Interpretable Feature Discovery from Unstructured Data

Application: using LLMs to grade and evaluate at scale

Sébastien Martin
with Baptiste Arnaudo & Amine Bennouna
INFORMS Annual Meeting · November 3, 2026

Collaborators



Baptiste Arnaudo

Stanford ICME · École Polytechnique



Amine Bennouna

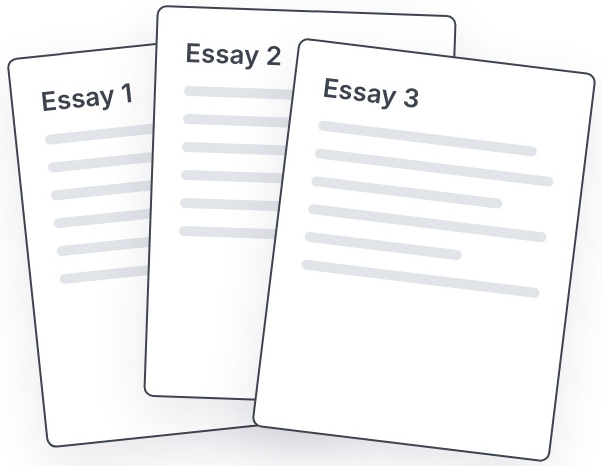
Northwestern Kellogg

Motivation

INTERPRETABLE FEATURE DISCOVERY FROM UNSTRUCTURED DATA

MOTIVATION

Imagine grading 200 essays



...and 197 more

	Argument	Evidence	Clarity	Grammar
Essay 1				
Essay 2				
Essay 3				
...				

A rubric: the features that matter.

MOTIVATION

What we want

A list of features e.g., the columns of the table before

NAME

Use of evidence

DESCRIPTION

Supports claims with concrete examples, data or sources, rather than generic statements.

SCORES

Essay 1 **2** Essay 2 **4** Essay 3 **3** ...

Enough for a human, or an AI, to understand what it means in this context, and to score it.

MOTIVATION

More generally: feature discovery on unstructured data

Grading

Hiring

Evaluating research papers

Machine learning on unstructured data

Post-hoc interpretability

Unstructured data
text, images...

→

Interpretable features
name + description

→

Scores
every item, every feature

MOTIVATION

Feature discovery, more precisely

Given n unstructured items $x_1, \ldots, x_n$

essays, cover letters, Reddit posts...

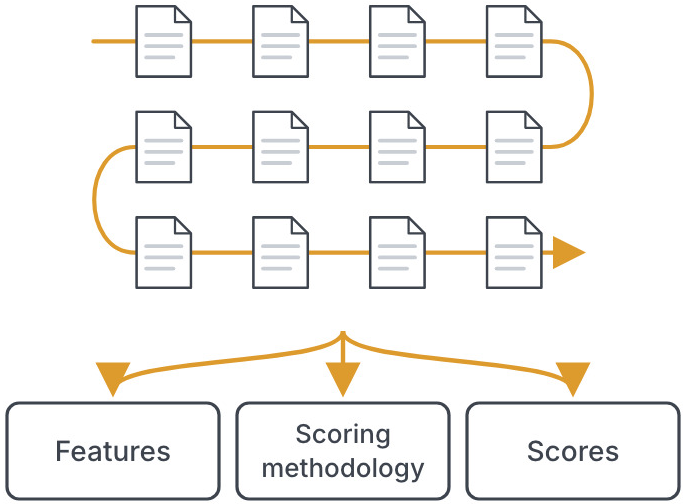
Find K features, each a name and a plain-language description

and a score matrix $S \in \mathbb{R}^{n \times K}$, where S_{ik} = score of item i on feature k

MOTIVATION

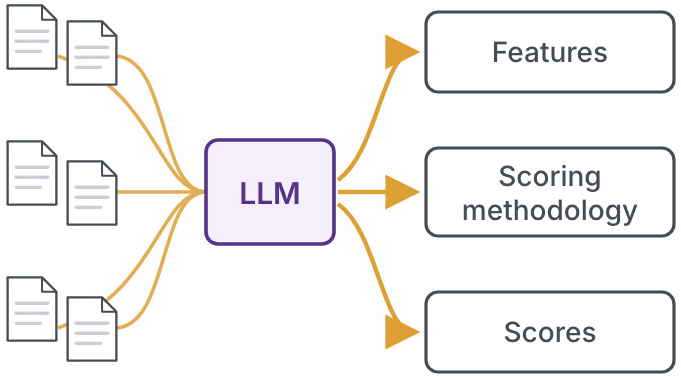
How could we do this?

