Error Detection by Language Models Depends More Heavily on the Vendor than the Prompt

Oliver Burch

School of Mathematical and Physical Sciences, University of Sheffield
Supervisors: Jonathan Potts and Wei Xing

University of Sheffield

The Question

Given animal movement data, does the quality of a language model's analysis depend on what it is told, or on which model you ask?

Approach

We asked four language models the same question, but gave them different information in three framings:

(A) The data alone, with the column names removed.
(B) The same data, plus the real study system named.
(C) The same data, plus a false claim that they came from roe deer in Britain [3].

We then ran a second test: no data at all, only a varying inventory of real and decoy dataset titles, and a request to design a study hypothesising why fewer elk now migrate. One real cause – supplemental winter feeding – was left off every list.

Models: Claude Sonnet 5, Claude Opus 4.8, Gemini 2.5 Flash, Gemini 3.1 Pro.

72

Responses: three runs of three framings, for each of four models in each of two tests.

The Data

- Some elk leave the Ya Ha Tinda range for the summer and others stay. This has been known for decades [1]. Can a language model find that split in the data?



Figure 1. An elk – the focal species of this project. Photo: Ryan Hagerty, U.S. Fish and Wildlife Service (public domain).



Figure 2. Ya Ha Tinda, on the eastern edge of Banff National Park in the Alberta Rockies. Boundaries and relief: Natural Earth (public domain).

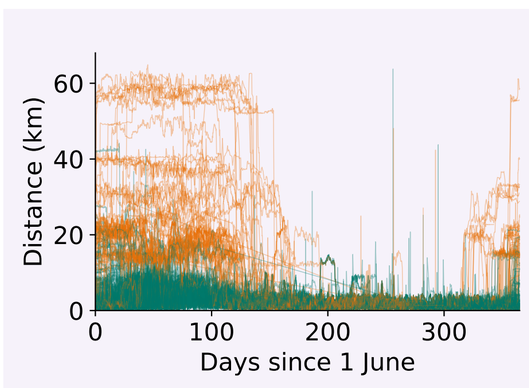


Figure 3. Every elk-year in the herd [2] – one elk followed through one year from 1 June. Orange leaves the valley for the summer; Teal stays near the winter range all year.

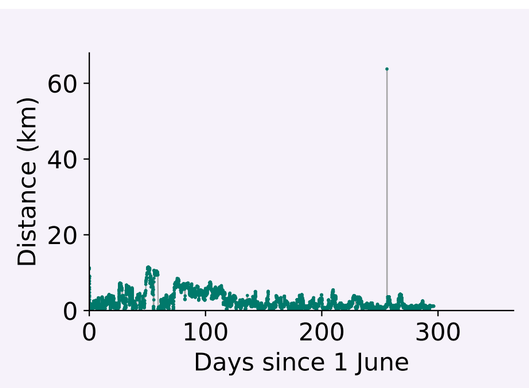


Figure 4. An elk that registered a single poor GPS fix, far from all its others. That distance is impossible in the time recorded. Several other collars showed similar spikes.

Results

(A) Given the data alone

- Nearly every response found that the herd splits into groups.
- 31 of 36 stopped at two groups. The data supports more.
- Only a single model placed the boundary between the most prominent groups correctly.

(B) Given the real study system

- Naming the system did not improve the analysis.
- It made models readier to assert what the data cannot support.

(C) Given a false claim about the species

- No Claude response questioned the false species. Half the Gemini responses did to at least a small extent.
- Yet only the Claude models asked whether extreme readings were real animals or bad fixes.

Across all conditions

- Framing moved scores by a few marks, in no consistent direction.
- The vendor gap held under every framing, and on both tests.
- None of the 36 responses proposed supplemental winter feeding – the one real cause withheld from every inventory.

Where the vendors differ

- Scepticism cannot be categorised as a single trait. A model can doubt the data and accept the framing, or the reverse.

	Claude	Gemini	*p*
Raises measurement error	7 / 12	0 / 12	0.005
Questions species plausibility	0 / 6	3 / 6	0.182

Table 1. Responses meeting each condition, pooled by vendor, with two-tailed Fisher exact tests. These are counts of responses, not marks. Only the measurement-error split is significant.

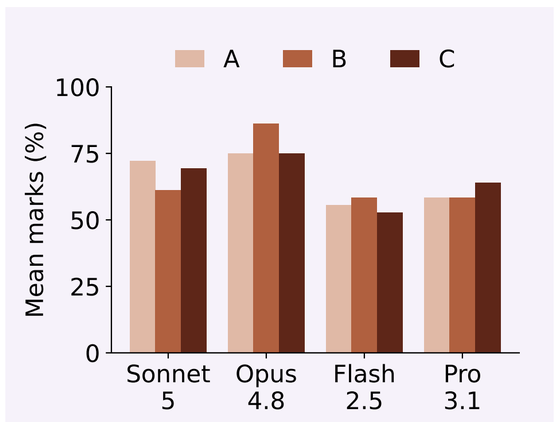


Figure 5. Reading the data by framing – (A) Data alone, (B) System named, (C) False label. The three means move in no consistent direction.

