

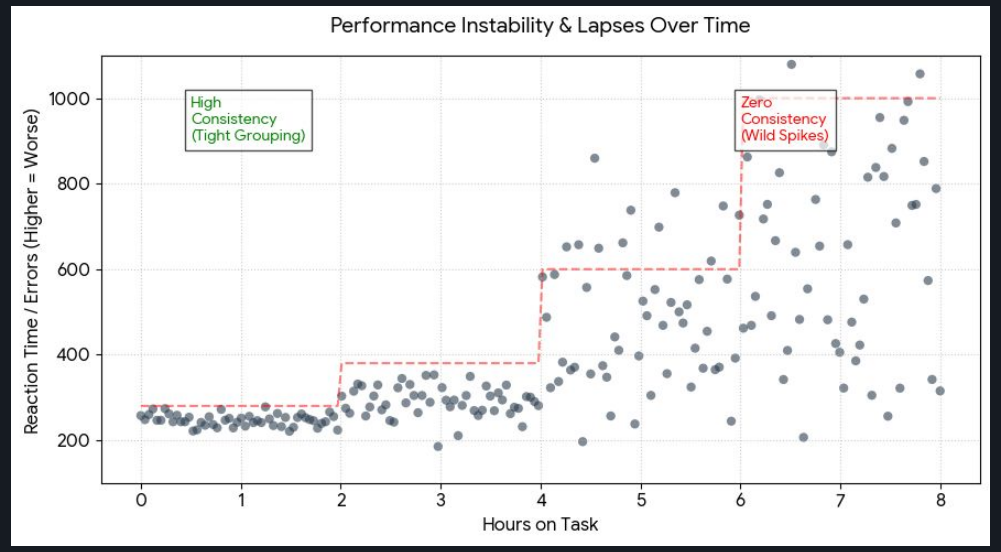
Active Genie

The Way to Consistency

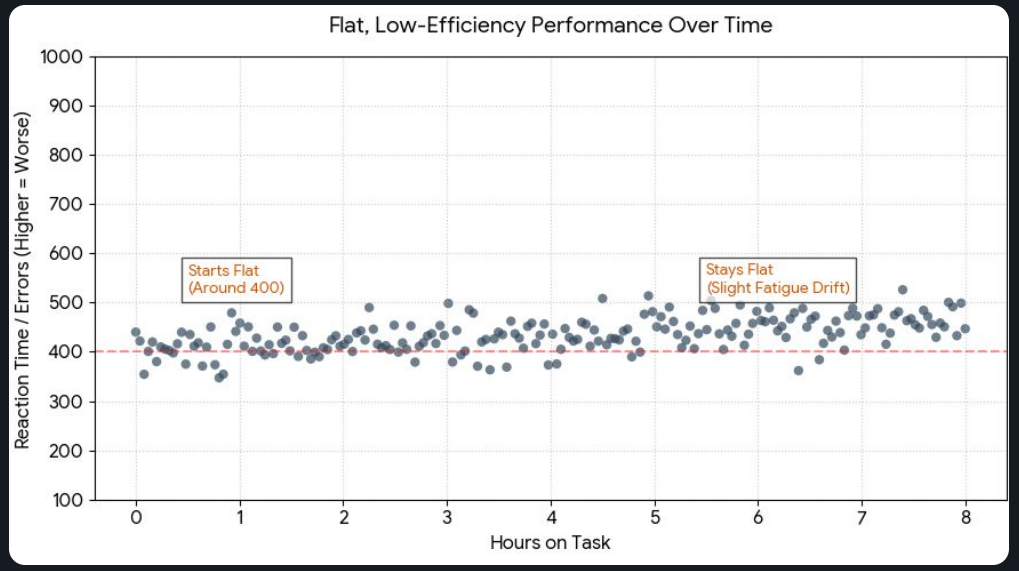
Engineering of Thinking

LLMs can be consistent enough

Humans



LLM



LLMs can be consistent enough, but we need to treat them as humans who never tire.
Welcome to the Engineering of Thinking.

The definition

Engineering of Thinking

practice of applying human mental models and procedures to machine reasoning

Human methodologies

We skip trying to **control the path** of the response. Instead, we use techniques like debate, consensus, and checklists to make the best response the only possible solution.

Structured Outputs

Uses **architectural constraints** instead of prompts. This limits drift and yields predictable output.

But first

Big thanks to Codeminer42





The Right Tool For the Job

GenAI doesn't solve everything

deterministic software will never die

Where Non-Determinism Shines

Deterministic software

Excels at **Structured & Hard Metrics**

- Price limits (e.g., < $150/night)
- Location coordinates/radius
- Amenities checkboxes (e.g., WiFi, Pool)

But there's a limit to what can be strictly quantified

Thinking

Excels at **Subjective** & Quality Metrics

- Evaluating abstract **taste** and atmosphere
- Gauging if a place is "**inspiring**" for writing
- Parsing review **sentiment** for isolation

Thrives when there are too many good options

The Inconsistency Trap

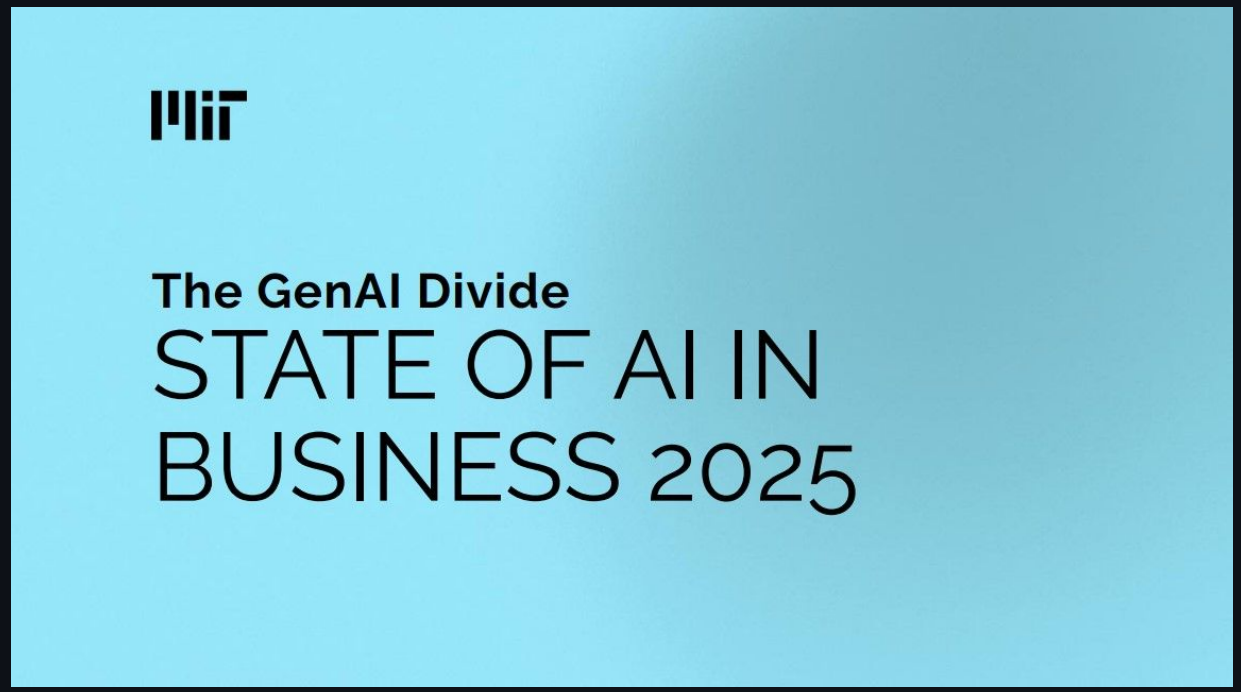


Subjectivity = Dice Roll

A ranking task that returns 10 different Top-3 outputs if asked 10 times, is impossible to scale and is impossible to maintain. No one can pay for that, but that’s the actual output for the majority of AI features.

```
Inconsistency is not a feature, it's an engineering liability.
```