By hand



- \+ Expert judgment
- – Biased
- – Not reproducible
- – Doesn't scale

One-shot LLM

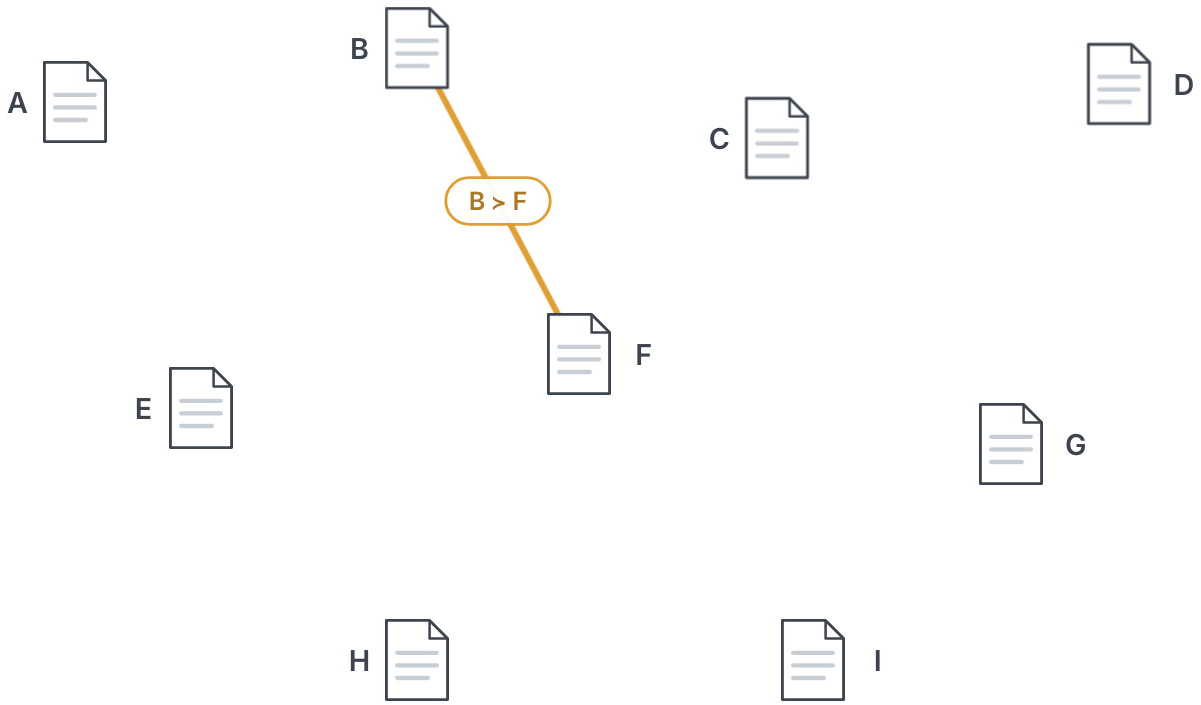


- \+ Automated, reproducible
- – Biased
- – Doesn't scale (context limit)
- – Quality issues

MOTIVATION

Another angle: comparative evaluation

Suppose we are given one feature to score, e.g. **Use of evidence**.



For each edge, ask the LLM:

Your goal is to evaluate which of two essays does better on the feature **"Use of evidence"**.

[detailed description of the feature]

Essay B: [...] Essay F: [...]

Which essay makes better use of evidence, B or F?

↓

LLM

↓

"B"

MOTIVATION

How do we derive scores from comparisons?

> Like **Elo ratings** in chess: one rating per player, from many games against different opponents. The canonical model: **Bradley–Terry**.

