Why purpose-built AI outperforms generalist models for property intelligence

A benchmark study by Nearmap,
a global property intelligence provider

Nearmap

Abstract

The rapid growth of LLMs and their increasing functionality is creating questions about their suitability for tasks previously handled by purpose-built AI systems. This paper presents the findings of a controlled benchmark study comparing Nearmap AI Gen 6 (a purpose-built aerial imagery detection system) against seven generalist models across two providers: four from Anthropic (Claude Fable 5, Opus 4.8, 4.6 and 4.5) and three from Google (Gemini 3.5 Flash, 3.1 Pro and 2.5 Pro). The paper features the headline results from the best performing model from each provider alongside Fable 5 across four representative property-intelligence tasks: three residential roof tasks — roof area, roof count and roof condition — plus swimming-pool identification.

The study evaluates accuracy (including failure-mode distribution), cost, and operational throughput across a dataset of 2,500 hand-labelled US residential properties. Fable 5, Anthropic's most capable model at time of testing, was evaluated during its brief 72-hour availability window before being suspended under a US government directive — making it the most current data available on that model's property intelligence performance. Findings indicate a material performance gap across all three dimensions, with implications for organisations making AI procurement and workflow decisions in property intelligence contexts.

Nearmap

1. Introduction

1.1 Background

The emergence of capable general-purpose AI models, including large language models (LLMs) and multimodal foundation models, has prompted organizations across industries to reassess their AI strategies.

Global data centre capital expenditure surged 57% in 2025 as AI deployments accelerated, with full-year 2026 capex forecast to surpass $1 trillion*. The scale of that investment has sharpened a practical question for every industry deploying AI at scale.

In property intelligence, where aerial imagery is analyzed at portfolio scale to extract actionable feature data, this question has particular relevance: can foundation models trained on broad datasets perform equivalently to models trained specifically for geospatial and property detection tasks?

Global data centre capital expenditure surged

57%

in 2025 as AI deployments accelerated

1.2 Scope and purpose

This paper presents the methodology and findings of a comparative benchmark study originally conducted by Nearmap in March 2026, updated in June 2026. The study does not seek to evaluate foundation models broadly, but to assess their performance on a set of four defined, representative property-intelligence tasks under real-world conditions, and to examine the structural factors that contribute to observed performance differences.

Full-year 2026 capex forecast to surpass

$1 trillion

1.3 Structure of this paper

Section 2 describes the methodology. Section 3 presents the results. Section 4 analyses the contributing factors. Section 5 sets out limitations. Section 6 considers implications across insurance, government and AECO. Section 7 offers a buyer's framework for evaluating property AI.

* https://www.delloro.com/market-research/data-center-infrastructure/data-center-capex/

2. Methodology

2.1 Task selection

Swimming pool detection was selected as one of our benchmark tasks. It is a well-defined property intelligence use case employed across insurance underwriting, government asset compliance, and property assessment. It is also the most straightforward: simply classifying the presence of something in an image, rather than attempting spatial reasoning. It offers a clear binary output (pool present or absent) and a well-validated ground truth label set. Its visual complexity (varying light conditions, shadow, occlusion, surface material diversity) makes it a meaningful proxy for the broader challenge of property feature detection from aerial imagery.

