What can a language model do for a movement ecologist?

Oliver Burch

School of Mathematical and Physical Sciences, University of Sheffield

Supervisors: Jonathan Potts and Wei Xing

OBurch1@sheffield.ac.uk



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?

A herd with a known answer





Ya Ha Tinda, Banff National Park · 234 elk-years, 2001–2020

Migrant/resident split published in **Hebblewhite et al. 2006**

Photo: Ryan Hagerty, USFWS (public domain). Map: Natural Earth (public domain).

The two tests

READING THE DATA

No context

Real system named

False species

NO DATA AT ALL

Dataset titles only

Some of them invented

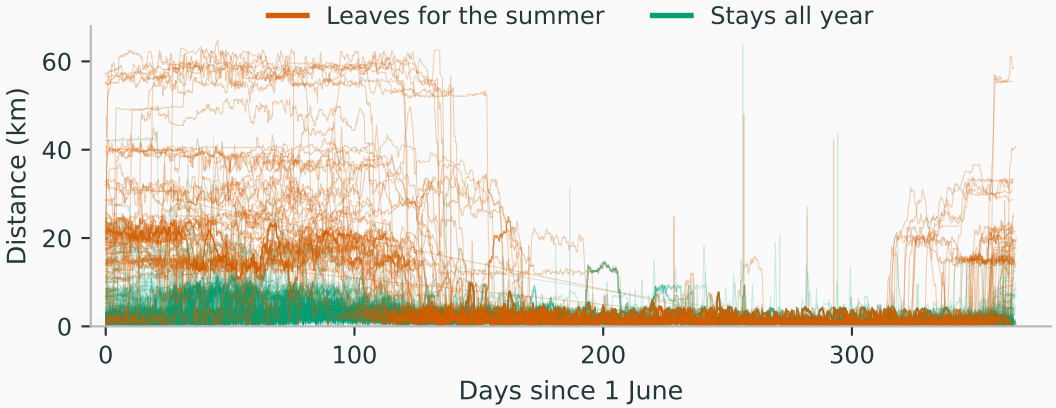
Models: Claude Sonnet 5 . Claude Opus 4.8 . Gemini 2.5 Flash . Gemini 3.1 Pro

Three runs of every condition

One rubric per test, six criteria each marked out of two

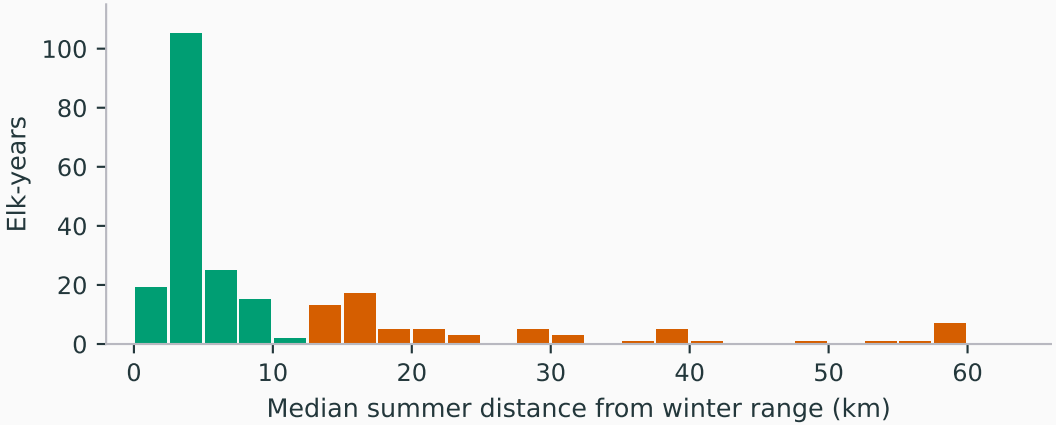
Flags marked independently of the scores

They find the split



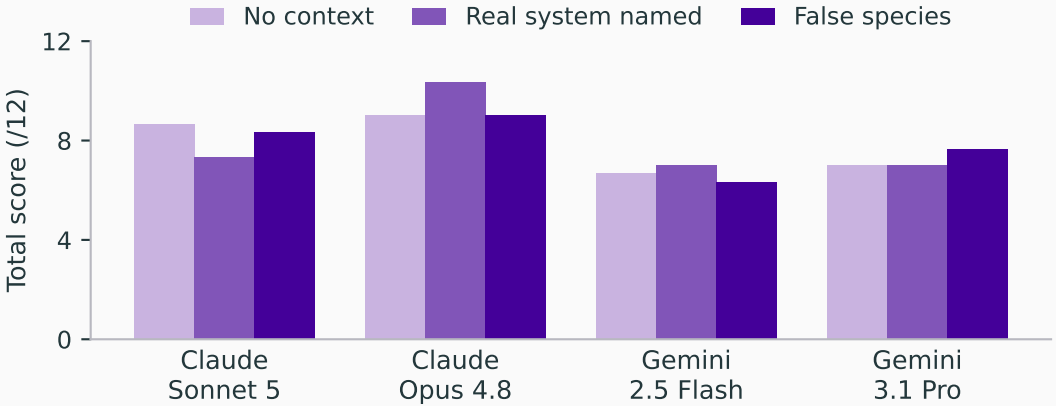
One line per elk-year · Only one response didn't find the structure