Model

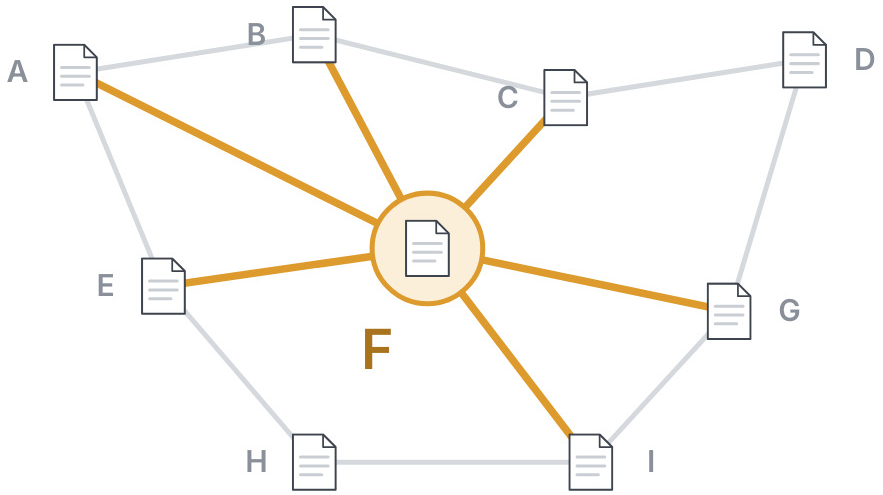
$$\mathbb{P}(i \text{ beats } j) = \sigma(s_i - s_j) = \frac{1}{1 + e^{-(s_i - s_j)}}$$

s_i: latent score of item i

Estimation: maximum likelihood

$$\hat{s} \in \arg\max_{s} \sum_{\text{comparisons } i \succ j} \log \sigma(s_i - s_j)$$

Advantages of Bradley–Terry



Item F appears in 6 comparisons.

Many comparisons per item

As a consequence, we can:

1. **Tolerate noise.** It is a statistical model: errors in individual LLM answers average out.

2. **Correct biases.** E.g. position bias: the LLM tends to favor the item shown first. Add one parameter β:

$$\mathbb{P}(i \text{ beats } j \mid i \text{ shown first}) = \sigma(s_i - s_j + \beta)$$

β is estimated by maximum likelihood, together with the scores.

How do we use the comparison idea to "discover" features, not only score them?