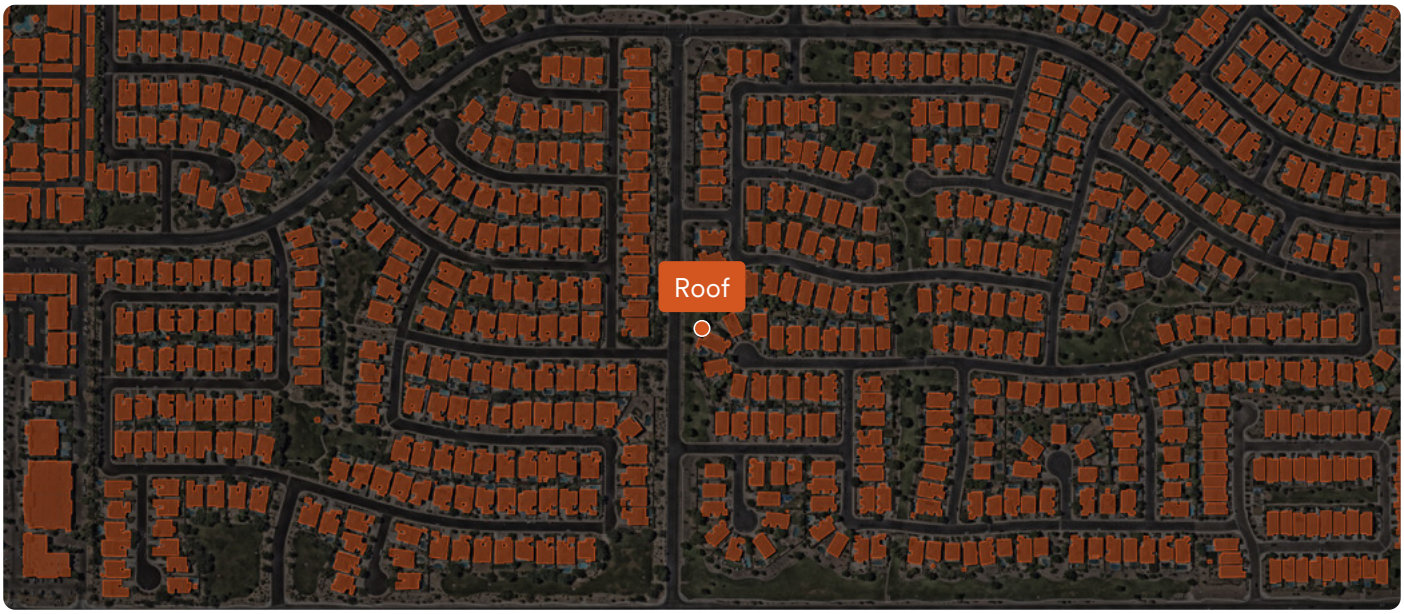
The three additional tasks, around roofs, ratchet the complexity up into a new domain. Gen AI models do not natively do well at quantitative reasoning: a roof area, count or condition level can be requested, but the results are exactly that — an estimate. This type of reasoning is critical in property intelligence, and is best achieved by a combination of AI and spatial mathematics written as traditional software. The four tasks were designed to encapsulate a range of realistic, valuable property intelligence and put these models through their paces.

2.2 Dataset

2,500 residential properties across the United States, with 158 confirmed swimming pools, were hand-labelled and reviewed by human experts with imagery sourced from Nearmap aerial captures. A pool prevalence rate of approximately 6.3% reflects realistic residential portfolio composition. Artificially balanced datasets tend to inflate model performance scores; testing against the real-world distribution produces a more accurate picture of production behaviour.

The set was almost exclusively individual residential homes, which makes the task much easier for the generalist models: houses fall within a fairly tight size range, and 97% of properties have between zero and three roofs. We also gave the models every reasonable advantage — they received not only our official ontology definitions, but the same imagery metadata on scale (the real-world dimensions of each image) to assist their reasoning. These results therefore likely flatter the LLMs relative to harder multi-structure or commercial properties.

Nearmap



2.3 Models evaluated

Nearmap AI Gen 6: a purpose-built deep learning model trained on proprietary Nearmap aerial imagery for property feature detection.

Claude Fable 5: Anthropic's newest flagship model, evaluated under the same protocol during its 72-hour availability window in June 2026. Fable 5 was subsequently suspended under a US government export control directive. Results are included as the most current available data on this model's performance on property intelligence tasks.

Gemini 3.1 Pro: evaluated using high-resolution Nearmap imagery as input, with maximum thinking mode enabled and a detailed natural-language definition of a swimming pool provided via prompt.
These conditions were selected to give the model the best reasonable opportunity to perform. The intent was a fair test, not a constrained one.

Claude Opus 4.8: evaluated under identical conditions to Gemini 3.1 Pro — same imagery, same prompt structure, same JSON output schema. Claude Opus 4.8 is the current production Claude model available via API.

