

Decision Models versus Reasoning Models: An Exploratory Evaluation of Perplexity’s Decisions API

Praveen Sirimanne

Independent article · October 2026

Status. This is an independent exploratory study, carried out in a personal capacity. It is not affiliated with or endorsed by any organisation. It has **not been peer reviewed**.

Use of AI. This article was prepared with the assistance of Claude (Anthropic), which supported the literature search, experiment code, data collection and drafting. The author designed the study, verified the ground truth and results, and takes full responsibility for the content. Because Claude is also one of the systems evaluated, all raw outputs are published so that results can be independently checked.

Abstract

Background. In September and October 2026, three providers released a new class of AI system: the *decision model*. Unlike conventional large language models (LLMs), a decision model does not generate text. It evaluates content against predefined answer options and returns a probability for each. **Aim.** This study compares Perplexity’s Decisions API with a conventional reasoning model, Claude Opus 5.5, in terms of accuracy, consistency, cost and speed. **Method.** The test task was *attribution*: distinguishing what a document proposes from what it merely cites from other sources, a task AI systems are known to struggle with. A recent public policy document was used as a real-world example, with 18 items whose status was verified against independent sources. Each system was run three times. **Results.** Asked an open-ended question, the reasoning model presented 11–12 of 14 inherited policies as new. On a closed list, the two systems achieved comparable accuracy (79–96% and 80%). The decision model was fully consistent across runs, approximately ten times cheaper and five times faster, and its confidence scores identified all of its own errors. **Conclusion.** Decision models do not replace reasoning models, but they offer a reliable, low-cost and self-flagging alternative for repetitive classification tasks. They perform best alongside a reasoning model, with human review of low-confidence results.

Keywords: decision models; large language models; classification; calibration; document analysis; AI evaluation

1 Introduction

1.1 Background

Organisations increasingly use LLMs to analyse policy, regulatory and compliance documents. These documents often discuss measures that originate elsewhere, such as earlier plans, legislation, national strategies or other governments’ schemes. A recurring failure is that an AI system asked to identify a document’s *new* proposals also lists measures the document only cites. The output is fluent and plausible, but wrong, and the reader has no indication of which items to doubt.

1.2 Decision models

Conventional LLMs generate text. When used for classification, their answers must be parsed from prose or structured output, and they give no reliable measure of confidence. Decision models take a different approach (Table 1).

Table 1: Conventional LLMs and decision models.

	Conventional LLM	Decision model
Input	A prompt	Content, plus questions with predefined answer options
Output	Generated text	A probability for each option
Explanation	Can describe its reasoning	Gives none
Typical uses	Drafting, extraction, analysis	Classification, routing, verification, scoring

In effect, a decision model is a fast, inexpensive system for answering multiple-choice questions. It cannot generate the options itself; these must be supplied by the user.

1.3 Emergence of the category

Three decision models were released within a few weeks of each other (Table 2). Perplexity’s Decisions API was selected because it was publicly available and fully documented at the time of the study [3, 6]. The

experiment was conducted on 2 October 2026, one day after its release.

Table 2: Decision models released in 2026.

Release	Key characteristics
Jev TypeSafe AI Sep. 2026	Described as the first “System One” decision model; hosted, early access [9]
Decisions API OpenAI 29 Sep. 2026	Announced at DevDay; built on OpenAI’s Luna model; limited preview [8]
Decisions API Perplexity 1 Oct. 2026	Model <code>pplx-decider-v1-27b</code> , fine-tuned from Qwen3.8-27B; open weights (Apache 2.0); USD 0.04 per million input tokens, output free [5, 6]

1.4 Research question

How does a decision model compare with a conventional reasoning model in accuracy, consistency, cost and speed, on a realistic task where AI systems are known to make errors?

The policy document serves only as a test case. The subject of evaluation is the AI systems, not the document or its policies.

2 Methodology

2.1 Test document

- **Document:** *Scotland’s Climate Change Plan: 2026–2040*, final main volume (44 pages; annexes excluded) [7].
- **Rationale:** the document is recent, publicly available and refers frequently to other strategies, legislation and earlier plans, which makes it well suited to testing attribution. Any comparable document could be used.
- **Preparation:** the PDF was converted to plain text (approx. 21,000 tokens), excluding figures. Both systems received identical text.

2.2 Ground truth

To avoid selection bias by the author, every item’s label was taken from an independent or originating source.

- **New items (4):** measures described as new by the Scottish Parliament Information Centre (SPICe), the Parliament’s non-partisan research service [2].
- **Cited items (14):** measures that appear in the document but originate elsewhere, each traced to its source (legislation, a national strategy, an earlier plan or UK Government regulation).
- **Verification:** each item carries a verbatim quotation from the document, with page number, and one from the supporting source. All quotations were automatically checked against saved copies of the sources.

