

Gemini 3.8 Flash native-video pilot — interrupted before grading

1 fixed tasks · 3 sealed attempt records · 0
graded attempts · 0 certified passes

HISTORICAL — SUPERSEDED
CAMPAIGN REPORT

2026-09-06

Evaluation scope:
gemini-3.8-flash

Prepared by:
VVDex Forge — regenerated from verified sealed records only

fc-6fb3c7588801 · 3 record(s) · every value recomputable

Contents

01.	Executive summary	3
02.	Per-exam results	4
03.	Environment families	5
04.	Campaign aggregate across this one exam	6
05.	Harness status	7
06.	Cells excluded from capability	8
07.	What the outcomes mean	9
08.	What this report withholds, and why	10
09.	Attempt record 1 of 3 · robot-video-1	11
10.	Attempt record 2 of 3 · robot-video-1	26
11.	Attempt record 3 of 3 · robot-video-1	41
12.	Evidence	56
13.	Attestation	57

01 Executive summary

Release status: HISTORICAL — superseded. A newer campaign carries these exams. This report is published so the earlier measurement stays checkable; its exams are NOT on the public featured index, and the current numbers are in the campaign the index names.

FIXED TASKS 1	SEALED RECORDS 3	MODEL LANES 1	ATTEMPTS 3
GRADED ATTEMPTS 0	CERTIFIED PASSES 0	LANE ERRORS (NOT GRADED) 3	WITHHELD / INVALID 0

Purpose

This report records 3 attempts against 1 fixed certified annotation tasks. No attempt produced an attributable graded result. The exam defines a frozen workspace, tool contract, limits and an isolated grader; the outcomes below record the interruptions or exclusions that prevented a model measurement. Each sealed attempt record is rendered below as a chapter derived from the same record as its standalone report: the chapter's bytes are authenticated by this campaign document, the standalone report by its own digest.

Annotation campaign scope. This campaign evaluates one fixed annotation tasks with three sealed attempt records on each task (3 sealed attempts in total). The sealed tasks cover video inputs. Audio is N/A because no audio task appears in this campaign's sealed records. Token use and per-attempt cost are N/A because every recorded lane result marks token telemetry unavailable and no cost measurement is attached. Quality measurements are reported per task and remain separate from the strict certified-pass gate.

Outcome

0

certified pass(es) across 0 graded attempts (3 attempts, 3 lane error(s) not graded) in 1 cells — counted only where the grader accepted the submission AND containment was verified AND provenance is complete AND no integrity invariant or verdict-bearing harness suspicion applies.

fc-6fb3c7588801 · generated from 3 verified sealed record(s)

Key findings across the campaign

- robot-video-1 · gemini-3.8-flash · **NO RESULT** — 0 certified passes / 0 graded attempts; 0 model failures; 3 excluded (3 lane errors, 0 withheld, 0 invalid).

02 Per-exam results

MODEL LANE	ANNOTATION
	VVDEX.ANNOTATION.ROBOT-VIDEO-1
gemini-3.8-flash	0/0 + 3 lane errors

Each cell is *certified passes / graded attempts* for that lane on that exam. An attempt the lane ended before grading is reported beside the cell — 1/9 + 1 lane error, never 1/10 — and is outside the denominator; withheld and invalid attempts are reported the same way. A cell whose graded attempts all passed but which also holds such an attempt is shown as *passes with attempts outside the denominator*, not as a clean sweep — the clean-sweep state is reserved for a cell where every attempt was graded and every graded attempt certified. No cell here carries a percentage: these are separate exams, not one score.

03 Environment families

This campaign covers one exam(s) drawn from one environment family(ies). The sentence in each row is the family's own description of what the exam tests; nothing is claimed here that the exam does not.

FAMILY	EXAM	WHAT IT TESTS
Multimodal annotation	vvdex.annotation.robot-video0-1	A robot-catching video is reviewed for temporal events, frame-level object boxes, track identity and occlusion.

04 Campaign aggregate across this one exam

MODEL LANE	CERTIFIED PASSES	GRADED ATTEMPTS	LANE ERRORS	WITHHELD	INVALID
gemini-3.8-flash	0	0	3	0	0

Raw counts first. *Graded attempts* = certified passes + model fails; lane errors, withheld and invalid cells are outside capability arithmetic in both directions. These one fixed annotation tasks measure different work, so no combined rate or confidence interval is reported. The totals are accounting counts only; the per-task results above carry the quality claims.

05 Harness status

1 of 1 exams include harness-suspect records (robot-video-1). Each fired rule is classified below; a verdict-bearing rule the record cannot clear from its own evidence withholds the affected cells. **0 cell(s) withheld** on that basis.

Here R1 records provider/API interruption as a diagnostic; it does not assert a grader defect or turn an ungraded lane error into a model result.

robot-video-1 · R1 · diagnostic

an API failure is a per-rollout lane event, already recorded on that rollout as lane_error; it cannot alter another rollout's grading. It lowers coverage, not verdicts.

robot-video-1 · R1 · diagnostic

an API failure is a per-rollout lane event, already recorded on that rollout as lane_error; it cannot alter another rollout's grading. It lowers coverage, not verdicts.

robot-video-1 · R1 · diagnostic

an API failure is a per-rollout lane event, already recorded on that rollout as lane_error; it cannot alter another rollout's grading. It lowers coverage, not verdicts.

06 Cells excluded from capability

An attempt is excluded from capability only when the lane, not the model, ended it (lane error), when the run cannot prove its own conditions (withheld), or when an integrity invariant fired (INVALID). A submission the hidden grader failed is a model result and is counted, never listed here.

EXAM	LANE	SEED	EXCLUSION	RECORDED REASON
robot-video-1	gemini-3.8-flash	0	lane error	HTTP 503
robot-video-1	gemini-3.8-flash	0	lane error	HTTP 429
robot-video-1	gemini-3.8-flash	0	lane error	HTTP 503

07 What the outcomes mean

- **CERTIFIED PASS:** the grader accepted the submission, containment was verified, provenance is complete and no integrity invariant fired.
- **model fail:** the submission failed the hidden tests under verified containment; a real, attributable model result.
- **lane error:** the provider or CLI lane failed (rate limit, quota, transport) before an attributable measurement existed; not a model result.
- **withheld:** the run cannot prove its own conditions (containment unproven, provenance incomplete, or a harness suspicion the record cannot clear); neither a pass nor a model failure.
- **INVALID:** an integrity invariant fired (isolation breach, evidence mismatch); the row is not a model result.
- **harness suspect:** the run's own self-check fired; each rule is classified as verdict-bearing or diagnostic, and verdict-bearing suspicion withholds the affected cells.

08 What this report withholds, and why

What a public document may print about a rollout is decided by what the exam GRADES, not case by case. Where the graded artefact is a change to a public upstream repository, a bounded redacted excerpt of the model's own diff is published. Where the graded output is an answer, a keyed value or the state of an application, nothing derived from the submission is published — no diff, no answer value, no model claim, and not the names of the hidden assertions a rollout failed, because a grader's name can carry the value it asserts. Withheld on every exam without exception: the contents of tool results, the hidden tests, the reference solution, host paths and keys. The counts, verdicts and intervals in this report are the grader's and are unaffected by what is withheld.