Claude Opus 4.5 and 4.6 and Gemini 2.5 Pro and 3.5 Flash were also evaluated under the same protocol. Version-to-version comparisons from this extended test set are referenced in Section 3.5.

2.4 Evaluation metrics

Performance on pool detection was assessed using the F1 score, the harmonic mean of precision (the proportion of positive detections that are correct) and recall (the proportion of actual positives detected), because both error types carry real-world operational cost.

The roof tasks use the metric appropriate to each: roof count by exact-match accuracy (the share of properties given exactly the right number of structures); roof area by mean absolute error in square metres; and roof condition by mean absolute error on the 0–100 Roof Spotlight Index. Two operational measures complete the picture: speed (latency) — the mean wall-clock time to return a complete property result across all four tasks, each sent as a separate API call — and throughput, properties processed per second at scale.

2.5 Throughput measurement

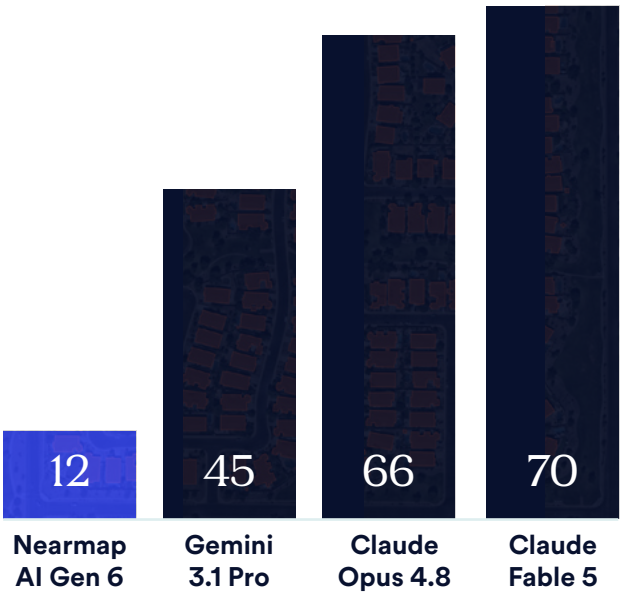
In addition to accuracy, processing throughput (properties per second) was measured for each model to assess operational viability at portfolio scale.

3. Results

3.1 Accuracy

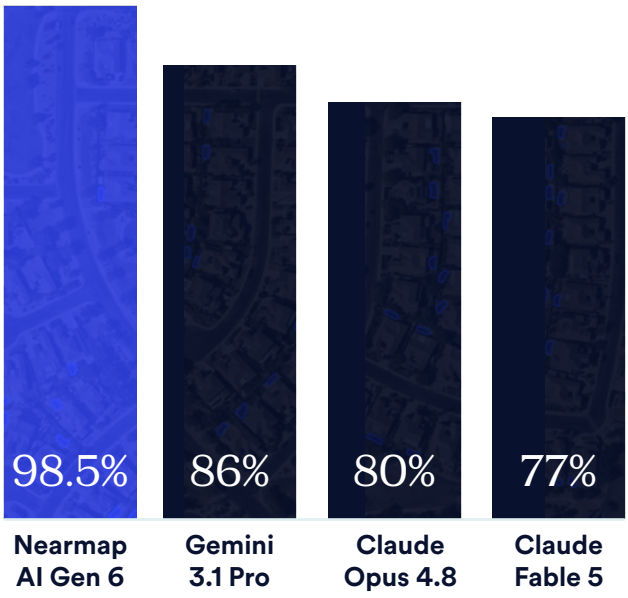
	Nearmap AI Gen 6	Claude Opus 4.8	Claude Fable 5	Gemini 3.1 Pro
Roof area — mean abs. error	**12 sqm**	66 sqm	70 sqm	45 sqm
Roof count — exact match	**80%**	62%	64%	70%
Roof condition — RSI mean abs. error	**reference**	18 pts	15 pts	12 pts
Pool presence — F1	**98.5%**	80%	77%	86%

Roof-area error (sqm)



On the roof tasks — the ones that matter most to underwriting and valuation — the gap was just as wide: roof-area error of 12 sqm against 45–70 sqm for the LLMs, and roof-count exact-match of 80% against 62–70%. Notably, both providers' newest models were worse at roof-area estimation than their predecessors.