- **Version control:** SPICe’s analysis covered the draft plan, so each item was checked against the government’s published record of changes in the final version [7].

Table 3: Illustrative ground-truth items.

ID	Item	Label	Source
N1	Heat in Buildings Strategy and Delivery Plan	New	SPICe: “a new commitment”
T2	Zero Emission Vehicle Mandate	Cited	UK Government, VETS Order 2023
T11	Product Stewardship Plan	Cited	Circular Economy and Waste Route Map (2024)
T13	Halt biodiversity loss by 2030	Cited	Scottish Biodiversity Strategy (2024)

2.3 Experimental conditions

Three conditions were tested (Table 4).

Table 4: Experimental conditions.

Condition	Task
A Claude Opus 5.5, open-ended	“List the newly proposed policies and priorities in this document.”
B Claude Opus 5.5, closed list	Classify each of the 18 items as <i>new</i> or <i>cited</i> , returning structured output
C Perplexity Decisions API	Classify the same 18 items in a single request, with identical wording and option definitions

Configuration. Claude was run with adaptive reasoning at “high” effort [1]. The Decisions API used one **choice** question per item, implemented exactly as specified in the provider’s documentation [4, 6].

Controls.

- Identical document text, item wording and option definitions for both systems.
- Neutral item identifiers, so labels could not be inferred.
- A fixed item order for all runs.
- Three runs per condition.

2.4 Evaluation measures

- **Accuracy:** reported separately for new and cited items, and as *balanced accuracy* (the mean of the two). This corrects for class imbalance; a system labelling everything as cited would otherwise score 78%.
- **Consistency:** the number of items receiving the same label in all three runs.
- **Cost:** token usage multiplied by each provider’s published list price.
- **Latency:** time per request.

Table 5: Summary of results across three runs per condition.

Measure	A: Reasoning, open-ended	B: Reasoning, closed list	C: Decision model
Cited items misclassified as new (of 14)	12 / 11 / 12	1 / 6 / 6	2 / 2 / 2
New items correctly identified (of 4)	4 / 4 / 4	4 / 4 / 4	3 / 3 / 3
Balanced accuracy	n/a	96% / 79% / 79%	80% / 80% / 80%
Items labelled consistently (of 18)	n/a	13	18
Cost per run (USD)	0.20	0.16	0.015
Latency per run	34 s	14 s	2.7 s
Confidence measure provided	No	No	Yes

- **Calibration** (decision model only): whether lower confidence corresponded to lower accuracy.

3 Results

3.1 Summary

Table 5 summarises the results. Values shown as “x / y / z” refer to runs 1, 2 and 3.

3.2 Open-ended extraction

Asked to list the document’s new proposals, the reasoning model returned approximately 60 items per run, including 11–12 of the 14 cited items. In several cases its own wording showed that an item was not new. The following, for example, were listed as newly proposed:

“Continue the Agricultural Reform Route Map...”
“Continue the Scottish Industrial Energy Transformation Fund.”

In one run, the model acknowledged that the document “doesn’t clearly separate brand-new measures from continuations of existing ones”. These results confirm the failure mode that motivated the study.

3.3 Closed-list classification

- **Reasoning model:** achieved the highest single-run accuracy (96%). In runs 2 and 3, however, the same five cited items were reclassified as new, reducing accuracy to 79%. Identical inputs produced different outputs.
- **Decision model:** returned identical labels in all runs, with probabilities consistent to two decimal places (80% balanced accuracy).
- **Common failure:** both systems misclassified the most ambiguous item, a woodland target that the document re-dates but which originates in a 2020 plan.

3.4 Calibration

All three of the decision model’s errors fell below a confidence of 0.70 (Table 6). A rule referring such items for human review would have caught every error while requiring review of only 4 of 18 items. The reasoning model provides no comparable signal.

Table 6: Decision model calibration (run 1).

Confidence in chosen answer	Items	Correct
≥ 0.80	6	6 (100%)
0.70–0.80	8	8 (100%)
< 0.70	4	1 (25%)

3.5 Cost structure

The decision model’s lower cost results from three factors, partly offsetting one another (Table 7). Because its cost scales with the number of questions asked of a document, the advantage narrows for long documents with many questions.

Table 7: Factors determining relative cost.

Factor	Effect
Input price: USD 0.04 vs 4.00 per million tokens	Approx. 100× lower
Output: 18 free tokens vs 1,238 billed (mostly reasoning)	No charge for reasoning or text
Document billed once per question (380,449 vs 34,149 input tokens)	Offsets part of the saving
Net effect	Approx. 10× lower

4 Discussion

4.1 Comparative assessment

Table 8 summarises the strengths and limitations of each approach.