The MIT State of AI in Business 2025 report



Despite **$30~40 billion** in enterprise investment into GenAI, this report uncovers a surprising result in that **95%** of organizations are getting **zero return**

https://mlq.ai/media/quarterly_decks/v0.1_State_of_AI_in_Business_2025_Report.pdf

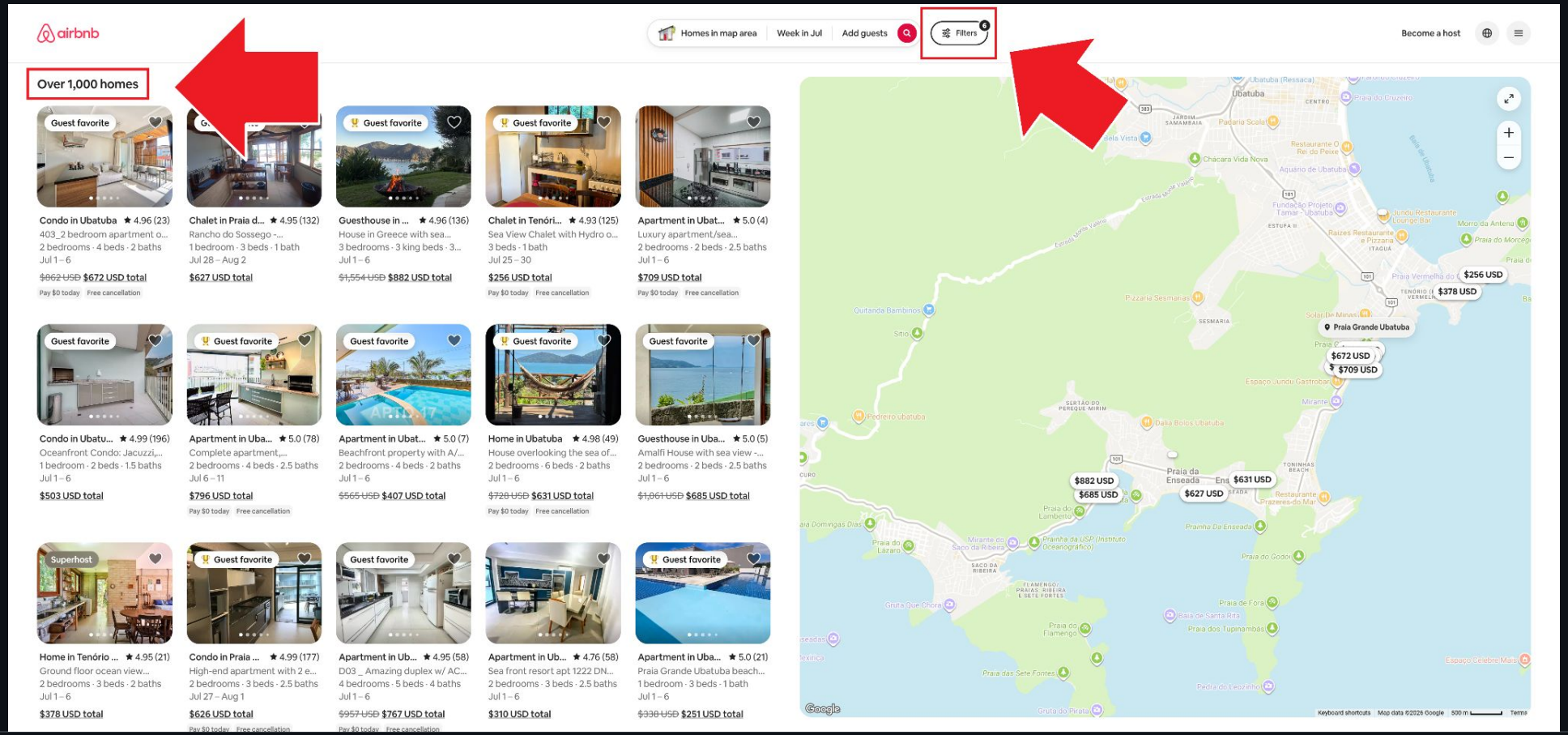


Real use case

