GEOOWL

SYNTHETIC DEMONSTRATION / v1.0

GEOINT AI EVALUATION

Building-change detection. An example client report.

A controlled synthetic comparison of two fictional prediction sets. The package demonstrates a traceable method, an editable scorecard, failure inspection, and a repeatable evidence trail.

48	96	2
EVALUATION CASES	REFERENCE EVENTS	FICTIONAL CANDIDATES

Document date: 18 November 2026. All example systems, observations, geometries, review times, and cost inputs are fictional. Numerical results are calculated from the supplied synthetic fixtures. No operational model was executed, no real imagery was evaluated, and no accreditation is represented.

02 / EXAMPLE CLIENT DELIVERABLE

Neither meets every gate.

The demonstration is deliberately constructed to reveal a tradeoff. Candidate B reduces false alerts by 71.4% relative to Candidate A, while missing four additional reference events.

75.0 / 87.5%

A: PRECISION / RECALL

90.9 / 83.3%

B: PRECISION / RECALL

Illustrative acceptance criterion	Candidate A	Candidate B
Precision at least 90%	75.0% - NOT MET	90.9% - MET
Recall at least 85%	87.5% - MET	83.3% - NOT MET
False alerts / km2-month at most 0.25	0.583 - NOT MET	0.167 - MET
Every condition recall at least 75%	Worst 75.0% - MET	Worst 62.5% - NOT MET
All four criteria satisfied	NO	NO

EXAMPLE DECISION

For a real program with these agreed criteria, neither result would support unconditional acceptance. Candidate B illustrates a promising reduction in false-alert burden, with a recall gap requiring focused investigation and retest.

These are assumed demonstration criteria. The fixture is not a statistical sample of a deployment population. No operational suitability, superiority, or safety claim follows from these results.

03 / EXAMPLE CLIENT DELIVERABLE

The evaluation charter.

Dimension	Frozen demonstration definition
Task	Identify newly observable building footprints in a case and observation period
Reference set	96 constructed new-building polygons across 48 synthetic cases
Case structure	12 fictional AOIs, four monthly observation periods, 1 km2 exposure per case
Period coverage	July through October 2026; time fields are scenario metadata
Condition groups	Clear, Seasonal, Registration, Obscured; 12 cases and 24 reference events per group
Candidate inputs	Two fictional v1.0 feature sets. No trained model or live inference produced them
Output boundary	One-to-one comparison of predicted and reference polygons within the same case
Geographic meaning	Artificial coordinates and geometries; no real facilities or sites are represented

What is intentionally outside this example

Real imagery interpretation, physical construction dates, collection availability, learned-model generalization, end-to-end API behavior, operational security assessment, and human-system performance are not tested.

REFERENCE DATA IS CONSTRUCTED TRUTH

The geometric fixture supplies known reference answers. Workload fields illustrate a process; they do not document a real analyst adjudication exercise.

04 / EXAMPLE CLIENT DELIVERABLE

Follow the evidence chain.

Stage	Evidence retained	Control
1. Define	Task, case catalog, condition labels, and thresholds	Protocol and data version recorded
2. Construct	Reference and candidate geometries	Seeded generation; all data synthetic
3. Match	Per-event TP, FP, and FN outcomes	Same case; same class; one-to-one IoU >= 0.50
4. Aggregate	Case-level counts and condition totals	Counts reconcile to source geometries
5. Report	Scorecard, figures, and decision criteria	Same canonical data feeds every format

Metric definitions

Precision = TP / (TP + FP). **Recall** = TP / (TP + FN). **F1** = 2TP / (2TP + FP + FN). Aggregate percentages are computed from pooled counts, not by averaging case percentages.

False-alert burden = FP / monitored exposure. The 48 one-square-kilometer monthly cases provide 48 km2-month of exposure per candidate, including cases with no reference events.

Matching uses polygon overlap in the fixture local-metre coordinate system. GeoJSON files are provided for interoperability with documented artificial geographic anchoring. Coordinate conventions and scoring units are explicit in the data dictionary.