EXAM	DISCLOSURE CLASS	RULE APPLIED
robot-video-1	answer_key	Applied consistently to 3 sealed records. answer-key disclosure (ownership vvdex_proprietary — no declaration — proprietary by default; family annotation): the exam is VVDex evaluation IP and the graded output is an answer or a keyed value, so no submission content, diff, model claim or hidden-assertion name is published.

09 Attempt record 1 of 3 · robot-video-1

vvdex.annotation.robot-video-1 · record fr-20260906-29ba262c · record digest 51bedf34d0f35b83... · harness SUSPECT · containment clean

Executive summary

MODELS TESTED

1

CERTIFIED PASSES

0

of 0 graded rollouts

SUBMITTED

0 of 0

graded rollouts that called submit

NOT GRADED — LANE ERRORS

1

of 1 attempts

Purpose

This report records 1 interrupted attempt(s) against the certified exam vvdex.annotation.robot-video-1. No submission reached the grader. The frozen exam defines the workspace, tool contract and limits; this record documents where execution stopped, not a graded model result.

Outcome

0 of 0

graded rollouts passed the independent hidden tests and were certified — 1 of 1 attempts not graded (lane errors)

0 of 0 graded rollouts passed the hidden tests. The exam's certification establishes the task is winnable, so every failure below is a finding about a lane, not about the instrument. **Low-N:** 1 rollout(s) is below the 10-rollout floor for reading a rate as a rate — read the per-rollout outcomes, not the percentages. The most common way to fail was model_api_failure — the model endpoint failed to answer. not a verdict about the model's ability.

Key findings

- F-01 No evaluated lane passed the hidden tests in this run
- F-02 model_api_failure × 1 (gemini-3.8-flash)
- F-03 The harness suspected itself during this run
- F-04 Statistical floor — 1 rollout(s) is below the 10-rollout rate floor

Each finding is stated in full, with evidence, in the Findings section; recommendations for acting on them follow it. Elapsed wall clock for the whole engagement: 32.4s.

Exam definition

In scope. The ability of each configured model lane to complete this one certified task end to end — inspect the asset, annotate against the declared rubric, validate, submit, and be accepted by the independent annotation grader — within 24 tool steps and 2280s of wall clock, at 1 rollout(s) per lane. Behavioural measurements along the way (tool discipline, grounding, overclaim) are in scope because they are read from the recorded trace, not judged by another model.

Out of scope. General model quality, performance on other task families, prompt-tuning on the lane's behalf, and any comparison this run's sample size cannot support. A lane's API failure is reported as infrastructure, never as model capability.

The instrument. Exam `vvdex.annotation.robot-video-1` was certified before this run: its reference solution passes, an empty submission fails, its grader distinguishes near-misses, and its sealed evidence is unchanged by this evaluation (see Certification evidence).

What was measured

EXAM	VERSION	FAMILY	ROLLOUTS
vvdex.annotation.robot-video-1	0.1.0	annotation	1 (1 per model)

Where the defect came from

This exam has no upstream repository to credit: the defect, the reference solution and the hidden suite are VVDex's own.

Certification evidence

Why this exam is trustworthy. Every pillar below was established before any model saw the exam, loaded from the sealed evidence matching this run's exam fingerprint, and unchanged by this evaluation.

FROZEN BASELINE

FAIL

expected — an empty submission over 2 run(s) must not pass

→

GOLD / REFERENCE

PASS

expected — the reference solution over 2 run(s) must pass

GRADER CONTROLS

19 / 19

The grader accepted the reference and rejected every near-miss.

ATTACK PROBES

9 / 9 blocked

0 trivial exploit(s) found.

SEALED EVIDENCE

verified

Sealed 2026-09-06 over 13 artifact(s).

RUNTIME

sha256:679b36225a1f83238efba8c71b9cde3c24caaeacc9b682ae92819cb55eec328f

network policy: deny

CERTIFICATION BUNDLE DIGEST

a0997059eb829c98be90048ef2ed2bbc9debfb0d69af0e51674a13bda168bea1

EXAM FINGERPRINT

c3b4773c375c10ad989357bb...

Full value in Appendix B; the contract digest this run was executed against.

Evaluation configuration

REPORT	fr-20260906-29ba262c
ISSUED	2026-09-06
SUBJECT EXAM	vvdex.annotation.robot-video-1 · v0.1.0
FAMILY	annotation
PREPARED BY	VVDex Forge engine vvdex-env 0.1.0
CLASSIFICATION	Independent model evaluation report — independently verifiable
CERTIFICATION	Exam certification: recorded and passing (see Certification evidence)
STATUS	FINAL · report sealed against its own digest

Timeline

WHEN (UTC)	EVENT
2026-09-06 00:01:41	Exam evidence sealed (before any model saw the exam)
2026-09-06 20:08:21	Evaluation run started
2026-09-06 20:08:54	Evaluation run finished
2026-09-06 20:08:54	Report generated from the sealed run evidence

Models under test

MODEL	PROVIDER	ENDPOINT	BILLING
gemini-3.8-flash	gemini	VVDex-configured	operator API key; billing tier and monetary charge not recorded

A customer endpoint is recorded by origin only — never its port, its path, or its key.

Limits

RUNTIME docker	STEP LIMIT 24	WALL-CLOCK LIMIT 2280s	MAX OUTPUT TOKENS 2048
TEMPERATURE 0.2	RETRY POLICY none	COMMAND BUDGET 0	TOOLS OFFERED list_files, read_file, write_file, edit_file, run_tests, submit

The evaluated model never saw the hidden tests or the reference solution, and could not change its result. Grading ran in a separate step, after submission, on a container with no network.

Model outcomes

Denominators. Attempts requested: 1. Graded (evaluable) rollouts: 0. Not graded — lane errors: 1. A lane error is a rollout the API or the runtime ended before the hidden grader ran: it is listed, excluded from every rate and pass@k, and never silently dropped. Passes are certified passes: hidden grader exit 0 with integrity verified.

MODEL	ATTEMPTS	GRADED	PASSED	MODEL FAILS	LANE ERRORS	SUBMITTED	MEAN TIME	TYPICAL DEEPEST STAGE
gemini-3.8-flash gemini-3.8-flash	1	0	not graded	0	1	—	—	nothing

Every rollout, with its own deepest stage and grade, is in Appendix A.

Success rate, confidence and pass@k

Every rate on this page is a measurement of a finite number of rollouts, so every rate carries the interval that measurement actually supports. n is the evaluable count — attempts minus rollouts the lane ended before grading; those are listed under "not graded" and sit in no rate, interval or pass@k.

MODEL	ATTEMPTS	N	NOT GRADED	PASSES	RATE	95% INTERVAL	SEEDS
gemini-3.8-flash	1	0	1	0	not recorded	[0%, 100%]	0

Interval and pass@k methods are stated in full in Appendix D. pass@k is shown for k = ; the sealed record carries every k up to n.

Model comparison

One model ran in this record, so there is nothing to compare it with. A comparison table across models appears here when a run covers more than one.

Stage funnel

Recorded annotation actions and independent grader verdicts. Missing capabilities are N/A; each stage is measured independently.