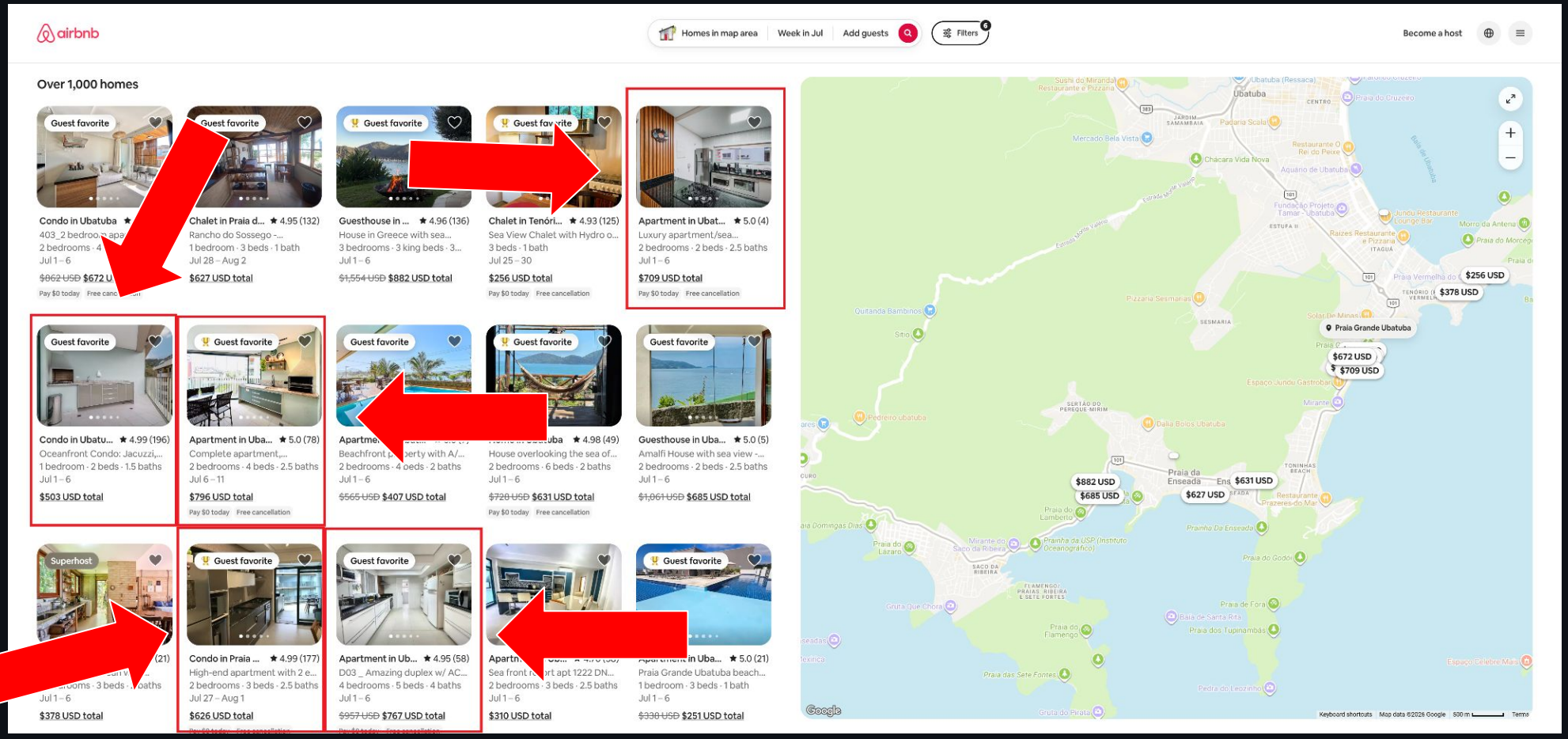
Let’s try create a real feature,

and prove it can work in the real world

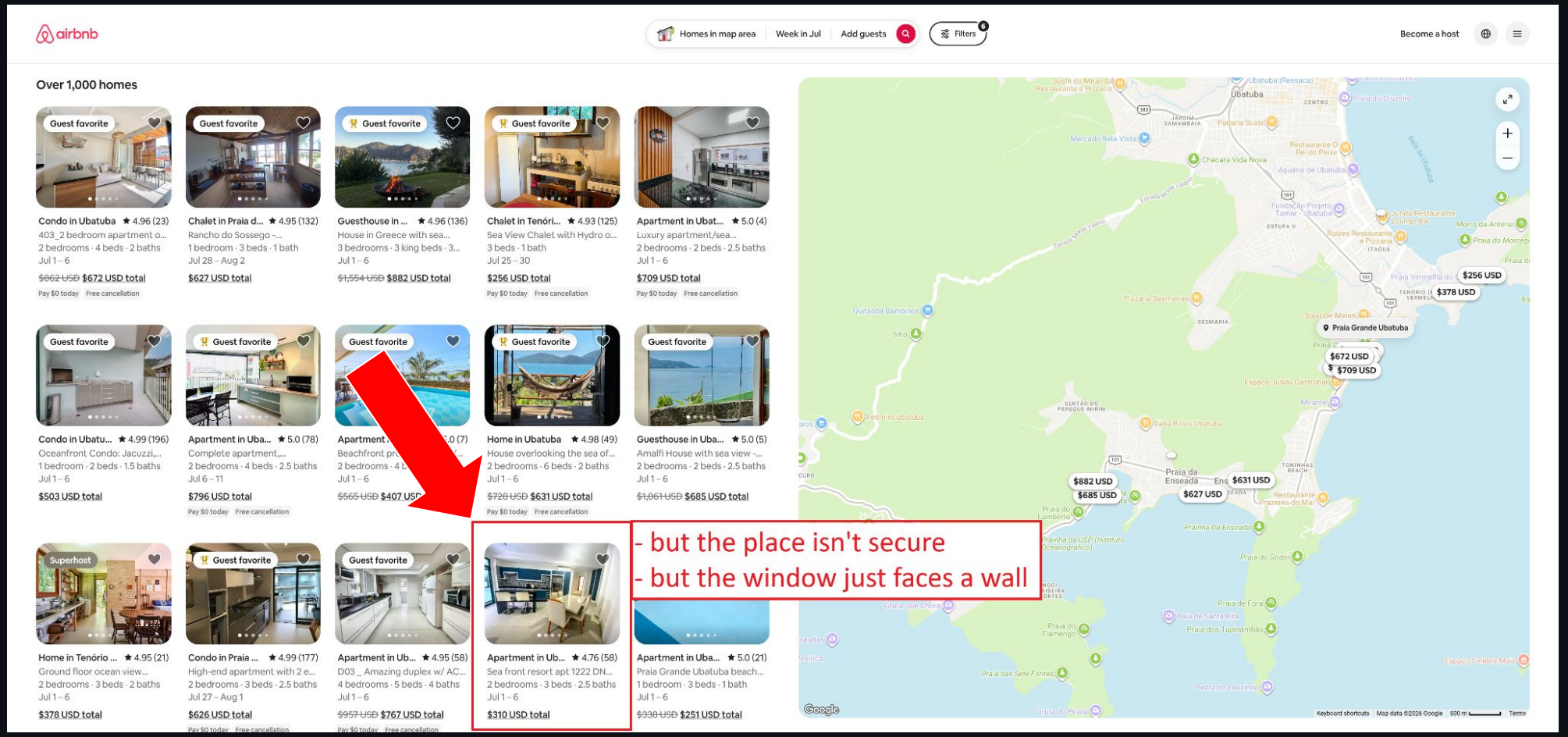
Search for a weekend beach place



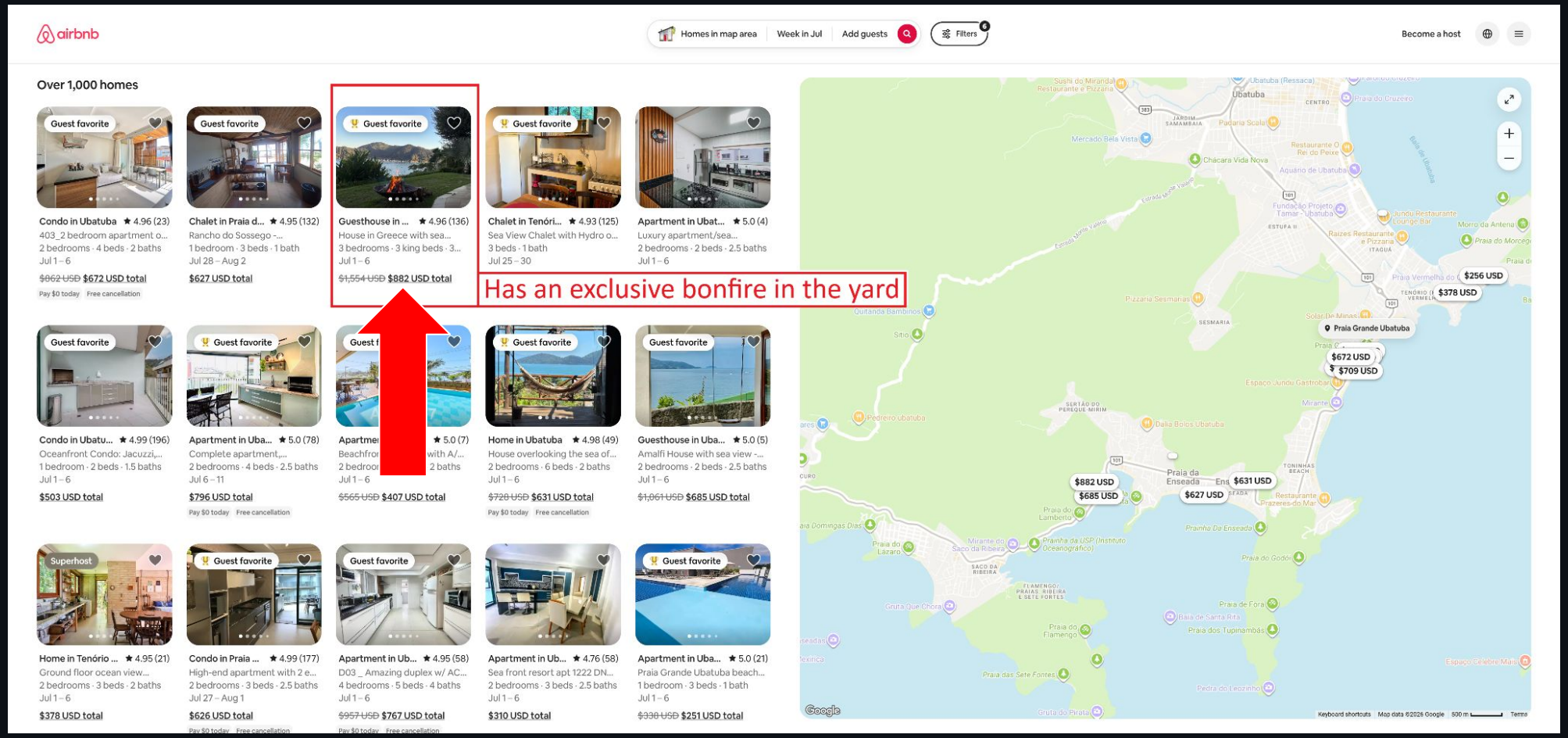
Bunch of similar places



Some clear bad ones



Some outstanding



Real case - Airbnb

How does a human solve it?

