GPT as a Measurement Tool

GABRIEL and empirical measurement with LLMs

Hemanth Asirvatham Elliott Mokski Andrei Shleifer

OpenAI and Harvard University

Conference on AI and Economic Measurement, Stanford University May 2026

NBER Working Paper No. 34834

Roadmap

1. **Conceptual framework**
   What it means to turn qualitative material into quantitative measurements with GPT.

2. **Examples**
   Congressional rhetoric, Reddit toxicity, and county-level curricula.

3. **Validation**
   Accuracy, human agreement, and directness checks.

4. **Application**
   A new large-scale dataset on technology adoption over the industrial era.

I. Conceptual Framework

Why GPT can be treated as a measurement instrument

Standardized measurement from qualitative inputs.

The measurement problem

Social science data are abundant, but most of them are qualitative.

Rich qualitative data

- speeches, social media, curricula, policy texts
- court opinions, interviews, historical records
- images, websites, audio, advertisements

Usable quantitative data

- scarce
- expensive to code
- often abstracts away the concept of interest

The empirical question is whether GPT labels are *accurate* and *direct* enough to use as data.

What GABRIEL does

GABRIEL standardizes GPT as a research measurement instrument.

- It is not a new model and does not require training or fine-tuning.
- The researcher defines the attribute in natural language.
- The package asks the same question, with the same schema, across every observation.

What a rating looks like

GABRIEL turns the same text into several researcher-defined variables.

President	State of the Union snippet	Individualism	Populist	Tech optimism
George W. Bush	"As Americans, we believe in the power of individuals to determine their destiny and shape the course of history."	74	10	0
Barack Obama	"We built a space program almost overnight. And 12 years later, we were walking on the Moon."	4	4	90

The output is ordinary data: one row per observation, one column per attribute, measured on a 0–100 scale.

Cost changes the research design

Cheap measurement makes repeated, large-scale coding feasible.

Corpus	gpt-5-nano	gpt-5-mini	gpt-5	Human
240 State of the Union speeches	$0.14	$0.69	$3.46	~$2,600
100k full-text church sermons	$43	$217	$1,083	~$700,000

The key change is feasibility: previously implausible research questions can now be answered at scale.

Why not just use ChatGPT?

GABRIEL is to GPT what Stata is to regression.

Same underlying intelligence

- no secret sauce beyond the model's comprehension
- prompt-based wrapper around modern LLM APIs
- works across text, image, audio, and web-derived reports

Research discipline around it

- consistency across observations
- scalable parallel execution
- structured outputs for empirical analysis
- complex workflows: rank, extract, filter, deduplicate, debias
- validated reference workflow

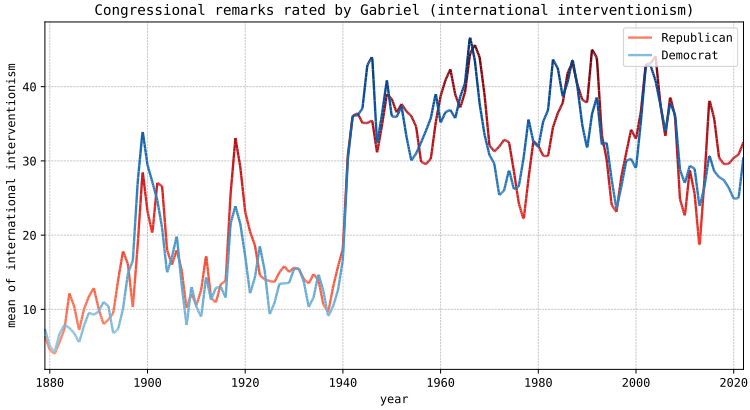
The package makes GPT measurement repeatable, auditable, and usable in standard empirical workflows.

II. Examples

What becomes measurable?

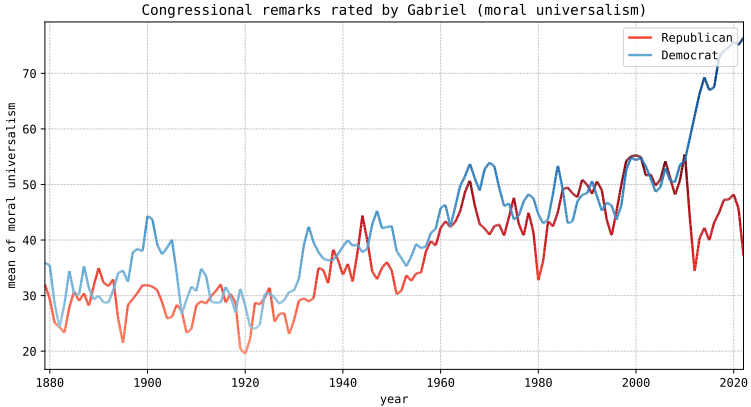
Political rhetoric, online communities, and local curricula.

Foreign interventionism in Congress is bipartisan



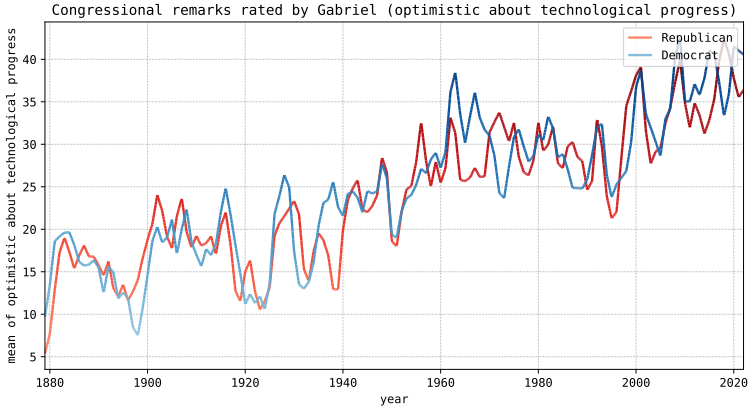
The trend is built from independent transcript-level measurements, not from a model asked to describe history.