Pool detection — F1



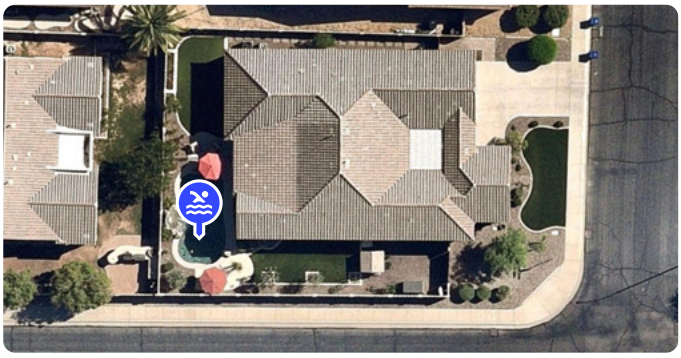
Nearmap AI Gen 6 led on every task. On pool detection it achieved an F1 score of 98.5%; Gemini 3.1 Pro was the strongest generalist at 86%, Claude Opus 4.8 scored 80%, and Claude Fable 5 — Anthropic's newest and most capable model — scored 77%, with high recall (96%) offset by significantly lower precision (64%). A more powerful model did not produce a more accurate result.

Nearmap

3.2 False negative analysis

We observed that Gemini 3.1 Pro and both Claude models produced a higher rate of false negatives (instances where pools were present but not detected). These errors concentrated in cases involving:

Deep shadow cast across pool surfaces.



Partial occlusion by overhanging vegetation or adjacent roof structures.

Non-standard pool finishes or surface treatments.

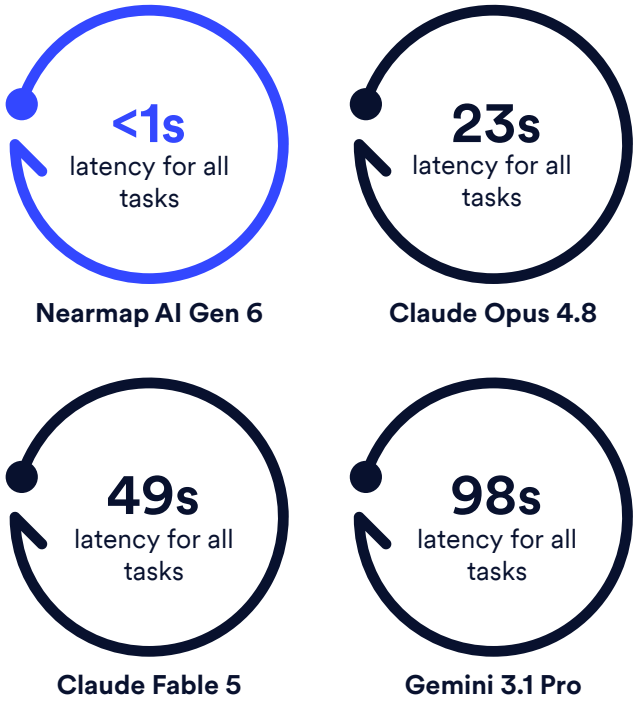
Nearmap AI Gen 6 correctly identified pools across these edge cases. This is consistent with a model calibrated on millions of labelled examples drawn from the same imaging conditions rather than one inferring pool characteristics from general visual training data.

3.3 False positive analysis

It appeared in our results that Gemini 3.1 Pro and both Claude models also produced false positives — detections on features that were not pools, in cases involving light-coloured paving, certain pool covers, and specific lawn textures. Claude Fable 5 showed the highest false positive rate of the models tested, consistent with its low precision score of 64%. Nearmap AI Gen 6 did not flag these features as pools. The distinction in surface texture at 5–7.5cm resolution is a learned characteristic in purpose-built models trained on aerial imagery, not reliably inferred from general visual or language training data.