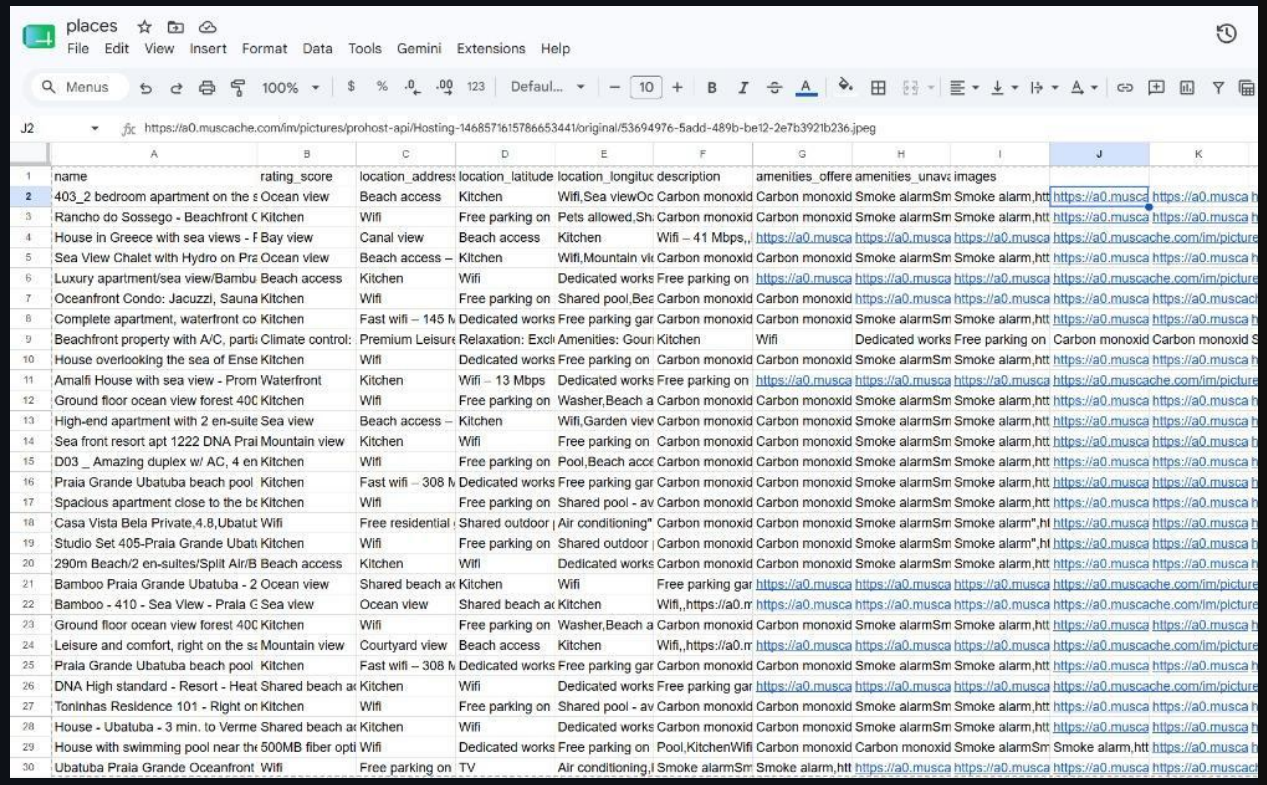
Human Execution

Humans click into reviews, analyze photos, and deduce qualitative details like noise and views. It takes significant manual time and focus. Highly prone to fatigue and subjective bias.

LLM Execution

Under structured process constraints, LLMs parse details at scale in seconds. Can replicate human judgment pathways. We just need to engineer the **process architecture**.

How does a machine solve it?



	A	B	C	D	E	F	G	H	I	J	K
1	name	rating_score	location_address	location_latitude	location_longitud	description	amenities_offere	amenities_unava	images		
2	403_2 bedroom apartment on the s	Ocean view	Beach access	Kitchen	Wifi,Sea viewOc	Carbon monoxid	Carbon monoxid	Smoke alarmSm	Smoke alarm,htt	https://a0.musca	https://a0.musca h
3	Rancho do Sossego - Beachfront (	Kitchen	Wifi	Free parking on	Pets allowed,Sh	Carbon monoxid	Carbon monoxid	Smoke alarmSm	Smoke alarm,htt	https://a0.musca	https://a0.musca h
4	House in Greece with sea views - F	Bay view	Canal view	Beach access	Kitchen	Wifi – 41 Mbps,,	https://a0.musca	https://a0.musca	https://a0.musca	https://a0.muscache.com/im/picture	
5	Sea View Chalet with Hydro on Pra	Ocean view	Beach access –	Kitchen	Wifi,Mountain vi	Carbon monoxid	Carbon monoxid	Smoke alarmSm	Smoke alarm,htt	https://a0.musca	https://a0.musca h
6	Luxury apartment/sea view/Bambu	Beach access	Kitchen	Wifi	Dedicated works	Free parking on	https://a0.musca	https://a0.musca	https://a0.musca	https://a0.muscache.com/im/picture	
7	Oceanfront Condo: Jacuzzi, Sauna	Kitchen	Wifi	Free parking on	Shared pool,Bea	Carbon monoxid	Carbon monoxid	https://a0.musca	https://a0.musca	https://a0.musca	https://a0.muscach
8	Complete apartment, waterfront co	Kitchen	Fast wifi – 145 M	Dedicated works	Free parking gar	Carbon monoxid	Carbon monoxid	Smoke alarmSm	Smoke alarm,htt	https://a0.musca	https://a0.musca h
9	Beachfront property with A/C, parti	Climate control:	Premium Leisure	Relaxation: Excl	Amenities: Gour	Kitchen	Wifi	Dedicated works	Free parking on	Carbon monoxid	Carbon monoxid S
10	House overlooking the sea of Ense	Kitchen	Wifi	Dedicated works	Free parking on	Carbon monoxid	Carbon monoxid	Smoke alarmSm	Smoke alarm,htt	https://a0.musca	https://a0.musca h
11	Amalfi House with sea view - Prom	Waterfront	Kitchen	Wifi – 13 Mbps	Dedicated works	Free parking on	https://a0.musca	https://a0.musca	https://a0.musca	https://a0.muscache.com/im/picture	
12	Ground floor ocean view forest 400	Kitchen	Wifi	Free parking on	Washer,Beach a	Carbon monoxid	Carbon monoxid	Smoke alarmSm	Smoke alarm,htt	https://a0.musca	https://a0.musca h
13	High-end apartment with 2 en-suite	Sea view	Beach access –	Kitchen	Wifi,Garden viev	Carbon monoxid	Carbon monoxid	Smoke alarmSm	Smoke alarm,htt	https://a0.musca	https://a0.musca h
14	Sea front resort apt 1222 DNA Prai	Mountain view	Kitchen	Wifi	Free parking on	Carbon monoxid	Carbon monoxid	Smoke alarmSm	Smoke alarm,htt	https://a0.musca	https://a0.musca h
15	D03 _ Amazing duplex w/ AC, 4 en	Kitchen	Wifi	Free parking on	Pool,Beach acc	Carbon monoxid	Carbon monoxid	Smoke alarmSm	Smoke alarm,htt	https://a0.musca	https://a0.musca h
16	Praia Grande Ubatuba beach pool	Kitchen	Fast wifi – 308 M	Dedicated works	Free parking gar	Carbon monoxid	Carbon monoxid	Smoke alarmSm	Smoke alarm,htt	https://a0.musca	https://a0.musca h
17	Spacious apartment close to the be	Kitchen	Wifi	Free parking on	Shared pool - av	Carbon monoxid	Carbon monoxid	Smoke alarmSm	Smoke alarm,htt	https://a0.musca	https://a0.musca h
18	Casa Vista Bela Private,4.8,Ubatut	Wifi	Free residential	Shared outdoor	Air conditioning"	Carbon monoxid	Carbon monoxid	Smoke alarmSm	Smoke alarm",ht	https://a0.musca	https://a0.musca h
19	Studio Set 405-Praia Grande Ubat	Kitchen	Wifi	Free parking on	Shared outdoor	Carbon monoxid	Carbon monoxid	Smoke alarmSm	Smoke alarm",ht	https://a0.musca	https://a0.musca h
20	290m Beach/2 en-suites/Split Air/B	Beach access	Kitchen	Wifi	Dedicated works	Carbon monoxid	Carbon monoxid	Smoke alarmSm	Smoke alarm,htt	https://a0.musca	https://a0.musca h
21	Bamboo Praia Grande Ubatuba - 2	Ocean view	Shared beach ac	Kitchen	Wifi	Free parking gar	https://a0.musca	https://a0.musca	https://a0.musca	https://a0.muscache.com/im/picture	
22	Bamboo - 410 - Sea View - Praia G	Sea view	Ocean view	Shared beach ac	Kitchen	Wifi,,https://a0.m	https://a0.musca	https://a0.musca	https://a0.musca	https://a0.muscache.com/im/picture	
23	Ground floor ocean view forest 400	Kitchen	Wifi	Free parking on	Washer,Beach a	Carbon monoxid	Carbon monoxid	Smoke alarmSm	Smoke alarm,htt	https://a0.musca	https://a0.musca h
24	Leisure and comfort, right on the sa	Mountain view	Courtyard view	Beach access	Kitchen	Wifi,,https://a0.m	https://a0.musca	https://a0.musca	https://a0.musca	https://a0.muscache.com/im/picture	
25	Praia Grande Ubatuba beach pool	Kitchen	Fast wifi – 308 M	Dedicated works	Free parking gar	Carbon monoxid	Carbon monoxid	Smoke alarmSm	Smoke alarm,htt	https://a0.musca	https://a0.musca h
26	DNA High standard - Resort - Heat	Shared beach ac	Kitchen	Wifi	Dedicated works	Free parking gar	https://a0.musca	https://a0.musca	https://a0.musca	https://a0.muscache.com/im/picture	
27	Tonlnhas Residence 101 - Right on	Kitchen	Wifi	Free parking on	Shared pool - av	Carbon monoxid	Carbon monoxid	Smoke alarmSm	Smoke alarm,htt	https://a0.musca	https://a0.musca h
28	House - Ubatuba - 3 min. to Verme	Shared beach ac	Kitchen	Wifi	Dedicated works	Carbon monoxid	Carbon monoxid	Smoke alarmSm	Smoke alarm,htt	https://a0.musca	https://a0.musca h
29	House with swimming pool near the	500MB fiber opti	Wifi	Dedicated works	Free parking on	Pool,KitchenWifi	Carbon monoxid	Carbon monoxid	Smoke alarmSm	Smoke alarm,htt	https://a0.musca h
30	Ubatuba Praia Grande Oceanfront	Wifi	Free parking on	TV	Air conditioning,I	Smoke alarmSm	Smoke alarm,htt	https://a0.musca	https://a0.musca	https://a0.musca	https://a0.muscach