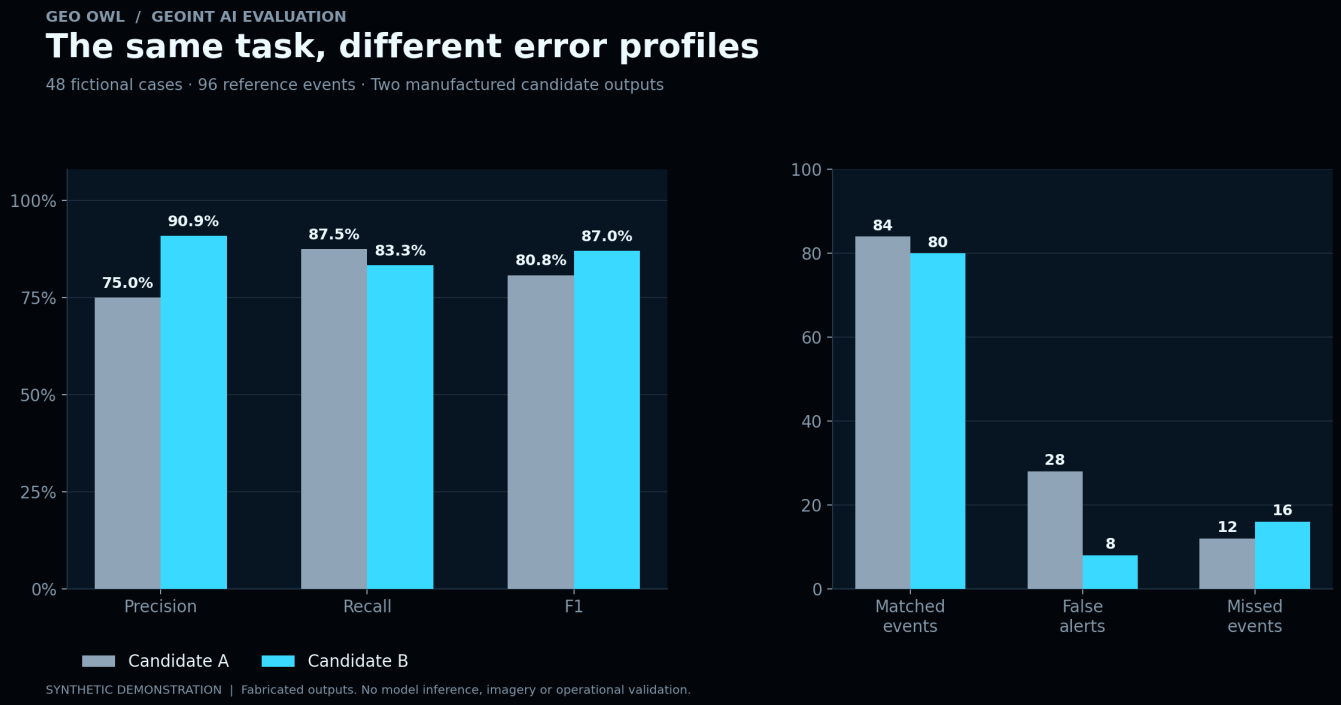
VERIFICATION STANDARD

Known-answer checks exercise perfect, empty, duplicate, and shifted predictions. Reconciliation checks confirm 48 cases per candidate, 96 reference events, and identical reference/exposure denominators.

This scorer supports only the axis-aligned rectangles in the fixture. It is not a production evaluator for arbitrary polygons. No confidence intervals or significance tests are reported: constructed fixtures do not justify population-level uncertainty estimates.

05 / EXAMPLE CLIENT DELIVERABLE

Precision and recall diverge.



Count or measure	Candidate A	Candidate B
Matched reference events / TP	84	80
False alerts / FP	28	8
Missed reference events / FN	12	16
Total predictions	112	88
F1	80.77%	86.96%

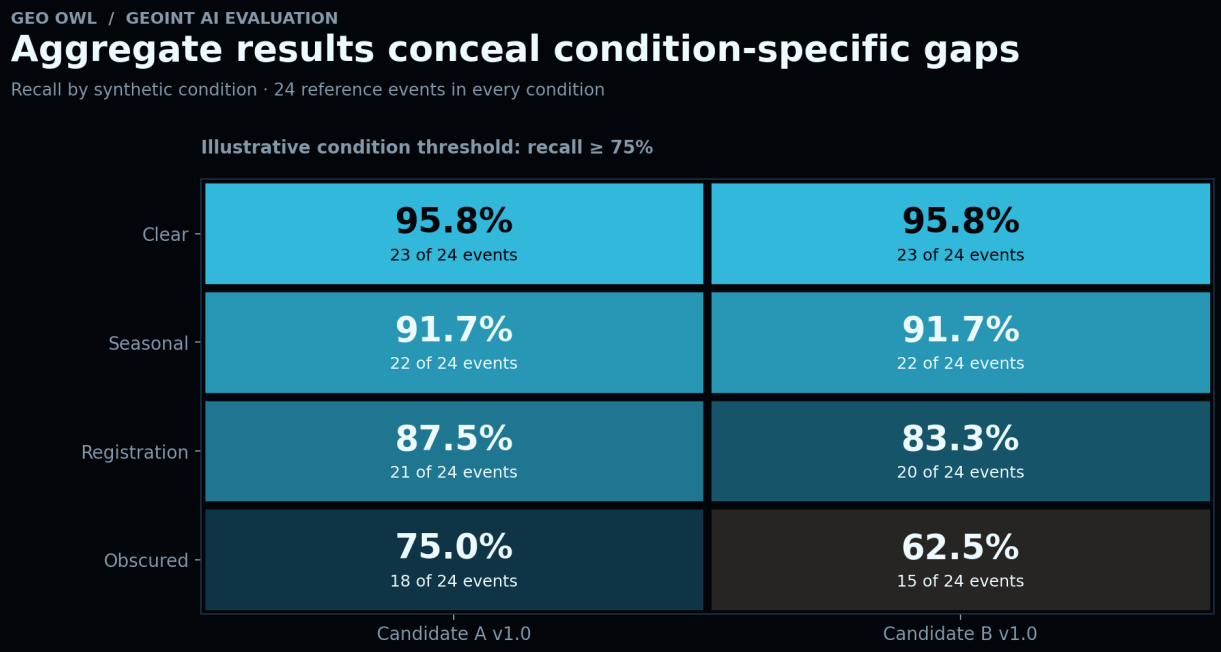
A HIGHER F1 DOES NOT OVERRIDE A REQUIREMENT

