

TODAYMECHANIC

ONE LLM CALL IS THE WRONG DEFAULT

Routing a multimodal problem to the model that can actually answer it

Geoff Niehaus · TodayMechanic



linkedin.com/in/geoffniehaus

WHO IS SAYING THIS

Geoff Niehaus

Founder, Today Mechanic

Before this: Executive Director of Software Development Engineering at CVS Health, via Oak Street Health, over a program of 100+ engineers.

Shipped applied AI in production: retrieval over millions of clinical documents into an LLM, a 30% lift in suspect finding, under HIPAA with human review in the loop.

Founding member of the Austin CTO Club. Patent pending on multi-tenant medical record access.

TodayMechanic

WHAT WE BUILD

Helping people get their vehicle fixed, without knowing anything about cars

Describe the problem in your own words

Find a shop, book it, approve the estimate

Every conversation is pinned to one vehicle

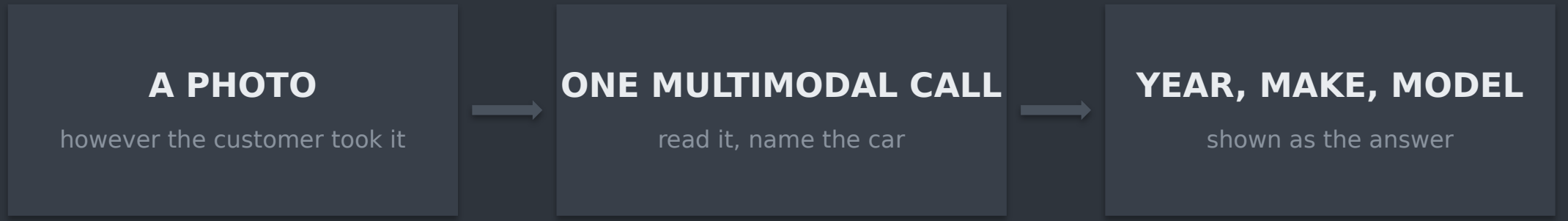
EVERYTHING IS PINNED TO ONE VEHICLE

So identification is the precondition, not a convenience

01	Chat is scoped to one vehicle
02	The estimate is written against one
03	The appointment books one into a bay
04	The parts are ordered for one

A wrong vehicle is invisible until a shop is already holding the wrong parts.

SO WE SHIPPED THE OBVIOUS THING



Seventeen seconds. $2.97 per thousand scans. Ships in an afternoon.

THEN WE MEASURED IT

A set of 150 actual VIN photos, rated by what appears to the customer

25%

of scans showed the customer the wrong vehicle

Right vehicle	**61**
Make only	**36**
Wrong vehicle shown	**37**
Told nothing	**16**

ONE PHOTO. FOUR RUNS.

The same Porsche windshield, four separate calls

Tesla Model X	Ford Transit	Audi Q7	Ford Expedition
5YJXCA2S9622S65048	7PDCA29922S650480	5YFCA2D09S5650480	5NPCA29225B650480

Same windshield. Same plate. Four manufacturers.

Not one of them is a Porsche.

AND CONFIDENCE CANNOT FILTER THEM

Median confidence, correct answers versus confabulated ones

CORRECT ANSWERS

0.72

INVENTED ANSWERS

0.72

THRESHOLD SWEEP

keep at 0.4

32/32 right · 20/22 wrong

keep at 0.9

7/32 right · 0/22 wrong

No cut-off keeps the good ones and drops the bad ones.

Raising the bar trades them roughly one for one.

SO WE BUILT THE FIX EVERYBODY BUILDS

Cheap reader first, fall back when it is not sure



Nothing controversial. This is the standard answer to an expensive model.

WHAT THAT ACTUALLY BOUGHT

Same 150 photos, cascade versus the single call

TOKEN SPEND / 1,000

$2.97

$1.91

36% cheaper

MEDIAN LATENCY

17.0s

14.2s

barely moved

WRONG ANSWERS

37

37

did not move at all

The cost graph looked like a win. Nothing else moved.

A CASCADE ONLY CHANGES THE PRICE.

A fallback still lets the big model answer the question it is worst at.

ONE PHOTO. TWO QUESTIONS.