Our approach & theory

INTERPRETABLE FEATURE DISCOVERY FROM UNSTRUCTURED DATA

Step 0: a dataset and a comparison budget

Data

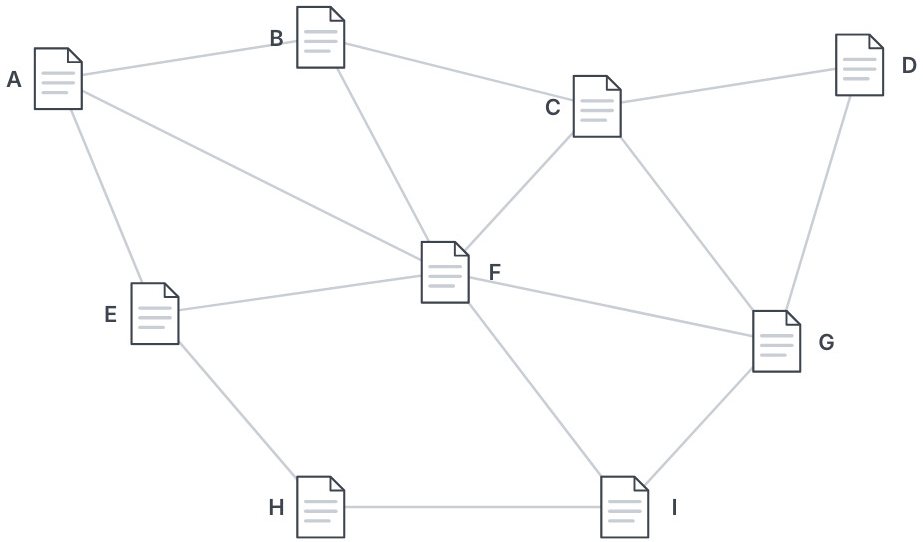
n items of unstructured data

e.g. $n = 200$ essays

Budget

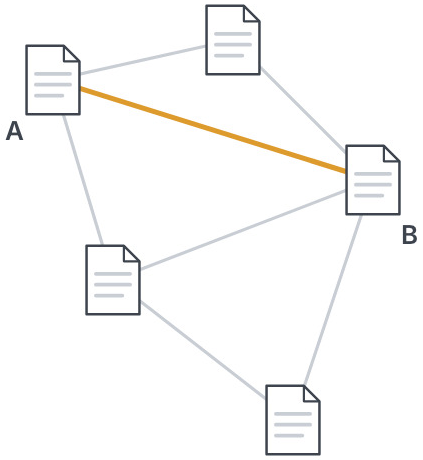
Each item is compared about d times, so about $dn/2$ LLM calls

e.g. $d = 10$: 1,000 calls. More comparisons, more accuracy.



random comparison graph

Step 1: ask the LLM how two essays differ



comparison graph

Two essays by MBA students in an operations class, that we will need to grade: [A], [B].

Find grading criteria on which they meaningfully differ. [...]

For each criterion: a name, a short description, and which essay is stronger.

(simplified prompt)

→ **LLM** →

Clear, engaging prose	A
Backs claims with statistics	B
Strong, focused thesis	A
...	

1. **Comparisons surface differences:** the features found are those that differentiate the items, which is our goal.
2. **Small context:** each call sees only two essays.
3. **Highly parallelizable.**

Step 2: group into K shared features

Raw phrases, from many pairs

clear, engaging prose · *readable* · *well-written* → **Clarity**

backs claims with statistics · *cites sources* → **Evidence**

strong, focused thesis · *convincing reasoning* → **Argument**

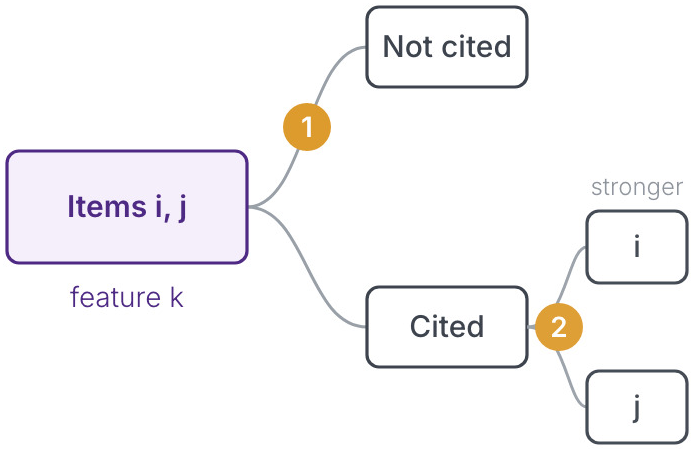
Grouping done by an LLM, or by clustering in an embedding space.

Verdicts pairs × K features

Pair	Clarity	Evidence	Argument
A vs B	A	B	A
A vs E	·	E	·
C vs D	·	·	C
…	…	…	…

· = not cited. Silence is also data.

A model of the LLM's two decisions



1 Does the LLM cite feature k when comparing i and j ?

$$\mathbb{P}(\text{cite } k \mid i, j) = \sigma\big(\,|S_{ik} - S_{jk}| + \gamma_k\,\big)$$

S_{ik}: intrinsic score of item i on feature k (unknown)

γ_k: **citation bias**

2 If k is cited, which item is stronger?

$$\mathbb{P}(i \text{ stronger on } k \mid i, j,\ k \text{ cited}) = \sigma\big(\,S_{ik} - S_{jk} + \beta\,\big)$$

β: **position bias** (i shown first)