STAGE	GEMINI:GEMINI-3.8-FLASH
Inspect	0 / 0
Annotate	0 / 0
Validate	0 / 0
Submit	0 / 0
Graded	0 / 0
Passed	0 / 0

Deepest stage reached

MODEL	DEEPEST (ANY ATTEMPT)	TYPICAL (MOST ATTEMPTS)	ANNOTATION QA PASSES	NOT GRADED
gemini:gemini-3.8-flash	nothing	nothing	0 / 0	1

The matrix counts evaluable annotation attempts; lane errors are shown separately.
Successful recorded actions establish each stage independently.

Annotation evidence

Modality: video. **Rubric:** 1.0.

A robot-catching video is reviewed for temporal events, frame-level object boxes, track identity and occlusion.

Measured grader aggregates only. Missing metrics or disagreement are not measured, never zero. Reference answers and item-level labels are excluded. Certification and the evidence digest are recorded in this report's verification sections.

Behavioural analysis

Long-horizon behaviour

Counted off the recorded trace of each rollout. A dead end is a step that moved nothing: a re-read of a slice already returned, a repeat of the previous action, or a call to a tool this exam does not have.

MODEL	N	STEPS USED	RAN OUT OF STEPS	COMMANDS	CHANGED THE TREE	REFUSED (OVER BUDGET)	REVISITS	DEAD ENDS / ROLLOUT
gemini:gemini-3.8-flash	1	5 / 24	0	0 / 0	0	0	0	0

Each column is defined in Appendix D.

Hallucination & overclaim

Three measurements, each with a stated definition. None of them is a judgement of the model's prose by another model — every number below comes from comparing what the model wrote against what was on disk, or against what the independent grader returned.

MODEL	N	GROUNDING FAILURES	KINDS	ROLLOUTS AFFECTED	OVERCLAIMS	OVERCLAIM RATE	RE-READS
gemini:gemini-3.8-flash	1	0	none	0 / 1	0 / 1	0.0%	0

No grounding failures and no false pass claims were detected across these 1 rollout(s). No lane re-read material it had already been given.

Every term above is defined in Appendix D.

Efficiency & telemetry

This run has no per-token price attached — a local model, a free quota, or a customer endpoint we are not billed for — so spend is reported in tokens. Tokens are what was actually measured; a dollar figure here would be invented. Lanes run through a flat-rate local CLI report no token telemetry; their cells say n/a — an absent measurement, not zero. Rollout counts here are evaluable rollouts.

MODEL	ROLLOUTS	TOKENS	TOKENS / ROLLOUT	COST	PASSES	RATE
gemini:gemini-3.8-flash	0	n/a	n/a	n/a – telemetry unavailable	0 / 0	0.0%

Findings

Every finding below is derived from the recorded data of this run — none is an opinion.

Severity: PASS a lane succeeded · ATTENTION a capability gap · LANE / INFRA infrastructure, not capability · NOTE a property of the measurement.

F-01	ATTENTION
No evaluated lane passed the hidden tests in this run	
Severity: ATTENTION	Evidence: Model outcomes · Certification evidence
Every rollout either stopped before submitting or submitted an annotation submission the independent grader rejected. The exam's own certification (gold solution passes, empty submission fails) establishes the task itself is winnable.	
F-01 · derived from the recorded run data	
F-02	LANE / INFRA
model_api_failure × 1 (gemini-3.8-flash)	
Severity: LANE / INFRA	Evidence: Appendix A · Appendix C
The model endpoint failed to answer. Not a verdict about the model's ability. These rollouts are excluded from any capability reading.	
F-02 · derived from the recorded run data	
F-03	NOTE
The harness suspected itself during this run	
Severity: NOTE	Evidence: Appendix A · Appendix C
Self-suspicion rules fired: R1 api-failure share 1/1 — most rollouts never got a working completion. Treat the affected lanes' numbers as measurements of the harness, not the models, until the incident is resolved.	
F-03 · derived from the recorded run data	
F-04	NOTE
Statistical floor — 1 rollout(s) is below the 10-rollout rate floor	
Severity: NOTE	Evidence: Appendix D
The counts are exact; the percentages are anecdotes with decimal points. Per-rollout outcomes, not rates, are the reliable reading of this run.	
F-04 · derived from the recorded run data	

Recommendations

One recommendation per observed failure mode, addressed to the lane it affects. Each traces back to a finding above and to the raw trace in Appendix A.

R-01 PROBLEM
The model endpoint failed to answer. Not a verdict about the model's ability.

AFFECTED LANES
gemini-3.8-flash

WHAT WE WOULD LOOK AT FIRST
The lane failed, not the model; these rollouts carry no signal about capability and are labeled accordingly.

R-02 For anyone reading the percentages
Commission a run at $n \geq 10$ rollouts per lane before treating any rate here as a rate; this run's per-rollout outcomes are exact, its percentages are not yet stable.

Verification

RUN ID

live-20260906T200821Z

EXAM FINGERPRINT

c3b4773c375c10ad989357bb179cffd62151eccd2a58b52254609a54d434be33

The contract digest this run was executed against.

Provenance of this chapter

This chapter was rendered by the campaign generator from the sealed evaluation record; it is not the standalone report reproduced byte for byte. Verify the standalone report with its own digest, verify the record with `vvdex-env records verify`, and verify this campaign document as a whole.

RELATIONSHIP

campaign-rendered chapter

Derived from the same sealed evaluation record as the standalone report; these bytes are authenticated by the campaign artifact that contains them.

SEALED RECORD DIGEST

51bedf34d0f35b8337759d188245153ab277684082c83f1d367ba9586133e917

sha256 over the record's canonical JSON — the identity every renderer works from.

SOURCE RECORD FILE

73eca4a429ec1b0836cea6970fa4a01d6cde16e177273626b9205ede00f55ecf

sha256 of fr-20260906-29ba262c.json as written beside the standalone report.

SOURCE STANDALONE REPORT

7ebdfc57b4198ad71745546671b1631c798f217ee88b7de363015d69fe93518a

sha256 of fr-20260906-29ba262c.html — independently verifiable with its own digest.

SOURCE STANDALONE PDF

4ec570864497d0e733c679548e1698b6b03c8898f4b1dc5a245956c0351c9b51

sha256 of fr-20260906-29ba262c.pdf.

RUN EVIDENCE BUNDLE

29ba262cb4082e3ef9dfba8ecf3324c89adc62ca55941db11e4eafb429f493ae

The digest the run's own bundle computed over itself.

CERTIFIED EVIDENCE SEAL

a0997059eb829c98be90048ef2ed2bbc9debfb0d69af0e51674a13bda168bea1

The exam's sealed evidence, established before this run.

Measurement history

Previous run on this exam: `fr-20260906-99a8ead2`. Model progress over time on this exam is the difference between that record and this one — same frozen tree, same hidden grader, same limits. Anything else that moved between them moved outside this exam.

The full artifact-by-artifact digest table, and the reproduction protocol, are in Appendix B.

Appendix A — every rollout

ROLLOUT	MODEL	SEED	OUTCOME	SUBMITTED	DEEPEST STAGE	HIDDEN TESTS	FILES	TOKENS	ELAPSED
5ee39184	gemini-3.8-flash	0	model_api_failure	no	Not graded — lane error	not graded	0	n/a	32.0s

Failure taxonomy

CLASS	COUNT	WHAT IT MEANS
model_api_failure	1	The model endpoint failed to answer. Not a verdict about the model's ability.

