

Gemini 3.6 Flash native-video exam — revision 0.2.0

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

HISTORICAL — SUPERSEDED
CAMPAIGN REPORT

2026-09-06

Evaluation scope:
gemini-3.6-flash

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

fc-2917108cf6be · 1 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 1 · robot-video-1	11
10.	Evidence	27
11.	Attestation	28

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Release status: HISTORICAL — superseded.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 1	MODEL LANES 1	ATTEMPTS 1
GRADED ATTEMPTS 1	CERTIFIED PASSES 0	LANE ERRORS (NOT GRADED) 0	WITHHELD / INVALID 0

Purpose

This report presents a controlled evaluation of 1 model lane(s) against 1 fixed certified annotation tasks, one sealed attempt record on each task. For each task, every lane received that task's frozen workspace, tool contract and limits; grading ran afterwards, in isolation, against hidden tests no lane ever saw; and every rollout carries three separate outcomes — grader, integrity, certified — of which only the certified outcome is counted. 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 one sealed attempt record on each task (1 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. Quality measurements are reported per task and remain separate from the strict certified-pass gate.

Outcome

0

certified pass(es) across 1 graded attempts (1 attempts, 0 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-2917108cf6be · generated from 1 verified sealed record(s)

Key findings across the campaign

- robot-video-1 · gemini-3.6-flash · **ATTENTION** — 0 certified passes / 1 graded attempts; 1 model failure; 0 excluded (0 lane errors, 0 withheld, 0 invalid).

02 Per-exam results

MODEL LANE	ANNOTATION
gemini-3.6-flash	VVDEX.ANNOTATION.ROBOT-VIDEO-1
	0/1

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-video-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-3.6-flash	0	1	0	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

No record in this campaign fired a harness self-suspicion rule. Every countable cell is a grader outcome under verified containment.

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.

No attempt in this campaign was excluded: every attempt produced a countable model result.

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 1 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 1 · robot-video-1

vvdex.annotation.robot-video-1 · record fr-20260906-76fdb6cd · record digest 59f431f7ed9ec521... · harness clean · containment clean

Executive summary

MODELS TESTED

1

CERTIFIED PASSES

0

of 1 graded rollouts

SUBMITTED

1 of 1

graded rollouts that called submit

NOT GRADED — LANE ERRORS

0

of 1 attempts

Purpose

This report presents a controlled evaluation of 1 model lane(s) against the certified exam vvdex.annotation.robot-video-1 (v0.2.0, family annotation). Every lane received the identical frozen workspace, tool contract and limits; grading ran afterwards, in isolation, against hidden tests no lane ever saw. The purpose is to state, with reproducible evidence, what each lane did and where it stopped.

Outcome

0 of 1

graded rollouts passed the independent hidden tests and were certified

0 of 1 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 submitted_failed — submitted annotations rejected by the independent grader.

Key findings

- F-01 No evaluated lane passed the hidden tests in this run
- F-02 submitted_failed × 1 (gemini-3.6-flash)
- F-03 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: 97.2s.

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	0.2.0	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

0a0a51360cde69737c9b5caef895b8b7c2a5e2d66da89617d8776ac724b019b9

EXAM FINGERPRINT

85e026007af17db692614084...

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

Evaluation configuration

REPORT	fr-20260906-76fdb6cd
ISSUED	2026-09-06
SUBJECT EXAM	vvdex.annotation.robot-video-1 · v0.2.0
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 21:19:43	Exam evidence sealed (before any model saw the exam)
2026-09-06 21:21:04	Evaluation run started
2026-09-06 21:22:41	Evaluation run finished
2026-09-06 21:22:41	Report generated from the sealed run evidence

Models under test

MODEL	PROVIDER	ENDPOINT	BILLING
gemini-3.6-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: 1. Not graded — lane errors: 0. 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-3.6-flash gemini-3.6-flash	1	1	0 / 1	1	0	1 / 1	96.8s	Graded

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	PASS@1
gemini-3.6-flash	1	1	0	0	0.0%	[0%, 79%]	0	0.0%

Interval and pass@k methods are stated in full in Appendix D. pass@k is shown for k = 1; 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-3.6-FLASH
Inspect	1 / 1
Annotate	1 / 1
Validate	1 / 1
Submit	1 / 1
Graded	1 / 1
Passed	0 / 1

Deepest stage reached

MODEL	DEEPEST (ANY ATTEMPT)	TYPICAL (MOST ATTEMPTS)	ANNOTATION QA PASSES	NOT GRADED
gemini:gemini-3.6-flash	Graded	Graded	0 / 1	0

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

Annotation evidence

Modality: video. Rubric: 1.1.

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

Submission 9fa86ff8-9ca9-4227-9817-2eecbc425dd7

Score	Items	Valid / invalid	Types
0.4211	6	6 / 0	box, classification, event, segment

Measured metrics

annotationCount	6	boxF1	0	boxIoU	0
boxPrecision	0	boxRecall	0	classificationAccuracy	1
classificationF1	1	classificationPrecision	1	classificationRecall	1
eventF1	0.5	eventPrecision	0.5	eventRecall	0.5
eventTemporalIoU	0.3132	matchedCount	4	mismatchCount	5
segmentF1	1	segmentPrecision	1	segmentRecall	1
segmentTemporalIoU	0.9	Track consistency on matched boxes	0		

Track consistency depends on successful box matching. A zero can reflect inaccurate boxes even when track IDs are correct.