Table 8: Strengths and limitations.

	Reasoning model (closed list)	Decision model
Strengths	Highest peak accuracy; explains answers; supports any task, including extraction	Fully consistent; confidence identifies errors; approx. 10× cheaper and 5× faster; machine-readable output
Limitations	Inconsistent between runs; no confidence measure; higher cost and latency	Slightly lower average accuracy; cannot extract or propose options; no explanation; per-question billing

4.2 Suitability by use case

Table 9 indicates which approach is better suited to common document-analysis tasks.

Table 9: Recommended approach by use case.

Use case	Recommended
Identifying candidate items within a document	Reasoning model
One-off analysis with expert review	Reasoning model
High-volume automated classification	Decision model
Repeatable outcomes (e.g. audit, compliance)	Decision model
Prioritising items for human review	Decision model

4.3 Recommended architecture

The findings point to complementary roles rather than a single preferred system:

1. A **reasoning model** extracts candidate items from the document.
2. A **decision model** classifies each item and attaches a confidence score.
3. A **human reviewer** examines items below a defined confidence threshold.

This approach combines the flexibility of reasoning models with the consistency, speed and cost efficiency of decision models.

5 Limitations

- **Scale:** one document, 18 items and three runs. The study is exploratory and should not be read as a benchmark; a five-point accuracy difference corresponds to roughly one item.
- **Class imbalance:** 4 new items against 14 cited items.
- **Ambiguous cases:** two items are open to reasonable dispute, notably targets the document revises but did not originate.
- **Source timing:** the independent analysis covered the draft document; this was mitigated by checking against the published record of changes.

- **Configuration:** a single reasoning model and setting were tested. Other models, prompts or settings may perform differently.
- **Maturity:** the decision model was tested one day after release. Its behaviour and pricing may change.

6 Conclusion

Decision models are a credible new tool, but not a replacement for reasoning models. In this exploratory study, Perplexity’s Decisions API achieved accuracy comparable to a leading reasoning model, within approximately one item on average. It was also fully consistent, about ten times cheaper and five times faster, and it reliably flagged its own errors through its confidence scores. Its limitations are equally clear: it cannot locate information independently and offers no explanation for its answers. For document analysis in policy and compliance settings, the strongest approach is to combine the two, assigning each the role it performs best.

Data and code availability

All materials required to verify or reproduce this study are published alongside this article: the ground-truth dataset with quotations, page references and source links; the experiment, scoring and verification scripts; and the raw outputs of all nine runs, including request identifiers. Total experimental cost was approximately USD 1.13.

References

- [1] Anthropic. Claude API documentation and pricing. <https://platform.claude.com/docs>, 2026. Accessed 2 October 2026.
- [2] Niall Kerr, Anna Brand, Alan Rehfish, Alexa Morrison, Rob Watts, and Kate Berry. Climate change plan: Policies, proposals and sector summaries. SPICe Spotlight, Scottish Parliament Information Centre, <https://spice-spotlight.scot/2025/12/01/climate-change-plan-policies-proposals-and-sector-summaries/>, December 2025. 1 December 2025. Accessed 2 October 2026.
- [3] Perplexity. Decisions API: Answer questions (API reference). <https://docs.perplexity.ai/api-reference/decisions-post>, 2026. Accessed 2 October 2026.
- [4] Perplexity. Triage support tickets with the decisions API. <https://docs.perplexity.ai/docs/cookbook/examples/decisions-api-ticket-triage/README>, 2026. Accessed 2 October 2026.
- [5] Perplexity. pplx-decider-v1-27b [model card]. Hugging Face, <https://huggingface.co/perplexity-ai/pplx-decider-v1-27b>, 2026. Accessed 2 October 2026.

- [6] Perplexity. Decisions API: Quickstart. <https://docs.perplexity.ai/docs/decisions/quickstart>, 2026. Accessed 2 October 2026.
- [7] Scottish Government. Scotland’s climate change plan: 2026–2040. <https://www.gov.scot/publications/scotlands-climate-change-plan-2026-2040-annexes/documents/>, March 2026. Main volume and Annex 4: Consultation and Scrutiny. Accessed 2 October 2026.
- [8] TechCrunch. OpenAI’s Jev clone could help the frontier lab stop its swarming agents. <https://techcrunch.com/2026/09/30/openais-jev-clone-could-help-the-frontier-lab-stop-its-swarming-agents/>, September 2026. 30 September 2026. Accessed 2 October 2026.
- [9] YourStory. Introducing Jev: Ex-OpenAI researchers build a new kind of AI model. <https://yourstory.com/ai-story/what-is-jev-typesafe-ai-decision-model>, 2026. Accessed 2 October 2026.