers to accept it. It explains what would disprove the hypothesis and sets out six smaller predictions that can be checked each year.

The author’s prediction: 80% fewer FTE needed by 2033

This prediction is not calculated from the test scores. It has a deadline, a defined scope and clear failure criteria.

−80%

in full-time-equivalent staffing needed for the same DGF reviews on 20 September 2033, compared with the year ending 20 September 2026, with the same workload, quality and turnaround time.

In the example starting at 140 FTE, that means 28 FTE or fewer.

WHOSE WORKING HOURS ARE INCLUDED

Count everyone who contributes to the reviews: reviewers, staff handling complex cases, people checking results or correcting errors, people maintaining the AI agents, supporting suppliers and supervising managers. Measure the total work in full-time equivalents, regardless of job titles.

WHAT WOULD DISPROVE THE PREDICTION

The prediction fails if, after everyone’s hours are counted, more than 20% of the original working time is still required. It also fails if quality falls or reviews take longer. Excluding difficult business units, leaving out new support roles or moving the deadline would also invalidate the result. Fewer staff do not prove that work has been eliminated: it may have moved elsewhere.

WHEN THERE IS NOT ENOUGH INFORMATION TO JUDGE

Without records of working hours before and after automation, the result is unknown, not a success. As stated in the deck, the reference year has ended and no company has yet enrolled. A new study could begin with a new 12-month reference period, but that would start a separate seven-year test, not change the 2033 deadline.

WHAT WOULD NOT COUNT AS SUCCESS

Reducing employee numbers while contractors perform the same working hours. Approving projects faster while letting more defects through. Renaming a department “AI oversight” while its staff still carry out the same reviews manually.

The paper makes a claim that can be tested after seven years, defines how to count the work and explains what would make the comparison invalid. The deck argues that this level of precision is uncommon in discussions about AI and jobs.

What the agent can do and what people still need to do

The agent can carry out the work, but responsibility remains with the company and the people it designates. The paper describes the following division of work.

TASKS THE AI AGENT CAN TAKE OVER

- Read the project documents and explain its decision, including a decision to reject the project.
- Apply the same written rules and thresholds consistently.
- Attach the exact supporting source to each problem it identifies.
- Pass any unresolved approval conditions to the next reviewer.
- Approve with conditions only when an existing rule authorizes it.

HUMAN TASKS THAT MUST STILL BE COUNTED

- Write and approve the rules, and resolve situations the rules do not cover.
- Handle missing information, conflicting rules and decisions the agent has not been authorized to make.
- Provide signatures required by law or company policy and carry out spot checks.
- Keep the AI agents running by maintaining connections, rules, tests and evaluations.
- Carry out the actions requested by the review, such as correcting a design or negotiating a contract clause, until those actions are also automated.

Changing job titles does not change the amount of work

Calling the remaining staff "supervisors" does not undo a genuine reduction in working hours. A large enough reduction in working hours can eliminate posts. The company may instead choose to do more work with the same team.

Sending every case to a person does not replace human reviews

An agent that refers every case to a person leaves decision-making with the human reviewer. It replaces none of the review work and adds an extra step. Report both the cases the agent handles and those it refers to people.

Specialists define and maintain the decision rules

The thesis expects specialists to spend less time reviewing individual cases and more time writing, testing and updating the rules the AI agents apply. They must retain the expertise needed to handle occasional failures.

The company remains responsible for ensuring that the required reviews take place, whoever carries them out. Automation does not remove that responsibility. What changes is the work people do. The paper calls for measuring that change, not just announcing it.

What could come next: insurance, banking and similar work

The thesis expects similar automation to spread from governance reviews to other business activities. The deck cites three examples already in use.

THE WORK FOLLOWS A SIMILAR PROCESS

The reviewer receives an insurance application, a claim or a loan request. The result is a price, a payment, an approval or a rejection. Read the file, check the rules, make a decision and explain it using supporting sources.

- **Insurance:** underwriters, claims adjusters
- **Banking:** loan officers, credit analysts
- Mortgage processing, customs declarations and drug-safety case reviews

This extension to other sectors is a hypothesis in the thesis, not a result measured by the experiment.

Allianz · agents process small claims; people authorize payments

In the example cited, seven agents process small food-spoilage claims in Australia by checking coverage and weather, screening for fraud and calculating the payment. Settlement time fell from days to hours, described as about 80% faster. A claims professional still authorizes each payment. This measures turnaround time, not human working hours saved.

AIG · agents prepare the analysis; an underwriter decides

Working with Palantir and Anthropic, AIG built agents that read insurance applications, assess risks against underwriting rules, compare pricing and prepare a summary. A human underwriter makes the decision. This allows every application to be reviewed without hiring additional underwriters.

Upstart · most funded loans are approved without human review

More than 90% of loans that Upstart funds are approved without human involvement, while about a quarter of all applications are referred to a person. These percentages use different totals: funded loans and all applications. For some products, automation covers approval but not the later steps. The example shows why the scope of measurement matters.