3.4 Throughput

Nearmap AI Gen 6 operates at sub-second inference, processing more than 1,000 properties per second. Claude Opus 4.8 averaged 23 seconds per property. Claude Fable 5, despite being the newer model, averaged 49 seconds — more than double — reflecting the additional cost of its default reasoning mode. Gemini 3.1 Pro, the most accurate generalist model tested, was also the slowest at 98 seconds per property. All generalist models are subject to rate-limiting at scale. At a portfolio of 100,000 properties, these throughput differences represent an operational gap of several orders of magnitude.



Nearmap

3.5 Output consistency

Commercial vision language models are not architected to guarantee identical outputs for identical inputs. The same image, submitted with the same prompt, can return a different result on each pass. For repeatable property intelligence workflows, like underwriting, compliance, and planning, an output that changes depending on when it is run cannot function as a reliable data source. Nearmap AI Gen 6 returns deterministic outputs. Same input, same result, every time.

	Pool	Roof count
Nearmap AI Gen 6	**0%**	**0%**
Claude Opus 4.8	0.6%	7.6%
Claude Fable 5	3.0%	18.1%
Gemini 3.1 Pro	3.0%	20.8%

3.6 Cost

Across the four tasks, cost per property ran from 2–5¢ (Gemini) to ~6¢ (Opus 4.8) to ~18¢ for Fable 5 — roughly 3× the cost of Opus for no accuracy gain. Over 150M US properties that is ~$3M (cheapest Gemini), ~$9.5M (Opus) and ~$28M (Fable 5) per pass. Nearmap AI produces 130+ attributes in a single inference pass, included in the subscription.

Cost of one pass over 150m US properties

3.7 How LLMs “measure”: the quantisation tell

LLMs generate plausible numbers rather than measuring geometry: 98–100% of Claude’s roof-area estimates were divisible by 5 sqm, and Opus 4.5 returned exactly “185 sqm” for roughly a quarter of all properties. For quantitative property analysis this is a structural limitation, not a tuning issue.

Model updates — and suspensions — introduce instability that purpose-built AI does not. General LLMs are updated for general capability, not specific property detection tasks. In June 2026, Claude Fable 5 was made available and suspended within 72 hours under a US government directive. A version update may improve or degrade performance on your use case without appearing in any relevant changelog — and a model can disappear entirely. **For auditable workflows, that risk is not theoretical.**

Nearmap

4. Analysis: Factors contributing to performance

4.1 Training data specificity

Nearmap AI Gen 6 is trained exclusively on aerial imagery captured at consistent resolution (5–7.5cm GSD), consistent geometry, and consistent sensor characteristics, all acquired by Nearmap and our proprietary camera system. This means the model's feature representations are calibrated to the precise visual conditions under which it operates.

Foundation models are trained to generalise across an enormous range of inputs. That generalisation is a core design property, not a deficiency. But it means that the visual signatures distinguishing a swimming pool from a pool cover, or a shadow-occluded pool from a dark paved surface, have not been systematically calibrated from millions of examples in the same imaging context.

4.2 Task scope and model stability

Nearmap AI Gen 6 operates within a tightly defined task scope. Its release cycle is governed by internal benchmarks against property intelligence tasks before any update ships. Performance changes are therefore tracked against the specific task domain.

Foundation models are designed to optimise across a broad range of capabilities simultaneously. That breadth is the point; but it means no single task, including property feature detection, anchors the development cycle. For organizations that need AI outputs to be stable, auditable, and consistent over time, that structural difference matters more than any individual benchmark result.

4.3 Resolution and geometry consistency

Aerial imagery at 5–7.5cm resolution presents a specific visual vocabulary (in texture, shadow geometry, surface spectral response) that differs substantially from ground-level photography or lower-resolution satellite imagery. Purpose-built models trained on this imagery learn representations specific to that vantage point. Foundation models trained across varied image types do not have the same depth of calibration.