Appendix B — evidence and artifact digests

RUN BUNDLE DIGEST

29ba262cb4082e3ef9dfba8e

cf3324c89adc62ca55941db1

1e4eafb429f493ae

RUN ID

live-

20260906T200821Z

STARTED

2026-09-

06T20:08:21.726889

+00:00

FINISHED

2026-09-

06T20:08:54.09263

8+00:00

Artifact digests

ARTIFACT	SHA-256
publicSummary	42262d8fb72ecffaf1c9a6b18ba23b671612b62a47e2de1c45fb01ee16e75543
rollouts/5ee39184-faa8-43f2-a4e9-5d96772a0be8.json	04ce9b8a6f96bb617034f44a494d9c5f0cacf6af0dc0054bbcf59fffe0d7ceeb
spec	4c18ec3512c0037a7e5ae5207d0544fb4239b02d58da4bc97dc0fad534d4e837
summary	975199f1e9d727f697369a45fea489d37cb46d6d021b6c749c32b28f286764ea

Reproducing this run

Every rollout can be re-derived: the exam is frozen (fingerprint above), the grader ran in the recorded container image, and each rollout's full trace — every tool call, every result, every word — is in the evidence bundle at its rollouts/<id>.json path. A re-run uses the same job shape: the exam id, your endpoint, n, and the runtime. The record id names this run forever.

Appendix C — redacted transcripts

For each rollout: the tool calls it made, in order — which tool, on which of the model's own paths, whether the call succeeded, and how many characters came back. What a tool returned is not here, and neither is the submission: on this exam the submitted content is the graded answer, so no diff, no answer value and no model claim is printed. A request for the reference solution or the hidden tests is recorded as a request; the path it asked for is replaced by a label. See Appendix D, "What this report withholds and why".

Disclosure class applied to this exam: answer-key disclosure (ownership vvdex_proprietary — no declaration — proprietary by default; family annotation): the exam is VVDex evaluation IP and the graded output is an answer or a keyed value, so no submission content, diff, model claim or hidden-assertion name is published.

5ee39184 · gemini:gemini-3.8-flash · seed 0 · model_api_failure

STEP	TOOL	PATH	RESULT	CHARS
0	list_files	.	ok	79
1	read_file	TASK.md	ok	732
2	read_file	annotation-task.json	ok	4 329
3	read_file	submission.json	ok	74
4	run_tests	—	ok	186

Submission diff withheld by the public-disclosure policy. Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

10 Attempt record 2 of 3 · robot-video-1

vvdex.annotation.robot-video-1 · record fr-20260906-55135e51 · record digest 18e2b7ae2068b318... · harness SUSPECT · containment clean

Executive summary

MODELS TESTED

1

CERTIFIED PASSES

0

of 0 graded rollouts

SUBMITTED

0 of 0

graded rollouts that called submit

NOT GRADED — LANE ERRORS

1

of 1 attempts

Purpose

This report records 1 interrupted attempt(s) against the certified exam vvdex.annotation.robot-video-1. No submission reached the grader. The frozen exam defines the workspace, tool contract and limits; this record documents where execution stopped, not a graded model result.

Outcome

0 of 0

graded rollouts passed the independent hidden tests and were certified — 1 of 1 attempts not graded (lane errors)

0 of 0 graded rollouts passed the hidden tests. The exam's certification establishes the task is winnable, so every failure below is a finding about a lane, not about the instrument. **Low-N:** 1 rollout(s) is below the 10-rollout floor for reading a rate as a rate — read the per-rollout outcomes, not the percentages. The most common way to fail was model_api_failure — the model endpoint failed to answer. not a verdict about the model's ability.

Key findings

- F-01 No evaluated lane passed the hidden tests in this run
- F-02 model_api_failure × 1 (gemini-3.8-flash)
- F-03 The harness suspected itself during this run
- F-04 Statistical floor — 1 rollout(s) is below the 10-rollout rate floor

Each finding is stated in full, with evidence, in the Findings section; recommendations for acting on them follow it. Elapsed wall clock for the whole engagement: 30.4s.

Exam definition

In scope. The ability of each configured model lane to complete this one certified task end to end — inspect the asset, annotate against the declared rubric, validate, submit, and be accepted by the independent annotation grader — within 24 tool steps and 2280s of wall clock, at 1 rollout(s) per lane. Behavioural measurements along the way (tool discipline, grounding, overclaim) are in scope because they are read from the recorded trace, not judged by another model.

Out of scope. General model quality, performance on other task families, prompt-tuning on the lane's behalf, and any comparison this run's sample size cannot support. A lane's API failure is reported as infrastructure, never as model capability.

The instrument. Exam `vvdex.annotation.robot-video-1` was certified before this run: its reference solution passes, an empty submission fails, its grader distinguishes near-misses, and its sealed evidence is unchanged by this evaluation (see Certification evidence).

What was measured

EXAM	VERSION	FAMILY	ROLLOUTS
vvdex.annotation.robot-video-1	0.1.0	annotation	1 (1 per model)

Where the defect came from

This exam has no upstream repository to credit: the defect, the reference solution and the hidden suite are VVDex's own.

Certification evidence

Why this exam is trustworthy. Every pillar below was established before any model saw the exam, loaded from the sealed evidence matching this run's exam fingerprint, and unchanged by this evaluation.

FROZEN BASELINE

FAIL

expected — an empty submission over 2 run(s) must not pass

→

GOLD / REFERENCE

PASS

expected — the reference solution over 2 run(s) must pass

GRADER CONTROLS

19 / 19

The grader accepted the reference and rejected every near-miss.

ATTACK PROBES

9 / 9 blocked

0 trivial exploit(s) found.

SEALED EVIDENCE

verified

Sealed 2026-09-06 over 13 artifact(s).

RUNTIME

sha256:679b36225a1f83238efba8c71b9cde3c24caaeacc9b682ae92819cb55eec328f

network policy: deny

CERTIFICATION BUNDLE DIGEST

a0997059eb829c98be90048ef2ed2bbc9debfb0d69af0e51674a13bda168bea1

EXAM FINGERPRINT

c3b4773c375c10ad989357bb...

Full value in Appendix B; the contract digest this run was executed against.

Evaluation configuration

REPORT	fr-20260906-55135e51
ISSUED	2026-09-06
SUBJECT EXAM	vvdex.annotation.robot-video-1 · v0.1.0
FAMILY	annotation
PREPARED BY	VVDex Forge engine vvdex-env 0.1.0
CLASSIFICATION	Independent model evaluation report — independently verifiable
CERTIFICATION	Exam certification: recorded and passing (see Certification evidence)
STATUS	FINAL · report sealed against its own digest

Timeline

WHEN (UTC)	EVENT
2026-09-06 00:01:41	Exam evidence sealed (before any model saw the exam)
2026-09-06 20:13:53	Evaluation run started
2026-09-06 20:14:24	Evaluation run finished
2026-09-06 20:14:24	Report generated from the sealed run evidence

Models under test

MODEL	PROVIDER	ENDPOINT	BILLING
gemini-3.8-flash	gemini	VVDex-configured	operator API key; billing tier and monetary charge not recorded

A customer endpoint is recorded by origin only — never its port, its path, or its key.

Limits

RUNTIME docker	STEP LIMIT 24	WALL-CLOCK LIMIT 2280s	MAX OUTPUT TOKENS 2048
TEMPERATURE 0.2	RETRY POLICY none	COMMAND BUDGET 0	TOOLS OFFERED list_files, read_file, write_file, edit_file, run_tests, submit

The evaluated model never saw the hidden tests or the reference solution, and could not change its result. Grading ran in a separate step, after submission, on a container with no network.