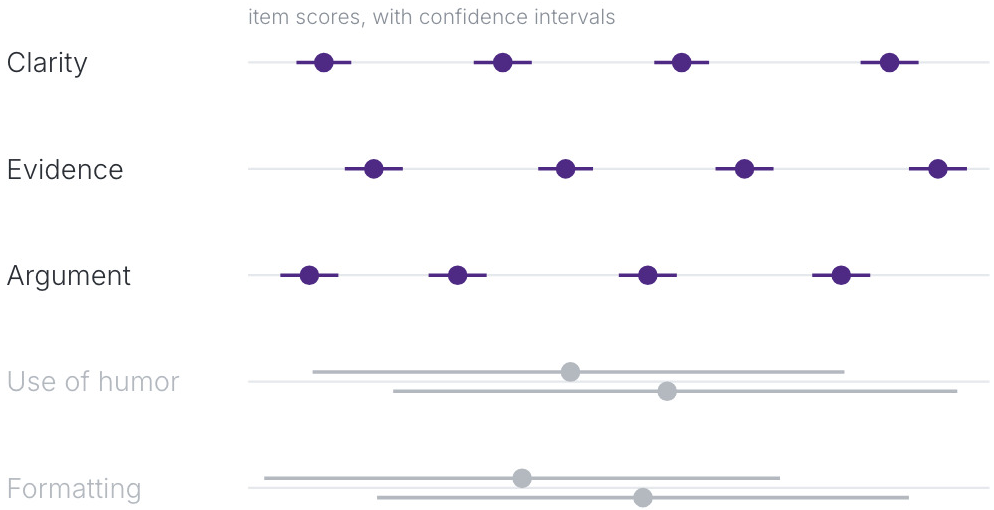
Step 3: estimate by maximum likelihood

Combining both decisions, we get our LLM choice model. Verdict $T_{ijk} \in \{+1, -1, 0\}$ (i stronger, j stronger, not cited):

$$\mathbb{P}(T_{ijk} = +1) = \sigma\big(|S_{ik} - S_{jk}| + \gamma_k\big)\, \sigma\big(S_{ik} - S_{jk} + \beta\big)$$
$$\mathbb{P}(T_{ijk} = -1) = \sigma\big(|S_{ik} - S_{jk}| + \gamma_k\big)\, \big(1 - \sigma(S_{ik} - S_{jk} + \beta)\big)$$
$$\mathbb{P}(T_{ijk} = 0) = 1 - \sigma\big(|S_{ik} - S_{jk}| + \gamma_k\big)$$

$$(\hat{S}, \hat{\gamma}, \hat{\beta}) \in \arg\max_{S,\,\gamma,\,\beta} \sum_{\text{pairs } (i,j)} \sum_{k=1}^{K} \log \mathbb{P}\big(T_{ijk}\big) \;-\; \frac{\lambda}{2}\, \|S\|_F^2$$

Step 4: drop unreliable features



✕ dropped

Every score has a confidence interval (asymptotic normality of the MLE). More comparisons, narrower intervals.