AN OCR PROBLEM

READ SEVENTEEN STAMPED CHARACTERS

The small model wins it outright, and can prove it did

A VISION PROBLEM

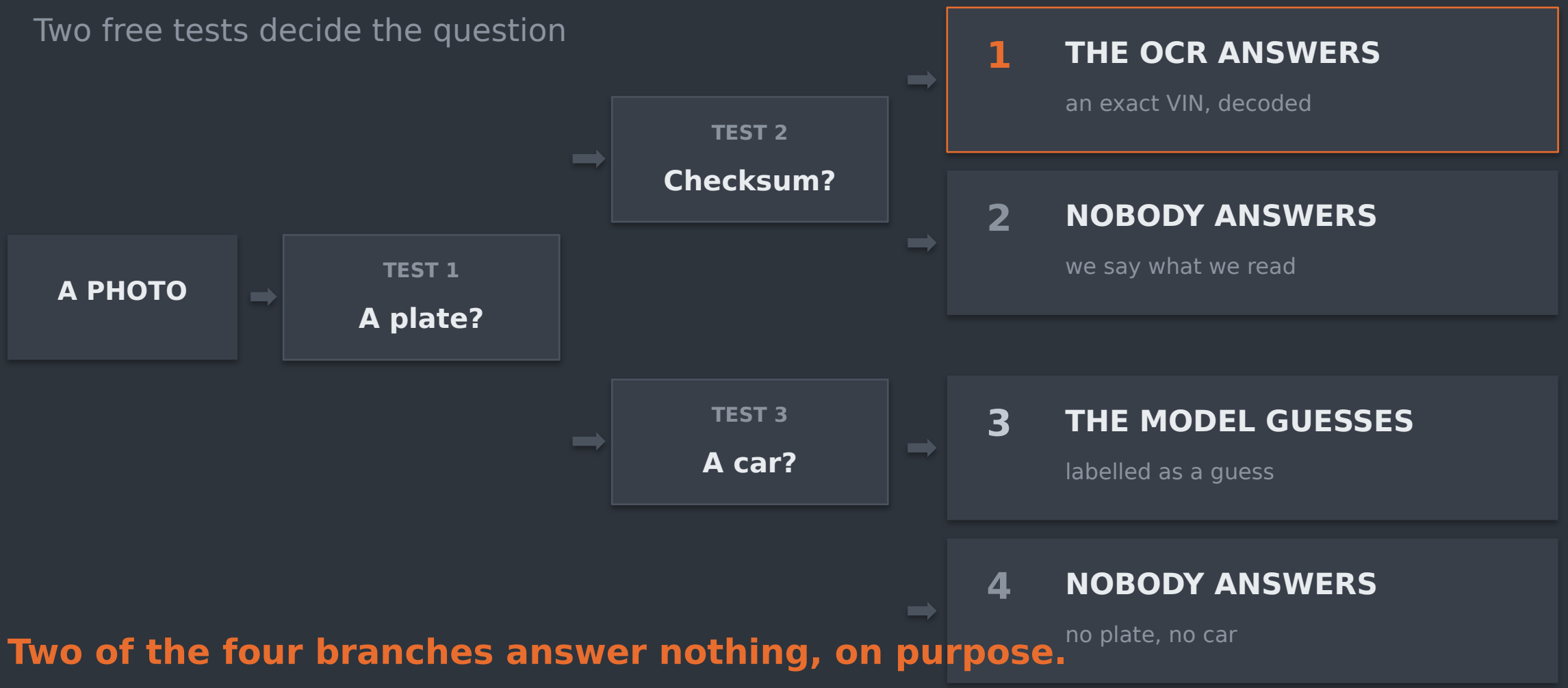
RECOGNISE A CAR FROM ITS BODYWORK

The OCR cannot attempt this one at all

Ask one model both, and the weaker answer arrives in the stronger one's voice.

ROUTE, DO NOT CASCADE

Two free tests decide the question



Two of the four branches answer nothing, on purpose.

WHY YOU CAN TRUST THE SMALL ONE

The cheap reader's answer validates itself, for free

1	H	G	C	M	8	2	6	3	3	A	0	0	4	3	5	2
								CHECK DIGIT								

Computed over the other sixteen. A clean read satisfies it; a mis-read almost never does.

Checksum alone

0.9796 precision

Checksum and a decoded model, what ships

1.0000 precision

THE BRANCH THAT VERIFIES

What the customer sees when we cannot read the plate

> "We read WP0AA298X7U7 off that plate, but it isn't a VIN we can look up. Pick your vehicle below."

01 Ask the model, then check its answer the way we check the OCR's

02 Two gates that fail in different places, so both must pass

03 A bad serial loses the VIN, not the car it still decodes to

THE CLASSIFIERS ARE FREE

Neither routing signal costs a model call

IS THERE A PLATE?

The OCR's own text detector

Already paid for. 7 of 300 cars misfire.

IS THERE A CAR?

A 3.6 MB object detector

10 ms per photo. It rescues plate misfires.

The model is never the first reader, and never the last word.

The win came from the classifier, not from the model.

THREE AXES, SAME DIRECTION

Same 22 photographs, both lanes, two runs

RIGHT VEHICLE

13 / 13

16 / 15

of 22, run one and two

WRONG VEHICLE SHOWN

8 / 8

5 / 6

the number that matters

MEDIAN SCAN

7.3s

5.0s

and 7.0s against 5.2s

Twenty two correlated images is a walk-through, not a benchmark.

WHERE IT IS STILL WRONG

The three things I would ask about if I were you

IT IS NOT ZERO ANY MORE

5 or 6 wrong, against 8

An earlier version refused when unsure and showed zero. This one answers, and is sometimes wrong. It still wins on the same photographs.

CHEAP IN TOKENS IS NOT FREE

1.38 CPU-seconds per scan

against 0.03 for the LLM path

Fixed infrastructure rather than per-call spend, and capped at two scans a pod.

THE ROUTING SIGNAL HAS ITS OWN ERROR RATE

7 of 300 whole-car photos trip the plate test on a badge or a dealer address. They now reach the model, not a refusal.

Volunteering the limits is what makes the rest of the numbers worth anything.

FIND YOUR OWN ROUTER

None of this is about VINs

01 What distinct questions is my input actually carrying?

One field, one call, does not mean one question

02 For each, is there a reader whose answer can be verified?

A checksum, a schema, a lookup that either resolves or does not

03 Is there a cheap signal that tells them apart?

Including one that tells you to answer nothing

YOUR CHEAP MODEL IS NOT A FALLBACK.

It is the right model for part of your traffic, and a free classifier is what tells them apart.

The most valuable branch we shipped is the one that calls nothing at all.



TodayMechanic

QUESTIONS

Geoff Niehaus
TodayMechanic