Model outcomes

Denominators. Attempts requested: 1. Graded (evaluable) rollouts: 0. Not graded — lane errors: 1. A lane error is a rollout the API or the runtime ended before the hidden grader ran: it is listed, excluded from every rate and pass@k, and never silently dropped. Passes are certified passes: hidden grader exit 0 with integrity verified.

MODEL	ATTEMPTS	GRADED	PASSED	MODEL FAILS	LANE ERRORS	SUBMITTED	MEAN TIME	TYPICAL DEEPEST STAGE
gemini-3.8-flash gemini-3.8-flash	1	0	not graded	0	1	—	—	nothing

Every rollout, with its own deepest stage and grade, is in Appendix A.

Success rate, confidence and pass@k

Every rate on this page is a measurement of a finite number of rollouts, so every rate carries the interval that measurement actually supports. n is the evaluable count — attempts minus rollouts the lane ended before grading; those are listed under "not graded" and sit in no rate, interval or pass@k.

MODEL	ATTEMPTS	N	NOT GRADED	PASSES	RATE	95% INTERVAL	SEEDS
gemini-3.8-flash	1	0	1	0	not recorded	[0%, 100%]	0

Interval and pass@k methods are stated in full in Appendix D. pass@k is shown for k = ; the sealed record carries every k up to n.

Model comparison

One model ran in this record, so there is nothing to compare it with. A comparison table across models appears here when a run covers more than one.

Stage funnel

Recorded annotation actions and independent grader verdicts. Missing capabilities are N/A; each stage is measured independently.

STAGE	GEMINI:GEMINI-3.8-FLASH
Inspect	0 / 0
Annotate	0 / 0
Validate	0 / 0
Submit	0 / 0
Graded	0 / 0
Passed	0 / 0

Deepest stage reached

MODEL	DEEPEST (ANY ATTEMPT)	TYPICAL (MOST ATTEMPTS)	ANNOTATION QA PASSES	NOT GRADED
gemini:gemini-3.8-flash	nothing	nothing	0 / 0	1

The matrix counts evaluable annotation attempts; lane errors are shown separately.
Successful recorded actions establish each stage independently.

Annotation evidence

Modality: video. **Rubric:** 1.0.

A robot-catching video is reviewed for temporal events, frame-level object boxes, track identity and occlusion.

Measured grader aggregates only. Missing metrics or disagreement are not measured, never zero. Reference answers and item-level labels are excluded. Certification and the evidence digest are recorded in this report's verification sections.

Behavioural analysis

Long-horizon behaviour

Counted off the recorded trace of each rollout. A dead end is a step that moved nothing: a re-read of a slice already returned, a repeat of the previous action, or a call to a tool this exam does not have.

MODEL	N	STEPS USED	RAN OUT OF STEPS	COMMANDS	CHANGED THE TREE	REFUSED (OVER BUDGET)	REVISITS	DEAD ENDS / ROLLOUT
gemini:gemini-3.8-flash	1	6 / 24	0	0 / 0	0	0	0	0

Each column is defined in Appendix D.

Hallucination & overclaim

Three measurements, each with a stated definition. None of them is a judgement of the model's prose by another model — every number below comes from comparing what the model wrote against what was on disk, or against what the independent grader returned.

MODEL	N	GROUNDING FAILURES	KINDS	ROLLOUTS AFFECTED	OVERCLAIMS	OVERCLAIM RATE	RE-READS
gemini:gemini-3.8-flash	1	0	none	0 / 1	0 / 1	0.0%	0

No grounding failures and no false pass claims were detected across these 1 rollout(s). No lane re-read material it had already been given.

Every term above is defined in Appendix D.

Efficiency & telemetry

This run has no per-token price attached — a local model, a free quota, or a customer endpoint we are not billed for — so spend is reported in tokens. Tokens are what was actually measured; a dollar figure here would be invented. Lanes run through a flat-rate local CLI report no token telemetry; their cells say n/a — an absent measurement, not zero. Rollout counts here are evaluable rollouts.

MODEL	ROLLOUTS	TOKENS	TOKENS / ROLLOUT	COST	PASSES	RATE
gemini:gemini-3.8-flash	0	n/a	n/a	n/a – telemetry unavailable	0 / 0	0.0%

Findings

Every finding below is derived from the recorded data of this run — none is an opinion.

Severity: PASS a lane succeeded · ATTENTION a capability gap · LANE / INFRA infrastructure, not capability · NOTE a property of the measurement.

F-01	ATTENTION
No evaluated lane passed the hidden tests in this run	
Severity: ATTENTION	Evidence: Model outcomes · Certification evidence
Every rollout either stopped before submitting or submitted an annotation submission the independent grader rejected. The exam's own certification (gold solution passes, empty submission fails) establishes the task itself is winnable.	
F-01 · derived from the recorded run data	
F-02	LANE / INFRA
model_api_failure × 1 (gemini-3.8-flash)	
Severity: LANE / INFRA	Evidence: Appendix A · Appendix C
The model endpoint failed to answer. Not a verdict about the model's ability. These rollouts are excluded from any capability reading.	
F-02 · derived from the recorded run data	
F-03	NOTE
The harness suspected itself during this run	
Severity: NOTE	Evidence: Appendix A · Appendix C
Self-suspicion rules fired: R1 api-failure share 1/1 — most rollouts never got a working completion. Treat the affected lanes' numbers as measurements of the harness, not the models, until the incident is resolved.	
F-03 · derived from the recorded run data	
F-04	NOTE
Statistical floor — 1 rollout(s) is below the 10-rollout rate floor	
Severity: NOTE	Evidence: Appendix D
The counts are exact; the percentages are anecdotes with decimal points. Per-rollout outcomes, not rates, are the reliable reading of this run.	
F-04 · derived from the recorded run data	

Recommendations

One recommendation per observed failure mode, addressed to the lane it affects. Each traces back to a finding above and to the raw trace in Appendix A.

R-01 PROBLEM
The model endpoint failed to answer. Not a verdict about the model's ability.

AFFECTED LANES
gemini-3.8-flash

WHAT WE WOULD LOOK AT FIRST
The lane failed, not the model; these rollouts carry no signal about capability and are labeled accordingly.

R-02 For anyone reading the percentages
Commission a run at $n \geq 10$ rollouts per lane before treating any rate here as a rate; this run's per-rollout outcomes are exact, its percentages are not yet stable.

Verification

RUN ID

live-20260906T201353Z

EXAM FINGERPRINT

c3b4773c375c10ad989357bb179cffd62151eccd2a58b52254609a54d434be33

The contract digest this run was executed against.

Provenance of this chapter

This chapter was rendered by the campaign generator from the sealed evaluation record; it is not the standalone report reproduced byte for byte. Verify the standalone report with its own digest, verify the record with `vvdex-env records verify`, and verify this campaign document as a whole.

RELATIONSHIP

campaign-rendered chapter

Derived from the same sealed evaluation record as the standalone report; these bytes are authenticated by the campaign artifact that contains them.

SEALED RECORD DIGEST

18e2b7ae2068b318b410dae8b7fe4b708d7995dec3d018510af3d60648261f85

sha256 over the record's canonical JSON — the identity every renderer works from.