I will give you a couple images from place to rent, describe the place as much as possible specially the good and bad aspects.

This looks like a modern, well-optimized beach apartment (the view in the third photo gives away that nice coastline). It's got a solid, contemporary design, but like any space, it has its trade-offs.

Here is a breakdown of the good and the bad based on the layout, furniture, and finishes.

The Good Aspects

- **Excellent Indoor-Outdoor Flow:** The large sliding glass doors completely open up the living room to the balcony. It creates a seamless transition, makes the apartment feel much larger, and lets in a ton of natural light.
- **Great Balcony & View:** You're right on the water with a clear view of the beach and a pool below. The wooden ceiling lining and the vertical slatted wood partition on the balcony add warmth and give it a premium, resort-like feel. Plus, having a full dining table outside is perfect for meals with a view.
- **Modern, Consistent Aesthetic:** The interior design is cohesive. The slatted wood TV panel matches the kitchen island and balcony ceiling, while the light grey floor tiles keep everything looking clean and airy.
- **Functional Kitchen Setup:** For a compact space, the kitchen is well-equipped. It features an

Ask Gemini Flash

The Wrong implementation

One-Shot Prompting

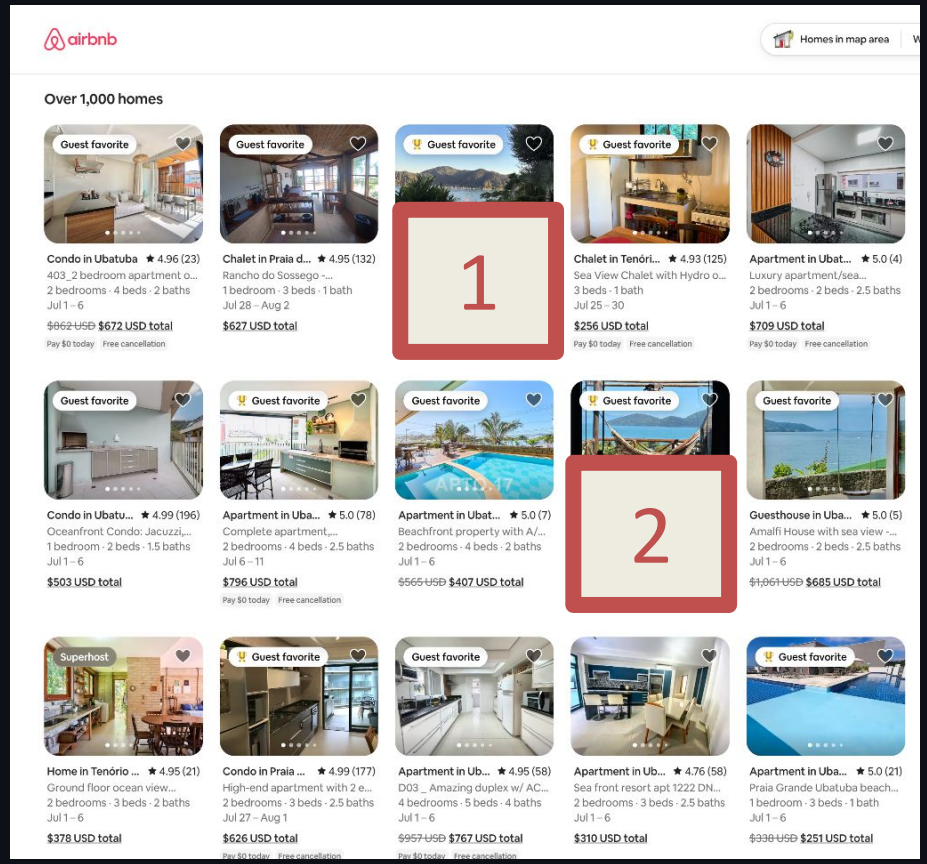
- Vibe prompt
- No context window control
- High bias contamination
- Each run yields a different result

```
16  prompt = <<~PROMPT
17  what are the top 5 best places in the list to: #{criteria}
18
19  #{all_places_descriptions.shuffle.join("\n\n")}
20  PROMPT
21
22  payload = {
23    model: "my-preference-model",
24    messages: [
25      { role: "user", content: prompt }
26    ]
27  }
```