Then stop at two groups



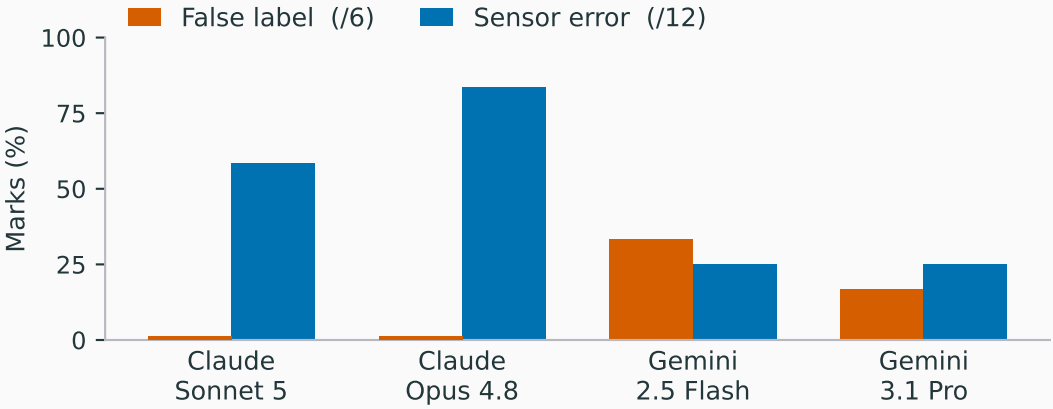
31 of 36 named exactly two groups · **4 of 36** responses placed the boundary correctly

Framing moved scores in no consistent direction



Naming the real system moved Sonnet **down** and Opus **up** · The Geminis barely moved

Scepticism is not one trait



The models most alert to bad readings never questioned the premise

The withheld cause

0 of 36

Responses proposed supplemental winter feeding, the withheld driver

What a usable hypothesis needs

S1 Mechanism named

S2 Direction predicted

S3 Disproof stated

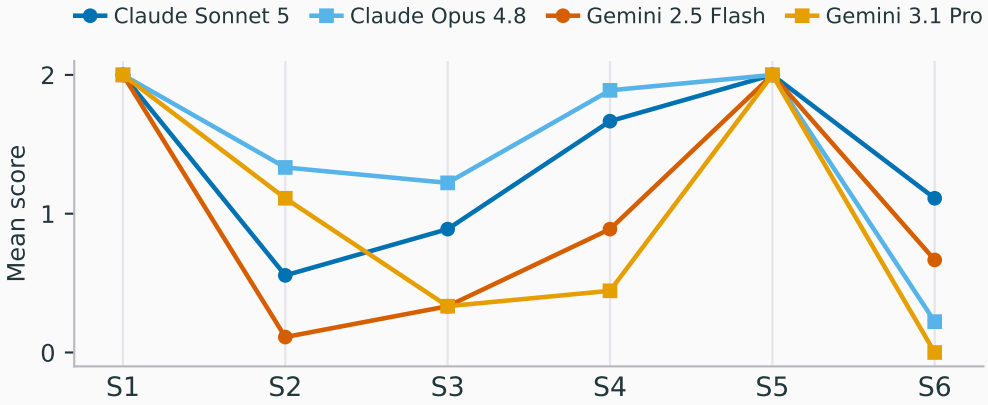
S4 Confounders named

S5 Measurement defined

S6 Overreach avoided

Two marks each, scored blind to which model wrote it

Where the gap actually sits



Two criteria separate the vendors · Two do not · Two are at ceiling

What can a language model do for a movement ecologist?

Oliver Burch

School of Mathematical and Physical Sciences, University of Sheffield

Supervisors: Jonathan Potts and Wei Xing

Which errors got caught tracked the vendor.

OBurch1@sheffield.ac.uk



Backup: the rubric

READING THE DATA

C1 Detects structure

C2 Number of groups

C3 Boundary placement

C4 What separates the groups

C5 Flags weak evidence

C6 Avoids unsupported claims

HYPOTHESISING

S1 Mechanism named

S2 Direction predicted

S3 Disproof stated

S4 Confounders named

S5 Measurement defined

S6 Overreach avoided

Backup: what three runs can carry

Three against three gives 20 arrangements, so the smallest possible p is **0.10**.

Framing comparisons cannot reach significance. The null on slide 7 is a limit of the design, not a result.

The flags pool across conditions instead: **12** responses per vendor, not three.

Sensor flag **7 against 0**, Fisher exact $p = \mathbf{0.005}$.

Backup: threats to the vendor claim

Two models per vendor cannot separate *vendor* from *these two models are alike*.

The vendors were not reached identically: Anthropic via OpenRouter, Gemini via the native API, sampling parameters unset against vendor default. Both asymmetries sit at vendor level.

Scoring: S6 is the criterion most exposed to prose style. No vendor split there, and the ordering interleaves, with a Claude model third of four.