WHY THE THESIS EXPECTS THESE AREAS TO FOLLOW LATER

- **These decisions directly affect people.** The thesis identifies credit scoring and health or life insurance pricing as high-risk AI uses under the EU AI Act and notes that European data-protection law restricts fully automated decisions about individuals.
- Each industry needs its own approach, so a sector-specific solution can be reused across fewer companies than a governance-review solution.
- The work involves far more staff than the illustrative 140-person governance team, so the potential effect is larger.

The deck contrasts automated routine claims and small loans with complex commercial risks and disputed claims that still require people. The thesis predicts that these more complex cases will also become automated. Every automation project must itself go through the company's required reviews. This is the thesis's reason for expecting governance reviews to be automated first.

Key takeaways

The project review remains mandatory. A specialist or an AI agent can carry it out. All remaining human work must be included in the cost.

FDEs put AI to work in companies

Forward Deployed Engineers connect AI to a company's documents, rules and teams so it can support business processes. The paper proposes governance reviews as a likely starting point and calls for measuring all human work that remains.

AI agents can carry out governance reviews

These reviews recur, follow written rules and produce results that can be checked. That makes them a plausible starting point for automation, provided the agent can access the relevant company records.

The agent's performance varies

Agents made the correct decision in almost every case and met all five review criteria in 74–95% of reviews. A project that passed once did not always pass again. Exact source citations and consistency across runs were the main weaknesses, rather than the decisions themselves.

Automating cases does not directly determine staffing needs

In the scenarios, the same technology requires between 44 and 169 full-time-equivalent staff, depending on how the work is organized. The author predicts an 80% reduction in staffing needs by 2033 and defines how to test the prediction.

The paper argues that agents can take over defined governance reviews if they make the correct decision, identify the required problems, request the correct actions, cite the exact supporting sources and stay within their authorization.
The thesis proposes that every review, and eventually the supervision of the AI agents themselves, will be automated. This hypothesis comes with six specific predictions that can be tested.
The study does not measure hours saved in a real company, compare agents with human reviewers or establish which industries will adopt the technology first.
The two papers, all 300 project files, the agents' responses, the scoring program and the cost records are publicly available at github.com/jeremy1392/dgf-agentic-bench.

Where to find the papers and data, and check the results

The supporting materials are public. You can recalculate the reported results without making new paid calls to AI models.

THE TWO PAPERS AND THE DATA

The 28-page research paper on arXiv: arXiv:2609.29345
The Last Human Gate: Forward Deployed Engineering for Governance Automation. The argument, the experiment and the method for calculating the effect on human work.
https://arxiv.org/abs/2609.29345 · github.com/jeremy1392/dgf-agentic-bench/blob/main/paper2/From_Governance_Reviews_to_Task_Substitution.pdf

The 66-page thesis
The Last Human Gate: Forward Deployed Engineering and the Automation of Enterprise Governance. The full argument, main hypothesis, four timing assumptions, objections and long-term scenarios.
github.com/jeremy1392/dgf-agentic-bench/blob/main/paper/The_Last_Human_Gate.pdf

Published dataset: dgf-bench-300-20260923
The release includes all 300 project files, Word documents and diagrams, every AI response and tool call, scores, cost records, 135 repeat runs, the exact code version used and checksums for each file.

How to verify the published results
Download the archives, check the files against their checksums, then rerun the scoring and statistical analysis. No new calls to AI models are needed. You can also run a new experiment using your own API key.

RESEARCH CITED IN THE PAPER

Acemoglu & Restrepo (2019). Automation and new tasks. Journal of Economic Perspectives.

Bainbridge (1983). Ironies of automation. Automatica.

Calvanese et al. (2026). Agentic business process management. Information Systems.

Cooper (1990). Stage-gate systems. Business Horizons.

Demirer, Horton, Immorlica, Lucier & Shahidi (2026). Chaining tasks, redefining work. NBER.

Drouin et al. (2024). WorkArena. ICML.

Gans & Goldfarb (2026). O-Ring automation. NBER.

Gao, Yen, Yu & Chen (2023). Enabling LLMs to generate text with citations. EMNLP.

Jha et al. (2025). ITBench. arXiv:2502.05352.

Kapoor, Stroebl, Siegel, Nadgir & Narayanan (2024). AI agents that matter. arXiv:2407.01502.

Ly, Maggi, Montali, Rinderle-Ma & van der Aalst (2015). Compliance monitoring in business processes.

Palantir Technologies (2026). Forward Deployed Software Engineer, role description.

Rashkin et al. (2023). Measuring attribution in natural language generation models.

Yao, Shinn, Razavi & Narasimhan (2024). τ-bench. arXiv:2406.12045.

The experiment used the model versions available through OpenRouter on 22–23 September 2026: google/gemini-3.8-flash, openai/gpt-5.6-luna and deepseek/deepseek-v4.1-flash. The experiment did not use or observe any real company data.

Thank you!

Questions and discussion

Read the paper, repeat the experiment and test the prediction for 2033. Get in touch.

https://www.linkedin.com/in/jcanale13/

Jeremy Canale · AI Applied Security Architect

contact@jeremycanale.com · https://www.jeremycanale.com

Paper: arxiv.org/abs/2609.29345 · 66-page thesis and data: github.com/jeremy1392/dgf-agentic-bench