Wrong implementation

Run 1

1. House in Greece;
2. House overlooking Enseada;
3. Bamboo Praia Grande



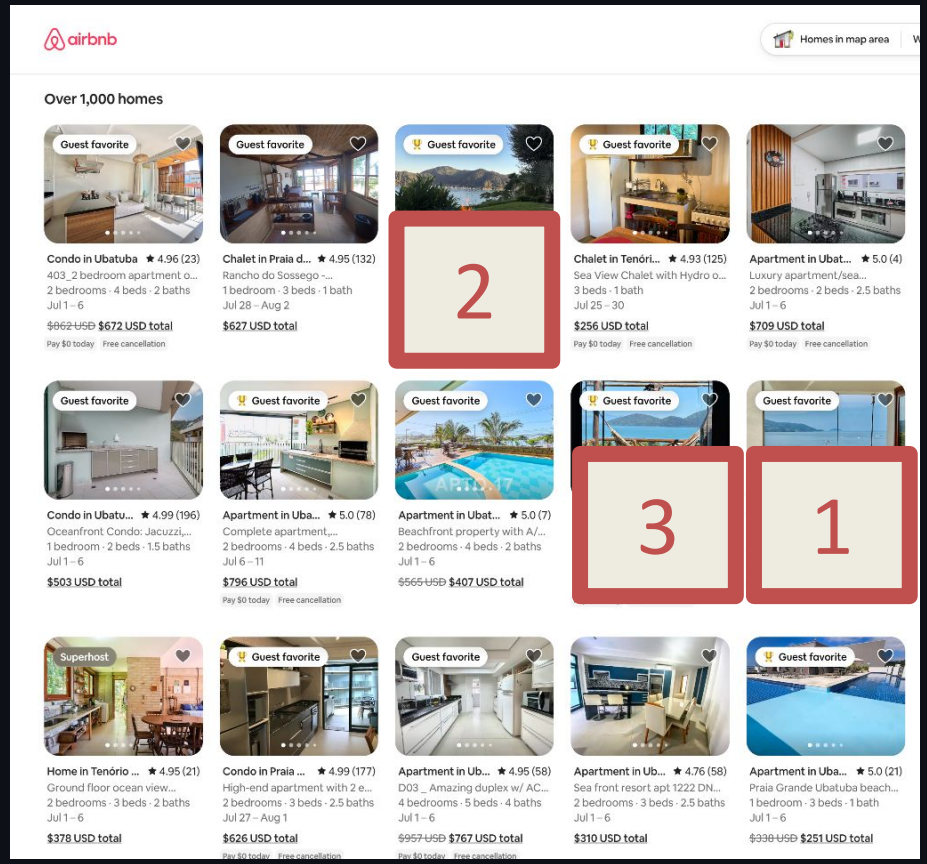
Wrong implementation

Run 1

1. House in Greece;
2. House overlooking Enseada;
3. Bamboo Praia Grande;

Run 2

1. Amalfi House;
2. House in Greece;
3. House overlooking Enseada



Wrong implementation

Run 1

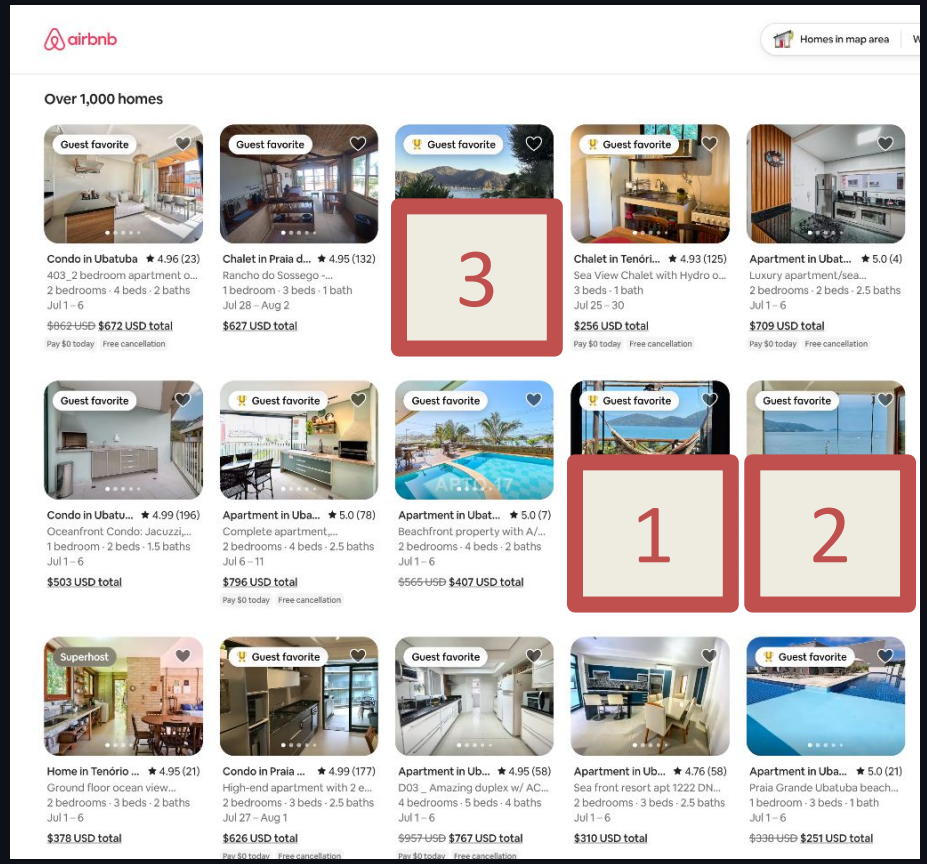
1. House in Greece;
2. House overlooking Enseada;
3. Bamboo Praia Grande;

Run 2

1. Amalfi House;
2. House in Greece;
3. House overlooking Enseada

Run 3

1. House overlooking Enseada;
2. Amalfi House;
3. House in Greece



How does a wrong implementation perform?

Runs

Position shift of top properties across 3 One-Shot runs:

· House in Greece: ranks [1, 2, 3] (Shift: 2)

· House overlooking Enseada: ranks [4, 4, 1] (Shift: 3)

· Amalfi House: ranks [2, 1, 2] (Shift: 1)

· Bamboo Praia Grande: ranks [3, 3, 5] (Shift: 2)

Severe Positional Bias

Changing input ordering yields different winners.

Stability metrics:

· House in Greece: **59.18%**

· House overlooking Enseada: **59.18%**

· Amalfi House: **37.64%**

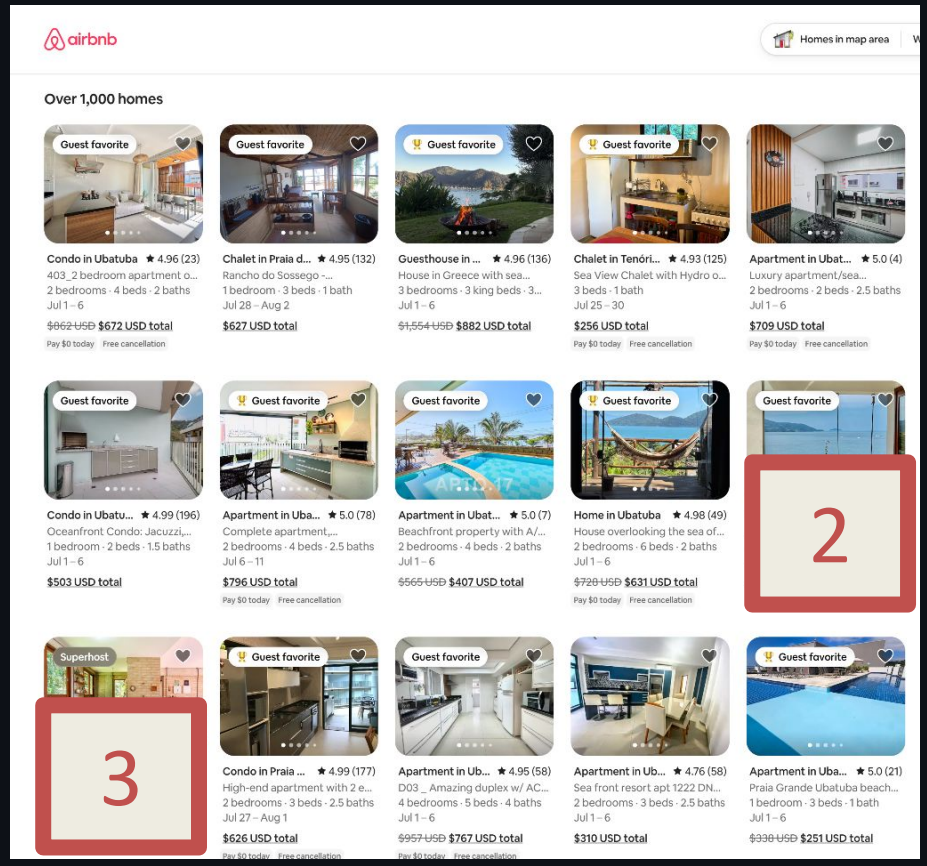
· Bamboo Praia Grande: **76.43%**

ActiveGenie Top 3 Finalists

1. 290m Beach / 2 en-suites
   ELO Score: 1019 | 100% Run Stability

2. Amalfi House with sea view
   ELO Score: 1018 | 100% Run Stability

3. Ground floor ocean view forest
   ELO Score: 1012 | 100% Run Stability



Real case - Airbnb

ActiveGenie Top 3 Finalists

1. 290m Beach / 2 en-suites
   ELO Score: 1019 | 100% Run Stability

2. Amalfi House with sea view
   ELO Score: 1018 | 100% Run Stability

3. Ground floor ocean view forest
   ELO Score: 1012 | 100% Run Stability

Engine Metrics

· Total Battles Simulated: **91 debates**

· Execution Time: **172 seconds**

· API Cost: **$0.087 (397k input + 112k output)**

Consistency is the only possible output

Run 1, Run 2, Run 3 output the exact same ranking.

The Consistency

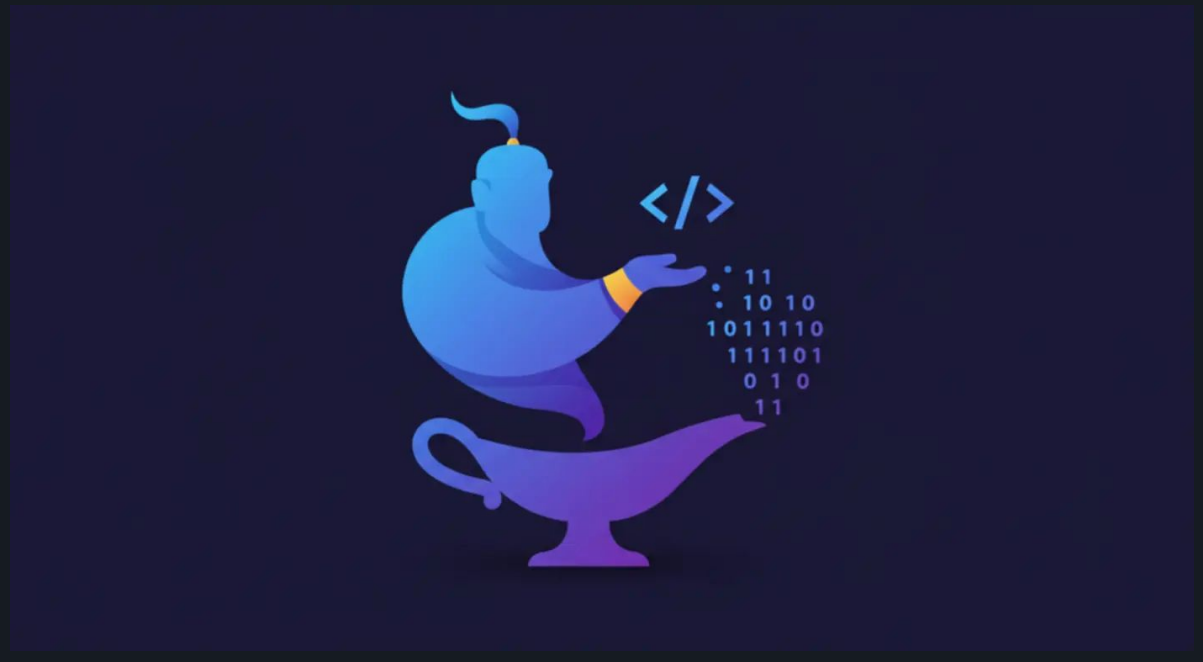


Value of Consistency

Why does predictability matter in GenAI?

- **Deterministic output** lets engineers build reliable interfaces.
- Eliminates random shuffling errors and bias.
- Turns subjective models into predictable API features.

Consistency is the bridge to enterprise GenAI value.



Active Genie

The Lodash for GenAI: Real Value + Consistent + Model-Agnostic

Architecture

1. Jury Bench

Individual Scoring

· 28 listings evaluated, 14 eliminated

· 3 LLM juries per listing

· 52s execution time

Filters outliers early.

2. ELO Tournament

3. Free-For-All

Architecture

1. Jury Bench

Individual Scoring

· 28 listings evaluated, 14 eliminated

· 3 LLM juries per listing

· 52s execution time

Filters outliers early.

2. ELO Tournament

Bracket Pruning

· Multi-tier bracket combats

· Fast elimination thresholds

· Keeps top 15 or less survivors

Reduces API costs by 90%.

3. Free-For-All

Architecture

1. Jury Bench

Individual Scoring

· 28 listings evaluated, 14 eliminated

· 3 LLM juries per listing

· 52s execution time

Filters outliers early.

2. ELO Tournament

Bracket Pruning

· Multi-tier bracket combats

· Fast elimination thresholds

· Keeps top 15 or less survivors

Reduces API costs by 90%.

3. Free-For-All

Unrestricted Finals

· 91 round-robin debates

· Deep 1v1 cross-argumentation

· 120s execution time

Produces definitive rankings.

The Jury Bench

```
// Run the jury assessment
reviewers = [
  "Budget-Conscious Tourist",
  "Mother of a Family",
  "Senior Hotel Administrator"
]

result = ActiveGenie::Scorer.call(
   place, criteria, reviewers
)

// Access structured consensus output
result.data
# => { "budget_conscious_tourist_score" => 85, ... }
```



Conflict as Calibration

We model **imperfect, conflicting perspectives** to force a structured, balanced consensus.

· Budget-Conscious Tourist (evaluates **cost/value**)

· Mother of a Family (**safety**, space, utilities)

· Senior Hotel Administrator (standard, **facilities**)

Produces a calibrated, multi-dimensional score.

How Jury Bench Works

```ruby
# active_genie/scorer/jury_bench.rb
def properties
  prop = {}
  juries.each do |jury|
    jury_key = underscore(jury)
    prop["#{jury_key}_reasoning"] = { type: 'string' }
    prop["#{jury_key}_score"] = { type: 'number' }
  end
  prop[:final_reasoning] = { type: 'string' }
  prop[:final_score] = { type: 'number' }
  prop
end
```

</> Structured Function Calling

No generic prompts. We enforce a **JSON Schema**.

- Force model to explain reasoning for each jury
- Force reasoning **before** writing the score fields
- Aligns LLM attention to reduce hallucinations

Yields predictable structured data classes.

Individual Scoring Highlight



Step 2 Highlighted

Eliminating high-variance and low-ranking stays. Only the most consistent options pass this gate. This focuses LLM resources on the hardest choices. Filters 28 properties down to 14

290m Beach/2 - 69/100

Beautiful Japandi design, excellent natural light, and mountain views, but docked because it lacks a clearly described ergonomic desk/chair

Studio Set 405-Praia Grande - 40/100

Great rooftop pool and modern look, but has no dedicated workspace, a very cramped layout, and the shared pool area poses noise distractions

Spacious apartment close to the beach - 25/100

Functional kitchen and close to beach, but lacks a dedicated workspace, natural light is not validated, and has very basic finishes with no view

ELO Tournament

```
# ELO tournament bottom-up filter
result = ActiveGenie::Ranker.by_tournament(
  listings,
  criteria,
  config: {
    ranker: {
      score_variation_threshold: 0.15,
      elo_initial: 1200
    }
  }
)

# => ["Luxury Studio", "Rustic Cabin", ...]
```



Tournament Pruning

ELO Tournament

```ruby
# ELO tournament bottom-up filter
result = ActiveGenie::Ranker.by_tournament(
  listings,
  criteria,
  config: {
    ranker: {
      score_variation_threshold: 0.15,
      elo_initial: 1200
    }
  }
)

# => ["Luxury Studio", "Rustic Cabin", ...]
```



Tournament Pruning

Why run a chess tournament?