Reliability of feature k

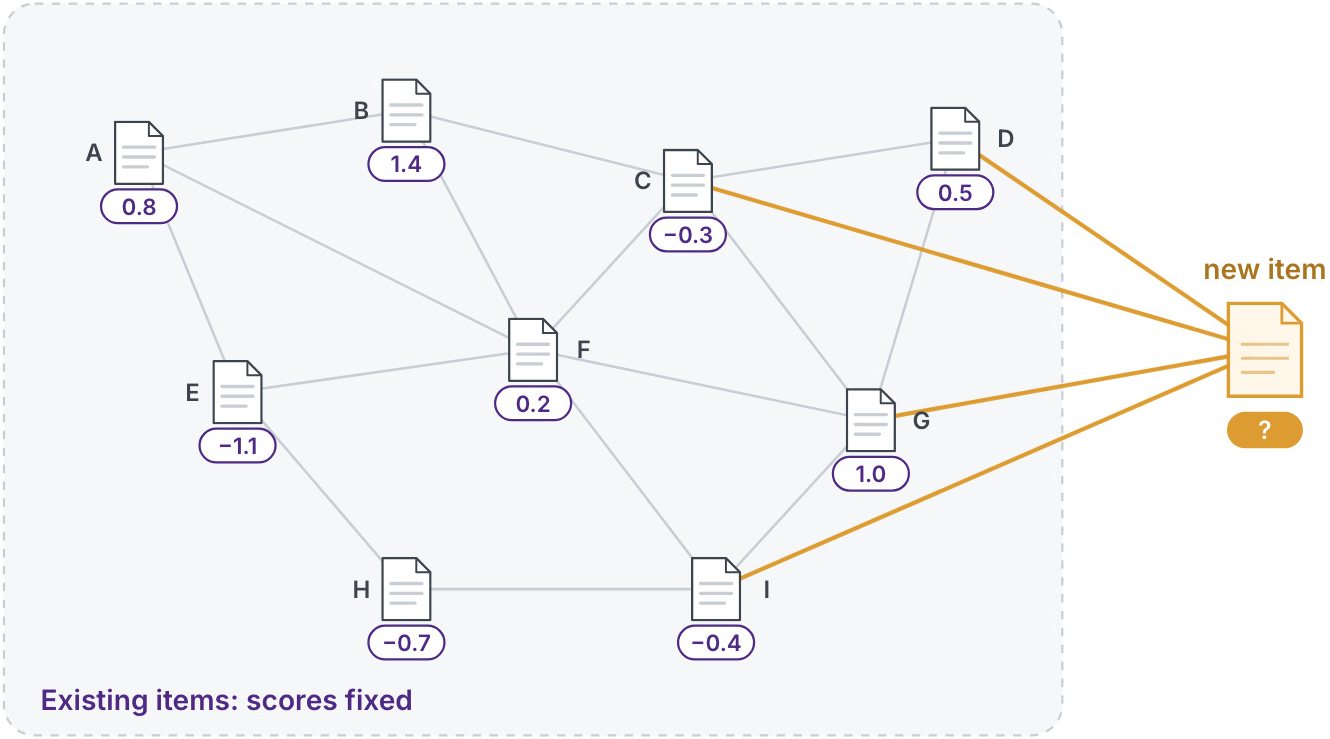
$$\rho_k = \frac{\widehat{\mathrm{Var}}(\hat{S}_{\cdot k}) - \nu_k}{\widehat{\mathrm{Var}}(\hat{S}_{\cdot k})}$$

$\widehat{\mathrm{Var}}(\hat{S}_{\cdot k})$: spread of the scores across items (the dots).

$\nu_k = \frac{1}{n}\sum_i \mathrm{Var}(\hat{S}_{ik})$: mean estimation variance (error bars).

Drop feature k if $\rho_k < \rho_{\min}$ (fixed in advance), then refit.

Scoring new items on the same scale



Recap: advantages of this approach

1

Leverages the strengths of LLMs.

2

Can control for some forms of LLM bias.

3

Statistics: we know how confident we are.

The estimator is well behaved

- Not a standard problem: the likelihood is non-convex, and the number of parameters grows with the number of items n.

Theorem (informal)

One feature; every pair compared at least once and a bounded number of times. With high probability, the likelihood has a local maximizer $\hat{s}$, unique near the truth, such that

$$\|\hat{s} - s^*\|_\infty \;\le\; C\sqrt{\frac{\log n}{n}}$$

Every item's score is consistent at once; the bias parameters γ, β converge at the same rate.

Empirical results

INTERPRETABLE FEATURE DISCOVERY FROM UNSTRUCTURED DATA

EMPIRICAL RESULTS

Synthetic data: cover letters with planted features

Dear Google Hiring Team,

Security work is the rare kind of engineering where the best outcome is that nothing happens. That inversion is what pulls me toward the Software Engineer role on the Information Security Engineering team: the chance to build systems whose success is measured in incidents that never occur.

My foundation is a ⟨EDUCATION⟩ in computer science from ⟨SCHOOL⟩ and ⟨EXPERIENCE YEARS⟩ years of professional software development at ⟨EMPLOYER⟩, where I have tested, maintained, and launched ⟨NUMBER OF LAUNCHES⟩ production services, writing Go alongside ⟨NUMBER OF LANGUAGES⟩ other languages.

⟨CORE UNIT 1⟩ ⟨CORE UNIT 2⟩
⟨CORE UNIT 3⟩ ⟨CORE UNIT 4⟩
⟨CORE UNIT 5⟩ ⟨CORE UNIT 6⟩

⟨NICHE UNIT 1⟩ ⟨NICHE UNIT 2⟩

I would be glad to talk in more depth about the team's challenges and where I could contribute from day one. I am fully authorized to work in Singapore and do not require visa sponsorship.