Candidate B has the higher pooled F1 in this fixture. Its recall remains below the assumed 85% gate, and one condition group also misses its recall gate. The report therefore preserves both the aggregate result and the acceptance gaps.

The results describe fabricated prediction sets. They are not scores of an actual commercial model or Geo Owl product.

06 / EXAMPLE CLIENT DELIVERABLE

Inspect the weak condition.



SYNTHETIC DEMONSTRATION | Fabricated outputs. No model inference, imagery or operational validation.

Condition	Reference events	A: TP / recall	B: TP / recall
Clear	24	23 / 95.83%	23 / 95.83%
Seasonal	24	22 / 91.67%	22 / 91.67%
Registration	24	21 / 87.50%	20 / 83.33%
Obscured	24	18 / 75.00%	15 / 62.50%

FOCUSED RETEST QUESTION

Can the candidate recover missed observable changes in the difficult condition without recreating the false-alert burden? A real retest would require representative, reserved cases and an agreed operating-domain definition.

Condition names describe designed scenario groups. They are not measurements of real cloud cover, image registration, seasonality, or sensor performance.

07 / EXAMPLE CLIENT DELIVERABLE

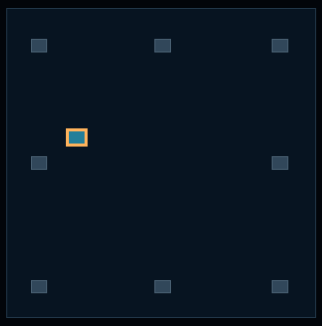
Make the failures inspectable.

GEO OWL / GEOINT AI EVALUATION

An evaluator must make errors inspectable

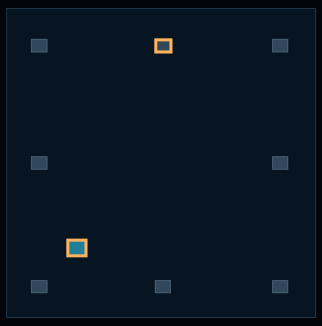
Geometric schematics only · Existing structures in slate · New references in cyan · Candidate outlines in amber

Matched prediction



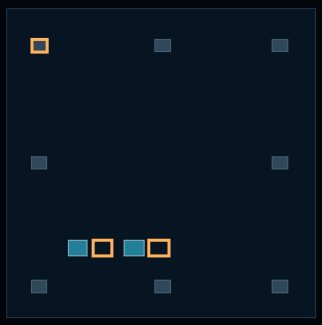
SYN-01-03 · Candidate A v1.0 · TP 1 / FP 0 / FN 0

False change on an existing building



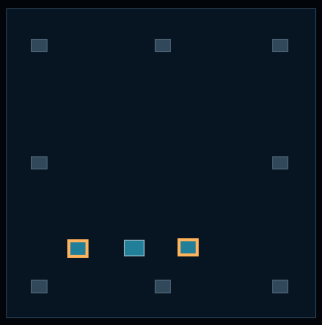
SYN-04-01 · Candidate A v1.0 · TP 1 / FP 1 / FN 0

Displaced prediction



SYN-07-01 · Candidate A v1.0 · TP 0 / FP 3 / FN 2

Omitted new building



SYN-10-01 · Candidate B v1.0 · TP 2 / FP 0 / FN 1

SYNTHETIC DEMONSTRATION | Fabricated outputs. No model inference, imagery or operational validation.