- **Reduces LLM calls** compared to brute-force combinations
- Eliminates low scorers in early rounds
- Filters out high-variance options.

Prunes the list down to 15 or fewer finalists

The ELO Debate Loop

```ruby
# active_genie/lib/active_genie/ranker/elo.rb
matches do |a, b|
   winner, loser = debate(a, b)
   update_players_elo(winner, loser)
end

# ELO update math
def calculate_new_elo(rating, opp_rating, score)
   expected = 1 / (1 + (10**((opp_rating - rating)/400)))
   rating + (K * (score - expected)).round
end
```



Debate by rounds

```
{
  "name": "comparation_through_debate",
  "description": "Evaluate a debate between player_a and player_b
  "parameters": {
    "type": "object",
    "properties": {
      "player_a_sell_himself": { ... },
      "player_b_sell_himself": { ... },

      "player_a_arguments": { ... },
      "player_b_counter": { ... },

      "player_a_adherence_score": { ... },
      "player_b_adherence_score": { ... },

      "player_a_quality_score": { ... },
      "player_b_quality_score": { ... },

      "player_a_risk_avoidance_score": { ... },
      "player_b_risk_avoidance_score": { ... },

      "impartial_judge_winner_reasoning": { ... },
      "impartial_judge_winner": { ... }
```

Debate

Impartial Judge is prompted to compare options:

- One player at a time
- Sequential logic of arguments and counters
- An impartial judge ensuring fairness

The deeper the debate goes, the more consistent the result becomes

Inside the Debates

Player A: House - Ubatuba - 3 min. to Vermelha Beach

Quiet rainforest isolation, spacious raised wooden deck with pergolas and white-fringed hammocks, multiple bedrooms (work-life separation), excellent natural light from floor-to-ceiling glass sliding doors and transom windows.

Player B: Rancho do Sossego - Beachfront Chalet

Unbeatable beachfront location (step straight from bed onto the sand), excellent indoor-outdoor flow via yellow-framed glass window walls, fire pit, hammocks

Verdict: Player A (House - Ubatuba - 3 min. to Vermelha Beach) Wins

The Impartial Judge's Decision: "a beachfront location's constant waves and beach activity present a risk of "overwhelming focus with continuous sensory stimulation" for a writer"

The Free-For-All Finals

Underdog Promotion

House - Ubatuba - 3 min. to Vermelha Beach

- Initial Jury Score: **55/100** | Initial Rank: **12th** out of 14 finalists
- Final rank: **6th** | Wins: **8 out of 13** matches, ELO: 1005

Why the Underdog Wins

Functional criteria override generic checklists:

· Other options suffered from **resort noise distractions** and beachfront tourist foot traffic.

· It offered complete **quiet isolation**, ideal for writing a book without distraction

The Cost-Time Efficiency

Execution Time

172 Seconds

91 parallel matches completed

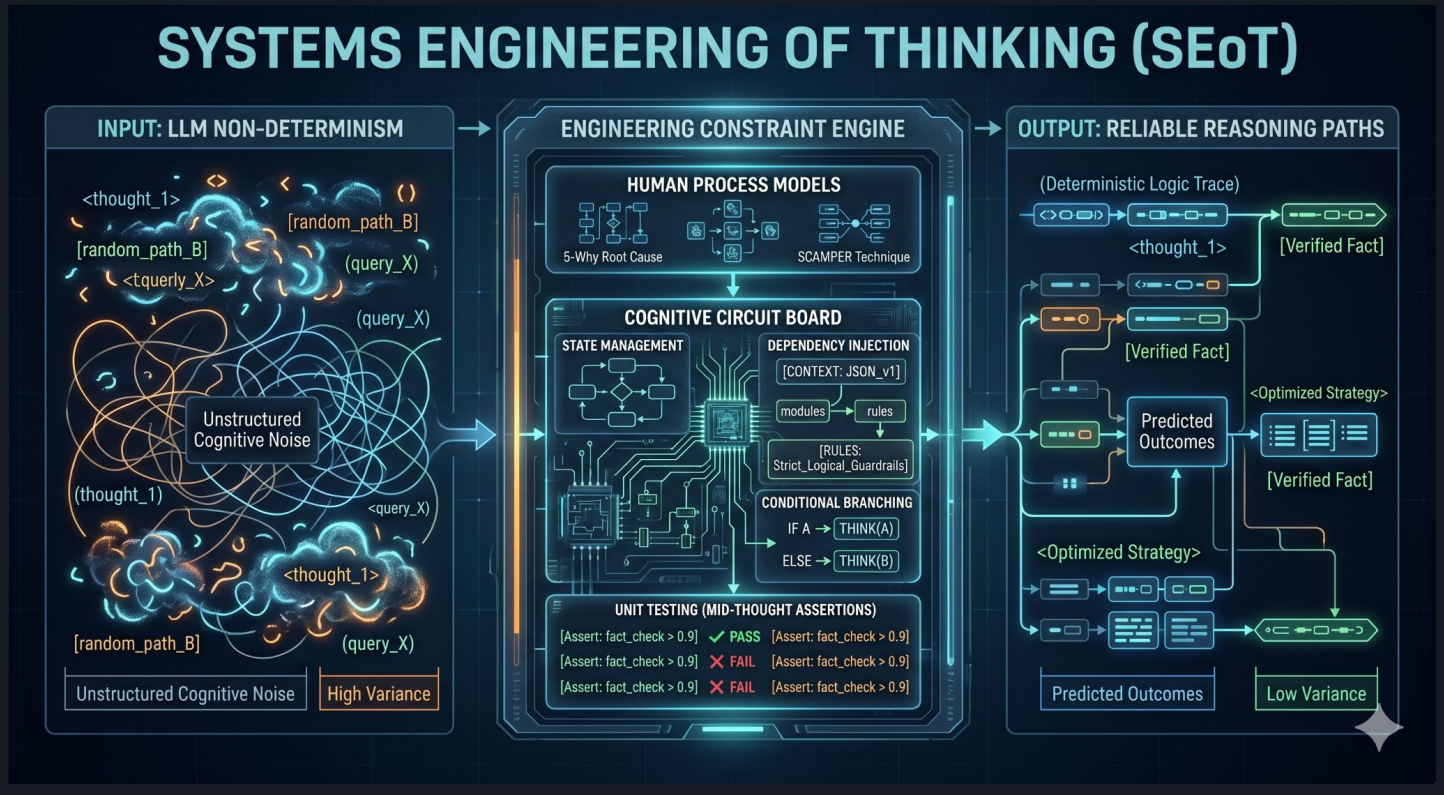
API Cost

~$0.08

397k input tokens + 112k output tokens

Can a human match this consistency?

Evaluating dozens of properties manually takes hours, yielding inconsistent choices every day.



Engineering of Thinking

Applying human cognitive models to architect deterministic LLM reasoning

The Human–AI Parallel

Human Processes

Decades of corporate practices are built to constrain human non-determinism:

- Peer reviews & diverse panel juries
- Tiered qualification tournaments
- Final 1v1 debate rounds

ActiveGenie Parallel

Processes built for humans work on LLMs, because models are trained to mimic humans:

- Jury Bench Scorer (Stage 1)
- ELO Tournament Bracket Pruning (Stage 2)
- Impartial Judge Free-For-All (Stage 3)

What Comes Next

Explore additional human mental models inside LLMs:

- **Eisenhower Matrix** — Urgent vs. Important categorization
- **RICE Scoring** — Reach, Impact, Confidence, Effort
- **Six Thinking Hats** — Parallel reasoning from forced roles
- **Pre-mortem Analysis** — Imagining failure paths to mitigate risk
- **OODA Loop** — Observe, Orient, Decide, Act loop

Check out
https://github.com/Roriz/active_genie