Sincerely,
⟨NAME⟩

One of 10 letter templates (from the paper). Each slot is filled according to the letter's planted features.

Items	200 cover letters
Planted	16 features relevant to the job

Ground truth known: we can check **which features each method recovers**.

EMPIRICAL RESULTS

Synthetic data: we recover all 16 planted features

Method	Cost (USD)	Features found	Planted features tracked	… and named right	Features tied to one planted feature	Leakage (lower is better)
Ours	**3.45**	**21**	**16 / 16**	**16 / 16**	**16 / 21**	**0.090**
Single LLM GPT-5.4 reads all letters, proposes features	0.26	21	**13 / 16**	13 / 16	13 / 21	0.109
Single LLM, many scorers same, scores averaged over several calls	**3.42**	21	**13 / 16**	13 / 16	14 / 21	0.118
Many LLMs each reads a subset; features merged	2.35	23	**12 / 16**	11 / 16	11 / 23	0.110
Many LLMs, many scorers	3.27	23	**12 / 16**	11 / 16	11 / 23	0.113

Tracked: a discovered feature's scores correlate with the planted feature. Named right: that feature's name also matches it. Leakage: correlation with the other planted features. Cost: batch-API prices.

Same cost, more scorers, still 13: the misses are a discovery failure, not a scoring failure.

EMPIRICAL RESULTS

Real data: which arguments change minds?

REDDIT R/CHANGEMYVIEW · ORIGINAL POST

CMV: Remote work is better than office work for most knowledge workers.

No commute, fewer interruptions, more focus [...] I just don't see what the office adds.

Reply A — Δ awarded

Think about junior employees. In my first job, I learned most of what I know by overhearing senior colleagues [...]

Reply B — no Δ

That's just wrong. Offices exist for a reason, and most people who want remote work [...]

Which reply changed the original poster's mind?

Illustrative example, shortened. Data: 1,200 pairs of replies (Tan et al., 2016).

Which characteristics of a reply can predict the original poster's decision?

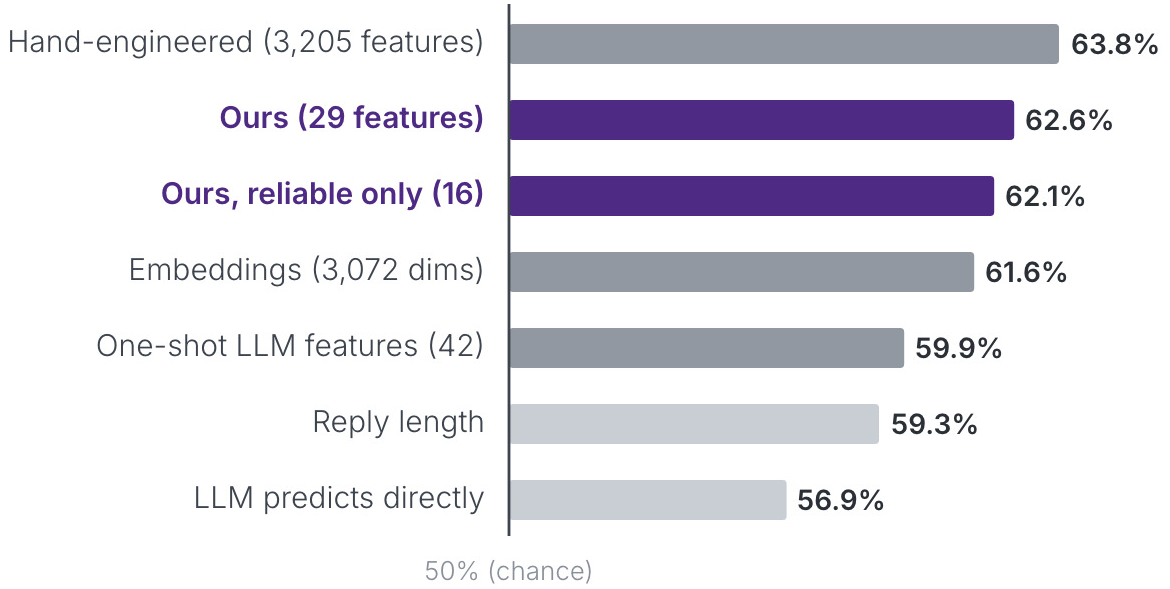
EMPIRICAL RESULTS

Real data: interpretable and predictive

Discovered features (examples)