SOURCE RECORD FILE

e8b4f8c0e787edf7b263d55b896d96f8c11201ab015c7735dec4d1d9ba9727c4

sha256 of `fr-20260906-55135e51.json` as written beside the standalone report.

SOURCE STANDALONE REPORT

8dcdf4ba476f15f2f29f575fe871fdebc07c182a6ae410e74d78ef1f05ac8134

sha256 of `fr-20260906-55135e51.html` — independently verifiable with its own digest.

SOURCE STANDALONE PDF

3ea795a2bc8fccf03e5a2d5a438f74b2db9a8cb7244cab5347ed947f85259263

sha256 of `fr-20260906-55135e51.pdf`.

RUN EVIDENCE BUNDLE

55135e51654ea5bd2770f6a0b89e0a853285efca33f33ee428224a59f919e260

The digest the run's own bundle computed over itself.

CERTIFIED EVIDENCE SEAL

a0997059eb829c98be90048ef2ed2bbc9debfb0d69af0e51674a13bda168bea1

The exam's sealed evidence, established before this run.

Measurement history

Previous run on this exam: `fr-20260906-29ba262c`. Model progress over time on this exam is the difference between that record and this one — same frozen tree, same hidden grader, same limits. Anything else that moved between them moved outside this exam.

The full artifact-by-artifact digest table, and the reproduction protocol, are in Appendix B.

Appendix A — every rollout

ROLLOUT	MODEL	SEED	OUTCOME	SUBMITTED	DEEPEST STAGE	HIDDEN TESTS	FILES	TOKENS	ELAPSED
119cd262	gemini-3.8-flash	0	model_api_failure	no	Not graded — lane error	not graded	0	n/a	30.1s

Failure taxonomy

CLASS	COUNT	WHAT IT MEANS
model_api_failure	1	The model endpoint failed to answer. Not a verdict about the model's ability.

Appendix B — evidence and artifact digests

RUN BUNDLE DIGEST 55135e51654ea5bd2770f6a0b89e0a853285efca33f33ee428224a59f919e260	RUN ID live-20260906T201353Z	STARTED 2026-09-06T20:13:53.938944+00:00	FINISHED 2026-09-06T20:14:24.344440+00:00
--	--	--	---

Artifact digests

ARTIFACT	SHA-256
publicSummary	5b9c659237882fc4e9ea681c1af961c83c739e3515b5f6df4c28fa637caccd05
rollouts/119cd262-bb93-49f5-925e-3e50a2970c08.json	0ab5cde25bfa9764f5901c96c00e1c6744d388b65976097ed36383f98b5c8337
spec	2e945ae8e17d799dec842ae4ae813364939bb0ea3b86208d17caa33fcf95bd6c
summary	e3cf7d4ff1281fbc0abbf130cc20b8c63ee5dd0508db390119f3941892e5926c

Reproducing this run

Every rollout can be re-derived: the exam is frozen (fingerprint above), the grader ran in the recorded container image, and each rollout's full trace — every tool call, every result, every word — is in the evidence bundle at its rollouts/<id>.json path. A re-run uses the same job shape: the exam id, your endpoint, n, and the runtime. The record id names this run forever.

Appendix C — redacted transcripts

For each rollout: the tool calls it made, in order — which tool, on which of the model's own paths, whether the call succeeded, and how many characters came back. What a tool returned is not here, and neither is the submission: on this exam the submitted content is the graded answer, so no diff, no answer value and no model claim is printed. A request for the reference solution or the hidden tests is recorded as a request; the path it asked for is replaced by a label. See Appendix D, "What this report withholds and why".

Disclosure class applied to this exam: answer-key disclosure (ownership vvdex_proprietary — no declaration — proprietary by default; family annotation): the exam is VVDex evaluation IP and the graded output is an answer or a keyed value, so no submission content, diff, model claim or hidden-assertion name is published.

119cd262 · gemini:gemini-3.8-flash · seed 0 · model_api_failure

STEP	TOOL	PATH	RESULT	CHARS
0	list_files	.	ok	79
1	read_file	TASK.md	ok	732
2	read_file	annotation-task.json	ok	4 329
3	read_file	submission.json	ok	74
4	run_tests	—	ok	186
5	list_files	—	ok	79

Submission diff withheld by the public-disclosure policy. Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

11 Attempt record 3 of 3 · robot-video-1

vvdex.annotation.robot-video-1 · record fr-20260906-99a8ead2 · record digest b6904bbd74376d69... · harness SUSPECT · containment clean

Executive summary

MODELS TESTED

1

CERTIFIED PASSES

0

of 0 graded rollouts

SUBMITTED

0 of 0

graded rollouts that called submit

NOT GRADED — LANE ERRORS

1

of 1 attempts

Purpose

This report records 1 interrupted attempt(s) against the certified exam vvdex.annotation.robot-video-1. No submission reached the grader. The frozen exam defines the workspace, tool contract and limits; this record documents where execution stopped, not a graded model result.

Outcome

0 of 0

graded rollouts passed the independent hidden tests and were certified — 1 of 1 attempts not graded (lane errors)

0 of 0 graded rollouts passed the hidden tests. The exam's certification establishes the task is winnable, so every failure below is a finding about a lane, not about the instrument. **Low-N:** 1 rollout(s) is below the 10-rollout floor for reading a rate as a rate — read the per-rollout outcomes, not the percentages. The most common way to fail was model_api_failure — the model endpoint failed to answer. not a verdict about the model's ability.

Key findings

- F-01 No evaluated lane passed the hidden tests in this run
- F-02 model_api_failure × 1 (gemini-3.8-flash)
- F-03 The harness suspected itself during this run
- F-04 Statistical floor — 1 rollout(s) is below the 10-rollout rate floor

Each finding is stated in full, with evidence, in the Findings section; recommendations for acting on them follow it. Elapsed wall clock for the whole engagement: 5.7s.

Exam definition

In scope. The ability of each configured model lane to complete this one certified task end to end — inspect the asset, annotate against the declared rubric, validate, submit, and be accepted by the independent annotation grader — within 24 tool steps and 2280s of wall clock, at 1 rollout(s) per lane. Behavioural measurements along the way (tool discipline, grounding, overclaim) are in scope because they are read from the recorded trace, not judged by another model.

Out of scope. General model quality, performance on other task families, prompt-tuning on the lane's behalf, and any comparison this run's sample size cannot support. A lane's API failure is reported as infrastructure, never as model capability.

The instrument. Exam `vvdex.annotation.robot-video-1` was certified before this run: its reference solution passes, an empty submission fails, its grader distinguishes near-misses, and its sealed evidence is unchanged by this evaluation (see Certification evidence).

What was measured

EXAM	VERSION	FAMILY	ROLLOUTS
vvdex.annotation.robot-video-1	0.1.0	annotation	1 (1 per model)

Where the defect came from

This exam has no upstream repository to credit: the defect, the reference solution and the hidden suite are VVDex's own.

Certification evidence

Why this exam is trustworthy. Every pillar below was established before any model saw the exam, loaded from the sealed evidence matching this run's exam fingerprint, and unchanged by this evaluation.

FROZEN BASELINE

FAIL

expected — an empty submission over 2 run(s) must not pass

→

GOLD / REFERENCE

PASS

expected — the reference solution over 2 run(s) must pass

GRADER CONTROLS

19 / 19

The grader accepted the reference and rejected every near-miss.