Case trace: Matched: SYN-01-03 / A. False change: SYN-04-01 / A.
Displaced: SYN-07-01 / A. Omitted building: SYN-10-01 / B.
Each schematic links to its case, geometry, and scoring record in the data package.

WHAT A REAL FAILURE REVIEW WOULD ADD

Source imagery, acquisition context, reference uncertainty, reviewer adjudication, operational consequence, and a tested explanation of the cause. A geometric miss alone does not establish why an actual model failed.

All drawings are synthetic geometric schematics. They are not satellite images, imagery interpretation, or evidence of a real construction event.

08 / EXAMPLE CLIENT DELIVERABLE

Expose the operating assumptions.

Workload and execution fields in this package are synthetic scenario inputs. The arithmetic below is reproducible, but it is not measured human effort, actual inference latency, or a vendor cost benchmark.

Illustrative workload measure	Candidate A	Candidate B
Review time, minutes	264.0	218.0
Review labor at $90/hour	$396.00	$327.00
Simulated inference expense	$0.23	$0.39
Modeled review + inference expense	$396.23	$327.39
Expense per matched reference event	$4.72	$4.09

Calculation

Modeled expense = review minutes / 60 x hourly review rate + simulated inference expense. Expense per matched reference event divides that amount by TP. The workbook exposes the review-rate assumption.

NOT A COMPLETE UNIT ECONOMICS MODEL

Imagery acquisition, reference labeling, engineering, hosting overhead, customer integration, rework, and the consequences of missed events are excluded. Lower modeled expense does not resolve a failed recall requirement.

A paid engagement would measure

Reviewer minutes by task; case failures and retries; elapsed time to a usable output; actual compute charges; and any operational impact of false or missed alerts. Model-only and human-corrected results would be reported separately.

09 / EXAMPLE CLIENT DELIVERABLE

Turn gaps into a retest plan.

Finding	Illustrative action	Evidence required next
A: precision 75%; false-alert burden 0.583	Inspect duplicate, misregistration, and no-change alerts; tune only on development cases	Fresh test shows precision >=90% and false-alert burden <=0.25
B: overall recall 83.33%	Review missed-event patterns and any confidence threshold tradeoff	Reserved cases meet recall >=85% without losing precision
B: obscured-group recall 62.50%	Define observability boundaries and review/abstention rules	Relevant condition groups meet >=75% recall or scope is explicitly revised
No operational deployment evidence	Run an authorized end-to-end evaluation in the customer environment	Source retrieval, timestamps, delivery, review, failures, and cost measured

Retest controls

Freeze a new candidate version and the revised test plan. Preserve the original results. Identify cases exposed during remediation. Evaluate reserved cases before making a generalization claim. Compare performance and review burden at the same task and exposure definitions.

EXAMPLE CUSTOMER DECISION

Proceed to a focused evaluation or remediation work package if the decision owner agrees. This sample does not recommend operational deployment and does not establish a real defect in any product.

The suggested investigations are examples of what a real failure review might consider. The fixture was deliberately manufactured; there is no learned-model root cause to diagnose.

10 / EXAMPLE CLIENT DELIVERABLE

The evidence package.

Artifact	How the client uses it
Evaluation report (this PDF)	Decision summary, scope, results, failure examples, limitations, and retest plan
Excel scorecard	Editable acceptance thresholds and labor-rate assumptions; traceable pooled and condition calculations
PowerPoint briefing	Concise, editable readout for a program owner or technical buyer
Plan and pilot scope PDF	Example charter, protocol, intake requirements, deliverables, and commercial boundaries
CSV / JSON / GeoJSON + code	Case records, reference and candidate features, metrics, schema, manifest, and reproducible scorer

Method references

NIST AI Risk Management Framework 1.0 - documented measurement, limitations, and lifecycle monitoring.
https://airc.nist.gov/airmf-resources/airmf/5-sec-core/

NIST ARIA Evaluation Planning Manual (2026) - distinguishing model, red-team, and user testing.
https://www.nist.gov/publications/aria-evaluation-planning-manual-elements-aria-style-ai-evaluations

DoW developmental test and evaluation guidance - intended operational conditions and evidence for decision makers.
https://www.cto.mil/dtea/te_aies/

INTERPRETATION BOUNDARY

This is a sales and delivery example, not a completed customer engagement. The numerical results come from generated polygons, not trained-model inference. No external imagery dataset was used, no human operational evaluation was performed, and no accreditation or authorization is claimed.

Prepared for Geo Owl as an illustrative package dated 18 November 2026. Source code and data documentation in the bundle specify how to reproduce the synthetic result.