- Concrete illustration
- Analogical reasoning
- Qualified rebuttal
- Concessionary rebuttal
- Structural explanation
- Confrontational tone

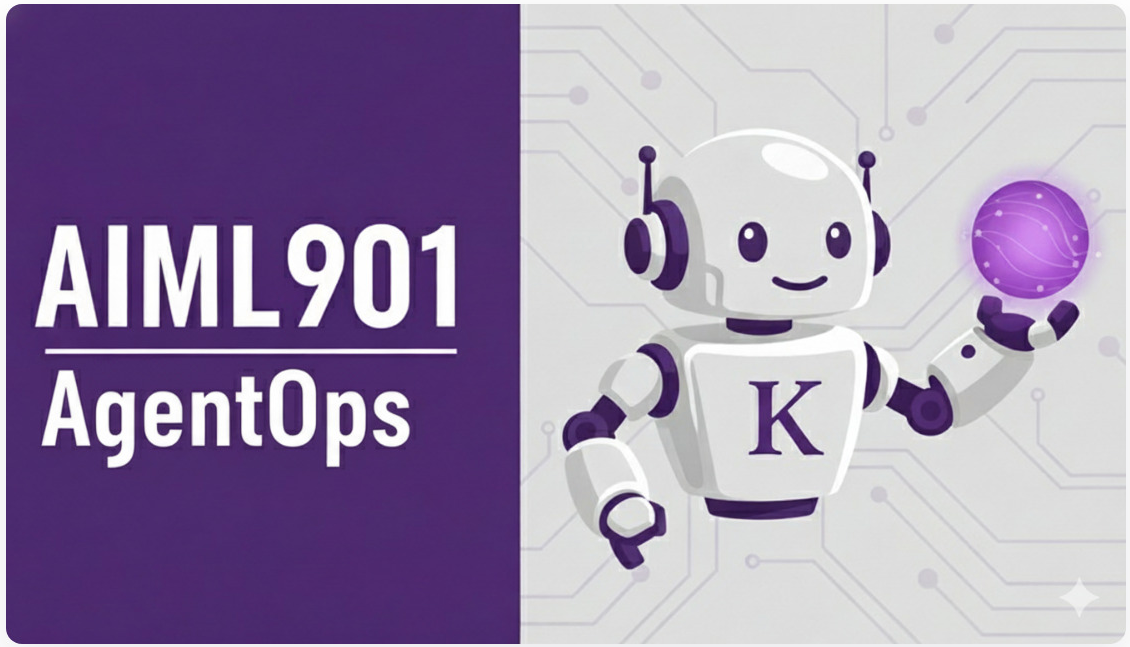
Held-out accuracy



Hand-engineered: Tan et al. (2016), "Winning arguments," WWW.
Ours: held-out replies scored pointwise by an LLM on the discovered features.

EMPIRICAL RESULTS

Anecdotal evidence



I used this approach to help grade my MBA students' project pitches.

Food for thought

1

LLMs × statistics

Structure and guarantees for tasks that were previously completely unstructured.

2

Choice models × LLMs

In my opinion, a hugely promising direction.

3

Automated feature discovery

We only scratched the surface: it could be a huge thing.

Thank you

Baptiste Arnaudo · Amine Bennouna · Sébastien Martin



arxiv.org/abs/2610.38082

BACKUP

Auditing an LLM judge, on real essays

- The judge picks the better essay **and justifies it** with features.
  $\mathbb{P}(i \text{ wins}) = \sigma\big(\sum_k \Delta_k + \beta\big)$: the overall score decomposes into named features.
- Two essay corpora with human grades (ELLIPSE, ASAP 2.0), four LLM judges.
- Judges **agree on which** features matter, **not on how much**.
- Fitted totals track human grades.
  Spearman 0.89–0.92 (ELLIPSE), 0.77–0.86 (ASAP 2.0).