ATTACK PROBES

9 / 9 blocked

0 trivial exploit(s) found.

SEALED EVIDENCE

verified

Sealed 2026-09-06 over 13 artifact(s).

RUNTIME

sha256:679b36225a1f83238efba8c71b9cde3c24caaeacc9b682ae92819cb55eec328f

network policy: deny

CERTIFICATION BUNDLE DIGEST

a0997059eb829c98be90048ef2ed2bbc9debfb0d69af0e51674a13bda168bea1

EXAM FINGERPRINT

c3b4773c375c10ad989357bb...

Full value in Appendix B; the contract digest this run was executed against.

Evaluation configuration

REPORT	fr-20260906-99a8ead2
ISSUED	2026-09-06
SUBJECT EXAM	vvdex.annotation.robot-video-1 · v0.1.0
FAMILY	annotation
PREPARED BY	VVDex Forge engine vvdex-env 0.1.0
CLASSIFICATION	Independent model evaluation report — independently verifiable
CERTIFICATION	Exam certification: recorded and passing (see Certification evidence)
STATUS	FINAL · report sealed against its own digest

Timeline

WHEN (UTC)	EVENT
2026-09-06 00:01:41	Exam evidence sealed (before any model saw the exam)
2026-09-06 20:07:21	Evaluation run started
2026-09-06 20:07:27	Evaluation run finished
2026-09-06 20:07:27	Report generated from the sealed run evidence

Models under test

MODEL	PROVIDER	ENDPOINT	BILLING
gemini-3.8-flash	gemini	VVDex-configured	operator API key; billing tier and monetary charge not recorded

A customer endpoint is recorded by origin only — never its port, its path, or its key.

Limits

RUNTIME docker	STEP LIMIT 24	WALL-CLOCK LIMIT 2280s	MAX OUTPUT TOKENS 2048
TEMPERATURE 0.2	RETRY POLICY none	COMMAND BUDGET 0	TOOLS OFFERED list_files, read_file, write_file, edit_file, run_tests, submit

The evaluated model never saw the hidden tests or the reference solution, and could not change its result. Grading ran in a separate step, after submission, on a container with no network.

Model outcomes

Denominators. Attempts requested: 1. Graded (evaluable) rollouts: 0. Not graded — lane errors: 1. A lane error is a rollout the API or the runtime ended before the hidden grader ran: it is listed, excluded from every rate and pass@k, and never silently dropped. Passes are certified passes: hidden grader exit 0 with integrity verified.

MODEL	ATTEMPTS	GRADED	PASSED	MODEL FAILS	LANE ERRORS	SUBMITTED	MEAN TIME	TYPICAL DEEPEST STAGE
gemini-3.8-flash gemini-3.8-flash	1	0	not graded	0	1	—	—	nothing

Every rollout, with its own deepest stage and grade, is in Appendix A.

Success rate, confidence and pass@k

Every rate on this page is a measurement of a finite number of rollouts, so every rate carries the interval that measurement actually supports. n is the evaluable count — attempts minus rollouts the lane ended before grading; those are listed under "not graded" and sit in no rate, interval or pass@k.

MODEL	ATTEMPTS	N	NOT GRADED	PASSES	RATE	95% INTERVAL	SEEDS
gemini-3.8-flash	1	0	1	0	not recorded	[0%, 100%]	0

Interval and pass@k methods are stated in full in Appendix D. pass@k is shown for k = ; the sealed record carries every k up to n.

Model comparison

One model ran in this record, so there is nothing to compare it with. A comparison table across models appears here when a run covers more than one.

Stage funnel

Recorded annotation actions and independent grader verdicts. Missing capabilities are N/A; each stage is measured independently.

STAGE	GEMINI:GEMINI-3.8-FLASH
Inspect	0 / 0
Annotate	0 / 0
Validate	0 / 0
Submit	0 / 0
Graded	0 / 0
Passed	0 / 0

Deepest stage reached

MODEL	DEEPEST (ANY ATTEMPT)	TYPICAL (MOST ATTEMPTS)	ANNOTATION QA PASSES	NOT GRADED
gemini:gemini-3.8-flash	nothing	nothing	0 / 0	1

The matrix counts evaluable annotation attempts; lane errors are shown separately.
Successful recorded actions establish each stage independently.

Annotation evidence

Modality: video. **Rubric:** 1.0.

A robot-catching video is reviewed for temporal events, frame-level object boxes, track identity and occlusion.

Measured grader aggregates only. Missing metrics or disagreement are not measured, never zero. Reference answers and item-level labels are excluded. Certification and the evidence digest are recorded in this report's verification sections.

Behavioural analysis

Long-horizon behaviour

Counted off the recorded trace of each rollout. A dead end is a step that moved nothing: a re-read of a slice already returned, a repeat of the previous action, or a call to a tool this exam does not have.

MODEL	N	STEPS USED	RAN OUT OF STEPS	COMMANDS	CHANGED THE TREE	REFUSED (OVER BUDGET)	REVISITS	DEAD ENDS / ROLLOUT
gemini:gemini-3.8-flash	1	0 / 24	0	0 / 0	0	0	0	0

Each column is defined in Appendix D.

Hallucination & overclaim

Three measurements, each with a stated definition. None of them is a judgement of the model's prose by another model — every number below comes from comparing what the model wrote against what was on disk, or against what the independent grader returned.

MODEL	N	GROUNDING FAILURES	KINDS	ROLLOUTS AFFECTED	OVERCLAIMS	OVERCLAIM RATE	RE-READS
gemini:gemini-3.8-flash	1	0	none	0 / 1	0 / 1	0.0%	0

No grounding failures and no false pass claims were detected across these 1 rollout(s). No lane re-read material it had already been given.

Every term above is defined in Appendix D.

Efficiency & telemetry

This run has no per-token price attached — a local model, a free quota, or a customer endpoint we are not billed for — so spend is reported in tokens. Tokens are what was actually measured; a dollar figure here would be invented. Lanes run through a flat-rate local CLI report no token telemetry; their cells say n/a — an absent measurement, not zero. Rollout counts here are evaluable rollouts.

MODEL	ROLLOUTS	TOKENS	TOKENS / ROLLOUT	COST	PASSES	RATE
gemini:gemini-3.8-flash	0	n/a	n/a	n/a – telemetry unavailable	0 / 0	0.0%

Findings

Every finding below is derived from the recorded data of this run — none is an opinion.

Severity: PASS a lane succeeded · ATTENTION a capability gap · LANE / INFRA infrastructure, not capability · NOTE a property of the measurement.

F-01	ATTENTION
No evaluated lane passed the hidden tests in this run	
Severity: ATTENTION	Evidence: Model outcomes · Certification evidence
Every rollout either stopped before submitting or submitted an annotation submission the independent grader rejected. The exam's own certification (gold solution passes, empty submission fails) establishes the task itself is winnable.	
F-01 · derived from the recorded run data	

F-02	LANE / INFRA
model_api_failure × 1 (gemini-3.8-flash)	
Severity: LANE / INFRA	Evidence: Appendix A · Appendix C
The model endpoint failed to answer. Not a verdict about the model's ability. These rollouts are excluded from any capability reading.	
F-02 · derived from the recorded run data	

F-03	NOTE
The harness suspected itself during this run	
Severity: NOTE	Evidence: Appendix A · Appendix C
Self-suspicion rules fired: R1 api-failure share 1/1 — most rollouts never got a working completion. Treat the affected lanes' numbers as measurements of the harness, not the models, until the incident is resolved.	
F-03 · derived from the recorded run data	

F-04	NOTE
Statistical floor — 1 rollout(s) is below the 10-rollout rate floor	
Severity: NOTE	Evidence: Appendix D
The counts are exact; the percentages are anecdotes with decimal points. Per-rollout outcomes, not rates, are the reliable reading of this run.	
F-04 · derived from the recorded run data	

Recommendations

One recommendation per observed failure mode, addressed to the lane it affects. Each traces back to a finding above and to the raw trace in Appendix A.