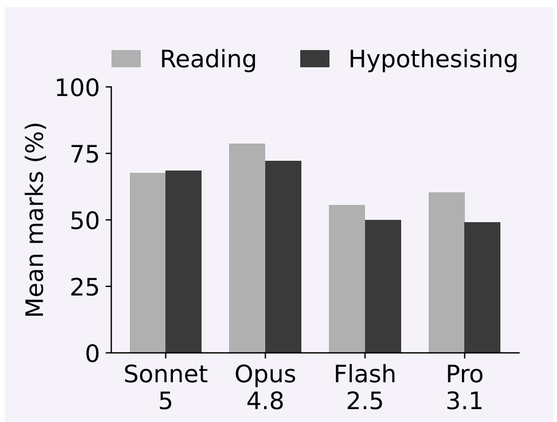


Figure 6. Scores on both tests. The Claude models lead across both.

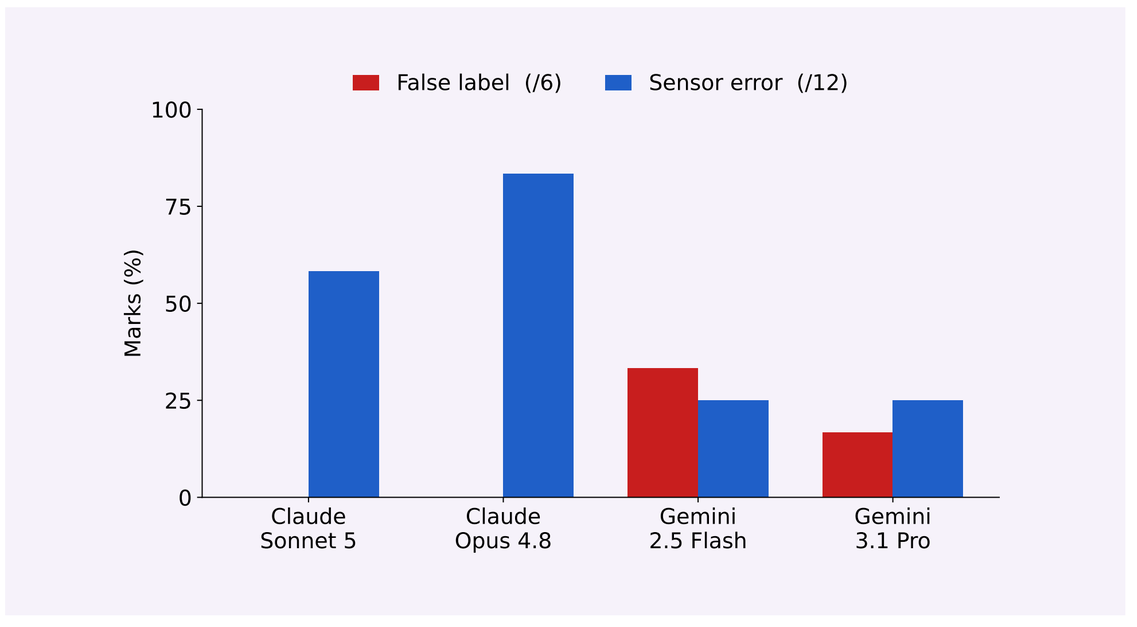


Figure 7. Two doubts, scored separately. **False label:** did the response question the roe deer claim? **Sensor error:** did it treat extreme readings as possible bad fixes? Marks as a percentage of each flag's own maximum; the two flags applied to different numbers of runs.

Conclusions

- Which errors a language model detects depends more on the vendor than on the question's framing.
- The models most careful with the data never questioned a false premise – and the reverse held too.
- The gap between vendors held on both tests – reading the data, and building hypotheses about it.

References

[1] Hebblewhite, M. *et al.* (2006). Is the migratory behavior of montane elk herds in peril? The case of Alberta's Ya Ha Tinda elk herd. *Wildlife Society Bulletin* 34(5), 1280–1294.

[2] Hebblewhite, M. *et al.* (2020). Data from: Ya Ha Tinda elk project, Banff National Park, 2001–2020 (females). Movebank, doi:10.5441/001/1.5g4h5t6c

[3] Mysterud, A. (1999). Seasonal migration pattern and home range of roe deer (*Capreolus capreolus*) in an altitudinal gradient in southern Norway. *Journal of Zoology* 247(4).

A language model's scepticism tracked its vendor more closely than the framing it was given. Judging what it failed to question is still the researcher's job.

Full report: www.olivergtburch.com

OBurch1@sheffield.ac.uk