Validation error codes

None recorded

Disagreement

Agreement review	Not measured
------------------	--------------

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.

Quality axes

Each axis was measured inside the grade box, on the submitted workspace, by the same deterministic script for every rollout. The reward is the conjunction of the required axes; the score is a weighted reading beside it, not the gate.

AXIS	GATE	ROLLOUTS PASSING	MEASURED	BUDGET	WEIGHT	WHAT IT MEASURES
boxPrecision	recorded	N/A	0	–	–	
boxRecall	recorded	N/A	0	–	–	
boxF1	recorded	N/A	0	–	–	
boxIoU	recorded	N/A	0	–	–	
classificationPrecision	recorded	N/A	1	–	–	
classificationRecall	recorded	N/A	1	–	–	
classificationF1	recorded	N/A	1	–	–	
classificationAccuracy	recorded	N/A	1	–	–	
eventPrecision	recorded	N/A	0.5	–	–	
eventRecall	recorded	N/A	0.5	–	–	
eventF1	recorded	N/A	0.5	–	–	
eventTemporalIoU	recorded	N/A	0.3132	–	–	
segmentPrecision	recorded	N/A	1	–	–	
segmentRecall	recorded	N/A	1	–	–	
segmentF1	recorded	N/A	1	–	–	
segmentTemporalIoU	recorded	N/A	0.9	–	–	
trackingConsistency	recorded	N/A	0	–	–	

QUALITY SCORE (MEAN)

0.4211

Weighted reading over the axes. It is not the reward.

REWARD COMPOSITION

Withheld from the public report.

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-3.6-flash	1	7 / 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-3.6-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-3.6-flash	1	23 820	23 820	not billed	0 / 1	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	ATTENTION
submitted_failed × 1 (gemini-3.6-flash)	
Severity: ATTENTION	Evidence: Appendix A · Appendix C
Submitted annotations rejected by the independent grader.	
F-02 · derived from the recorded run data	

F-03	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-03 · 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

Submitted annotations rejected by the independent grader.

AFFECTED LANES

gemini-3.6-flash

WHAT WE WOULD LOOK AT FIRST

Review the measured annotation metrics and validator error codes against the declared rubric.

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-20260906T212104Z

EXAM FINGERPRINT

85e026007af17db692614084f12ff40c5bbf2348bcfda0a01deb95aad8928f4c

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

59f431f7ed9ec5214737424a5c886cbe65f3caf49b1a11807bc27be17428b613

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

SOURCE RECORD FILE

7866a5ac234ee1701448f860b29b540a64a4539e6757a0e284b59038f547bb6b

sha256 of fr-20260906-76fdb6cd.json as written beside the standalone report.

SOURCE STANDALONE REPORT

e8ab74b0502248879a947364d62090a2f1f9350b8336376c23d8bacb60ed7157

sha256 of fr-20260906-76fdb6cd.html — independently verifiable with its own digest.

SOURCE STANDALONE PDF

53cdace6e4d9ee4e5343ab3cf5dbbc47a9918550c558c696e6734b0122605318

sha256 of fr-20260906-76fdb6cd.pdf.

RUN EVIDENCE BUNDLE

76fdb6cdcc186ebb6aa7d8a7daed7005b091a4676d57e0f3a1e16782872cc9c6

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

CERTIFIED EVIDENCE SEAL

0a0a51360cde69737c9b5cae f895b8b7c2a5e2d66da89617d8776ac724b019b9

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
9fa86ff8	gemini-3.6-flash	0	submitted_failed	yes	Graded	fail	1	23 820	96.8s

Failure taxonomy

CLASS	COUNT	WHAT IT MEANS
submitted_failed	1	Submitted annotations rejected by the independent grader.

Appendix B — evidence and artifact digests

RUN BUNDLE DIGEST

76fdb6cdcc186ebb6aa7d8a7daed7005b091a4676d57e0f3a1e16782872cc9c6

RUN ID

live-20260906T212104Z

STARTED

2026-09-06T21:21:04.395002+00:00

FINISHED

2026-09-06T21:22:41.548941+00:00

Artifact digests

ARTIFACT	SHA-256
publicSummary	c71502ca0b14aee6699b51d8e890adad77a0777b600795389d6ce004a9f5b8b3
rollouts/9fa86ff8-9ca9-4227-9817-2eecbc425dd7.json	39af27f7e11c3bfe3cbcd0d7818355d36619bac09680a0052b31e7a45b9189be
spec	c77ec4d2d23615e0ebcfc74fc08254cc54540a1a5cf379da38dea1543029e097
summary	450021ad8b722386568260b00dc438b95bdc5d87a3a5999cfdb38c9fb00d97d8

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.

9fa86ff8 · gemini:gemini-3.6-flash · seed 0 · submitted_failed

STEP	TOOL	PATH	RESULT	CHARS
0	list_files	.	ok	79
1	read_file	TASK.md	ok	1 119
2	read_file	annotation-task.json	ok	4 867
3	read_file	submission.json	ok	74
4	write_file	submission.json	ok	64
5	run_tests	—	ok	160
6	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (61 line(s) changed). 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 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-20260906-76fdb6cd` · `recordDigest` `59f431f7ed9ec5214737424a5c886cbe65f3caf49b1a11807bc27be17428b613`

11 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.