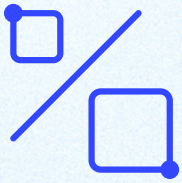
5-7.5cm
ground sample distance (GSD)

Nearmap

5. Limitations: Three boundaries are worth acknowledging

Task scope

The study covers four defined property tasks (three roof, one pool). The findings should not be extrapolated to tasks involving greater visual complexity, multi-class outputs, or feature types not represented in this study.



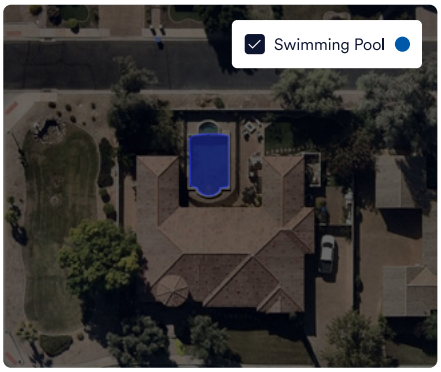
Model selection

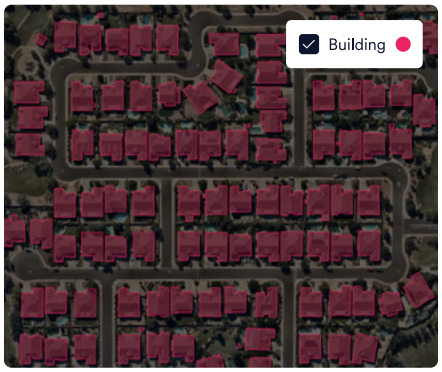
Seven generalist models across two providers were tested in total. Four are featured in this paper: Gemini 3.1 Pro, Claude Opus 4.8, and Claude Fable 5 as the headline generalist results, alongside Nearmap AI Gen 6. Other foundation models or configurations may produce different results. Notably, few-shot learning techniques applied to open-source foundation models represent an approach that may achieve closer parity on specific, well-defined detection tasks. This study does not evaluate that category of model. Claude Fable 5 results reflect testing conducted during its 72-hour availability window prior to suspension and represent the most current data available on that model.

Dataset composition

The 2,500-property dataset comprised US residential properties drawn from Nearmap aerial captures. Results may not generalise to commercial properties, other land use types, or imagery sourced from providers with different resolution, sensor characteristics, or capture geometry.

6. Implications by sector





6.1 Insurance

Roof condition and roof area are the primary drivers of underwriting and insurance valuation; pools are a secondary signal. Both affect risk pricing. A false negative (asset class not recorded) produces a policy priced without complete property information. A false negative (a feature misidentified) may trigger unnecessary inspection workflows or inaccurate renewal quotes. At portfolio scale, error rates that appear low in isolation accumulate into material underwriting exposure and operational cost.

The findings in Section 3 make this concrete. A material gap in F1 score between purpose-built and generalist AI, concentrated in false negatives on shadow-occluded or non-standard pools, translates directly into policies written without complete property data. Before embedding any AI model into underwriting or renewal workflows, require F1 scores, precision and recall figures separately, and error distribution data against a validated ground truth dataset. The evaluation framework in Section 6 provides the questions to ask.

6.2 Government

Local government agencies using aerial intelligence for asset compliance, infrastructure planning, or rate assessment require outputs that are accurate, auditable, and consistent over time. Model instability (whether from a foundation model version change or an edge-case failure on atypical properties) introduces risk in compliance workflows and planning decisions.

The consistency finding in Section 3.5 is particularly relevant here. A model that does not guarantee identical outputs for identical inputs cannot serve as a reliable basis for compliance records or rate assessment decisions. Agencies embedding AI into auditable workflows should require documented evidence of output determinism and clear communication of how model updates are tested against their specific use case before deployment.

6.3 Architecture, Engineering, Construction and Operations (AECO)

AECO workflows rely on accurate feature detection for site analysis, construction change detection, and planning documentation. Throughput at scale, that is the ability to process large property datasets without manual triage, is operationally significant. Models that require extended processing time or produce inconsistent results on atypical sites introduce friction into time-sensitive workflows.