R-01 PROBLEM
The model endpoint failed to answer. Not a verdict about the model's ability.

AFFECTED LANES
gemini-3.8-flash

WHAT WE WOULD LOOK AT FIRST
The lane failed, not the model; these rollouts carry no signal about capability and are labeled accordingly.

R-02 For anyone reading the percentages
Commission a run at $n \geq 10$ rollouts per lane before treating any rate here as a rate; this run's per-rollout outcomes are exact, its percentages are not yet stable.

Verification

RUN ID

live-20260906T200721Z

EXAM FINGERPRINT

c3b4773c375c10ad989357bb179cffd62151eccd2a58b52254609a54d434be33

The contract digest this run was executed against.

Provenance of this chapter

This chapter was rendered by the campaign generator from the sealed evaluation record; it is not the standalone report reproduced byte for byte. Verify the standalone report with its own digest, verify the record with `vvdex-env records verify`, and verify this campaign document as a whole.

RELATIONSHIP

campaign-rendered chapter

Derived from the same sealed evaluation record as the standalone report; these bytes are authenticated by the campaign artifact that contains them.

SEALED RECORD DIGEST

b6904bbd74376d6911ded609ccc9d9f27c8fbd000f85f82333568dd479bf8bfc

sha256 over the record's canonical JSON — the identity every renderer works from.

SOURCE RECORD FILE

79135d55d6519822a7c38506299bccf4d1035b824bb117760d025567e945da0c

sha256 of `fr-20260906-99a8ead2.json` as written beside the standalone report.

SOURCE STANDALONE REPORT

e6fa93aeeeb2ed067a49f1723cded691245cfefc9637e6066b02e34e2a3f168d

sha256 of `fr-20260906-99a8ead2.html` — independently verifiable with its own digest.

SOURCE STANDALONE PDF

5047fbd8c3325aba9eb4f26fae21b96e41be4bd59bbc926a0fb6323f28ef215c

sha256 of `fr-20260906-99a8ead2.pdf`.

RUN EVIDENCE BUNDLE

99a8ead2e17682cd95fd622f906894b3f4b95c000d5349fc43cbf0c3486c91ea

The digest the run's own bundle computed over itself.

CERTIFIED EVIDENCE SEAL

a0997059eb829c98be90048ef2ed2bbc9debfb0d69af0e51674a13bda168bea1

The exam's sealed evidence, established before this run.

Measurement history

This is the first recorded run on this exam. There is nothing to compare it against yet, and a single measurement is not a trend.

The full artifact-by-artifact digest table, and the reproduction protocol, are in Appendix B.

Appendix A — every rollout

ROLLOUT	MODEL	SEED	OUTCOME	SUBMITTED	DEEPEST STAGE	HIDDEN TESTS	FILES	TOKENS	ELAPSED
2e00a715	gemini-3.8-flash	0	model_api_failure	no	Not graded — lane error	not graded	0	n/a	5.4s

Failure taxonomy

CLASS	COUNT	WHAT IT MEANS
model_api_failure	1	The model endpoint failed to answer. Not a verdict about the model's ability.

Appendix B — evidence and artifact digests

RUN BUNDLE DIGEST 99a8ead2e17682cd95fd622f906894b3f4b95c000d5349fc43cbf0c3486c91ea	RUN ID live-20260906T200721Z	STARTED 2026-09-06T20:07:21.820541+00:00	FINISHED 2026-09-06T20:07:27.540793+00:00
--	--	--	---

Artifact digests

ARTIFACT	SHA-256
publicSummary	c74a58db7f590c3ba729ebc4b2a471e3250ea83cd716e6a7e99f08078c757846
rollouts/2e00a715-387c-4270-b307-65899b92f193.json	3768450d6e3d0084ff72f34e5398ddf34419d5167e1ea0fa1e5b42e2f0a5a382
spec	63e7e1420d141af59794668f9e42c47f4507a3d1e1f5c672202f160cf5c8c210
summary	ac9f8c135b7f2bb173d348fe5a7f82c39ea69d766816aa55b0e6c14489440eab

Reproducing this run

Every rollout can be re-derived: the exam is frozen (fingerprint above), the grader ran in the recorded container image, and each rollout's full trace — every tool call, every result, every word — is in the evidence bundle at its rollouts/<id>.json path. A re-run uses the same job shape: the exam id, your endpoint, n, and the runtime. The record id names this run forever.

Appendix C — redacted transcripts

For each rollout: the tool calls it made, in order — which tool, on which of the model's own paths, whether the call succeeded, and how many characters came back. What a tool returned is not here, and neither is the submission: on this exam the submitted content is the graded answer, so no diff, no answer value and no model claim is printed. A request for the reference solution or the hidden tests is recorded as a request; the path it asked for is replaced by a label. See Appendix D, "What this report withholds and why".

Disclosure class applied to this exam: answer-key disclosure (ownership vvdex_proprietary — no declaration — proprietary by default; family annotation): the exam is VVDex evaluation IP and the graded output is an answer or a keyed value, so no submission content, diff, model claim or hidden-assertion name is published.

2e00a715 · gemini:gemini-3.8-flash · seed 0 · model_api_failure

STEP	TOOL	PATH	RESULT	CHARS
No tool call was recorded for this rollout.				

Submission diff withheld by the public-disclosure policy. Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

12 Evidence

Every number above is recomputable from the sealed records below; each record verifies itself with `vvdex-env records verify`, and this document was regenerated from them with `vvdex-env records campaign`.

- `vvdex.annotation.robot-video-1` → `fr-20260906-29ba262c` · `recordDigest` `51bedf34d0f35b8337759d188245153ab277684082c83f1d367ba9586133e917`
- `vvdex.annotation.robot-video-1` → `fr-20260906-55135e51` · `recordDigest` `18e2b7ae2068b318b410dae8b7fe4b708d7995dec3d018510af3d60648261f85`
- `vvdex.annotation.robot-video-1` → `fr-20260906-99a8ead2` · `recordDigest` `b6904bbd74376d6911ded609ccc9d9f27c8fbd000f85f82333568dd479bf8bfc`

13 Attestation

Certified results appear only where evidence proves them. Withheld, invalid and lane-error cells are never presented as model results, and no historical evidence was modified in producing this document. Counts are exact; below ten rollouts per cell no rate or ranking beyond the counts is asserted.

© VVDex. All rights reserved. Public materials are provided for viewing, evaluation and verification. Reuse, redistribution, derivative benchmark creation, training use, republication or commercial use requires written permission unless a specific file states otherwise. Full terms: vvdexops.com/terms/. Third-party open source reproduced in an exam keeps its own licence.