The throughput finding in Section 3.4 sets the operational baseline. At 100,000 properties, the difference between sub-second inference and 7-second processing per property is not marginal, it determines whether AI-assisted workflows are viable without significant manual intervention. Evaluate processing rates and rate-limiting thresholds at portfolio scale, not sample scale, before committing to a model for production use.

7. Questions to ask when evaluating property AI

1. Accuracy methodology

Can you provide F1 scores against a validated ground truth dataset specific to the feature class we need to detect? What are the precision and recall figures separately?

In this study: 98.5% (Nearmap AI Gen 6) vs. 87% (Gemini 3.1 Pro), 80% (Claude Opus 4.8), and 77% (Claude Fable 5) on the same dataset. Section 3.1.

2. Training data provenance

Was your training data purpose-captured for aerial property intelligence, or assembled from general image sources? How consistent is the imaging geometry, resolution, and sensor characteristics across your training set?

Purpose-captured imagery at consistent resolution drove the edge case performance gap. Section 4.1.

3. Throughput at operational scale

What are your demonstrated processing rates at portfolio scale, not sample scale? At what point do throughput constraints or rate limits become operationally prohibitive?

1,000+ properties per second (Nearmap AI Gen 6) vs. 23 seconds per property (Claude Opus 4.8), 49 seconds (Claude Fable 5), and 98 seconds (Gemini 3.1 Pro). At 100,000 properties, the difference is orders of magnitude. Section 3.4.

4. Model stability and versioning

How are model updates managed and tested against property intelligence benchmarks before release? How are performance changes communicated, and what recourse exists if an update degrades performance on our specific use case?

Foundation model updates may degrade task-specific performance without appearing in a relevant changelog. Section 4.2.

5. Edge case performance

What is the model's performance on edge cases relevant to our dataset, shadow, occlusion, surface variation, and regional imagery characteristics? Can you provide error distribution data, not just aggregate scores?

False negatives concentrated in shadow, occlusion, and non-standard finishes. Section 3.2.

6. Output consistency

Does the model guarantee identical outputs for identical inputs across repeated runs?

Commercial VLMs do not. Same prompt, same image, different result. Section 3.5.

7. Financial implications

What are the cost implications of model updates, new feature attributes, or vendor pricing changes over time? How does the total cost of ownership compare when accuracy errors (missed features, false detections) are factored into downstream operational costs?

At portfolio scale, low aggregate error rates accumulate into material underwriting exposure and operational cost. Section 5.1.

8. Regulatory and governance

What audit trail does the model provide for its outputs? For general-purpose foundation models, how is accountability established when a detection error affects a regulated decision? What documentation exists to support compliance or dispute resolution?

Output consistency and model versioning directly affect auditability. Section 3.5, 4.2.

Conclusion

This study demonstrates a material performance gap between purpose-built AI and general-purpose foundation models on four well-defined property-intelligence tasks , across accuracy, cost and operational throughput. The gap is attributable to structural differences unlikely to be resolved by improvements in general capability alone. Even Anthropic's most powerful model, in the brief window it was public, did not change the conclusion. Foundation models remain a significant capability where flexible reasoning is required — and Nearmap uses them extensively for exactly that.

Connect with our AI experts to see how purpose built AI works for your org.

Connect Now

Methodology Note

Original March 2026 results are based on Nearmap internal testing, using an aerial imagery dataset of 2,500 residential properties from the USA, hand-labelled and reviewed by human experts, containing 158 swimming pools. All models were tested against the same dataset under the conditions described. The June 12 2026 update to this study added Claude Opus 4.8, Claude Fable 5, and Gemini 3.5 Flash under the identical protocol across all four tasks. Claude Fable 5 was tested during its 72-hour availability window prior to suspension under US government export control directive EO-2026-AI-7. Results reflect API defaults at time of testing, including default reasoning mode for models that reason by default. Given the pace of model releases and policy changes in this space, readers are encouraged to conduct independent evaluation against current model versions before making procurement decisions.

Nearmap