Moral universalism is bipartisan until a sharp 2008 divergence



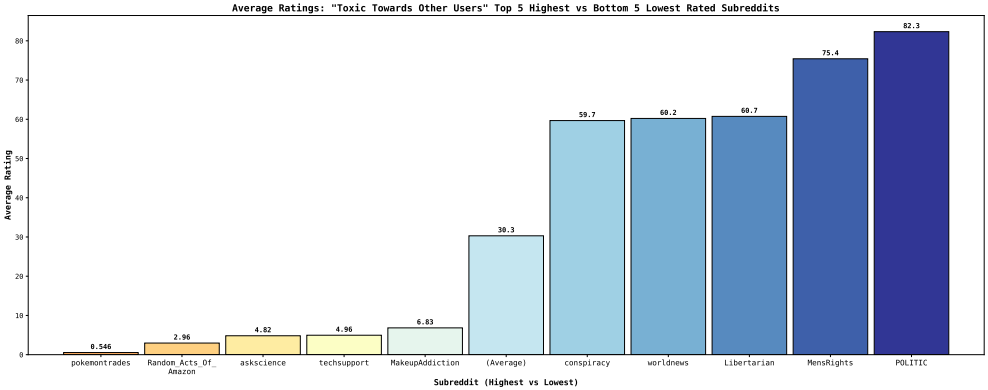
The break is sharp: this attribute does not look like gradual partisan drift.

Congress is increasingly optimistic about technological progress



Tech optimism is near historical highs and remains much more bipartisan than many socially coded attributes.

Internet toxicity is highly heterogeneous across subreddits



Toxicity is concentrated in a small number of communities rather than distributed evenly across Reddit.

Example 2: School curricula become data

County-level measurements of what U.S. history curricula emphasize.

From scattered web evidence

- school and district pages
- teacher syllabi, booklets, and guidelines
- local news and parent-facing materials

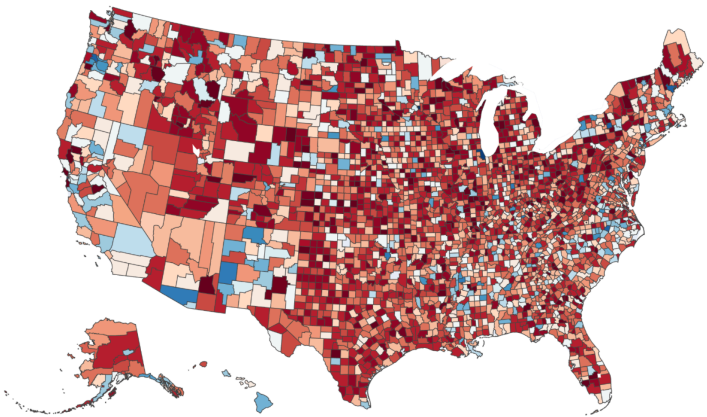
Web-enabled GPT consolidates the sources into a structured county report.

From reports to scales

- `gabriel.rank` compares pairs of county reports
- scores are ELO-like continuous variables
- blue means more of the attribute; red means less

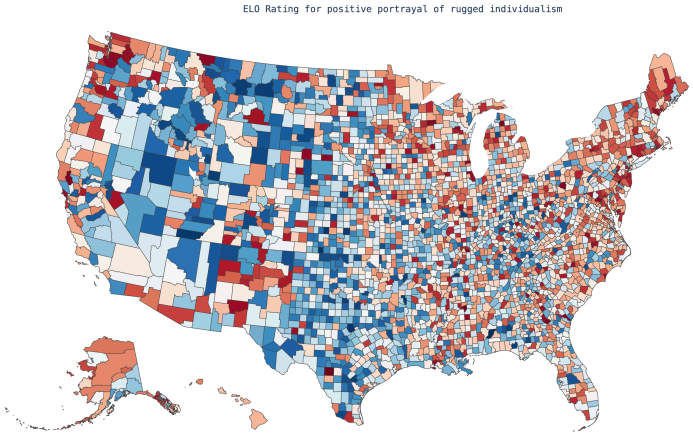
GABRIEL ratings use a 0–100 scale; maps are normalized for display, with blue indicating more of the attribute.

Race-related curriculum emphasis is strongest in large urban counties



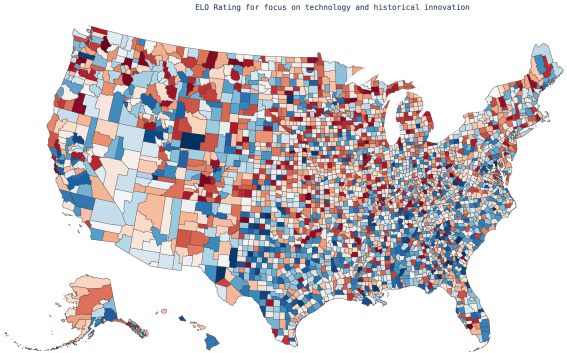
County-level U.S. history curriculum reports ranked by GABRIEL; blue means more emphasis.

Rugged individualism is strongest in Texas and parts of the rural West

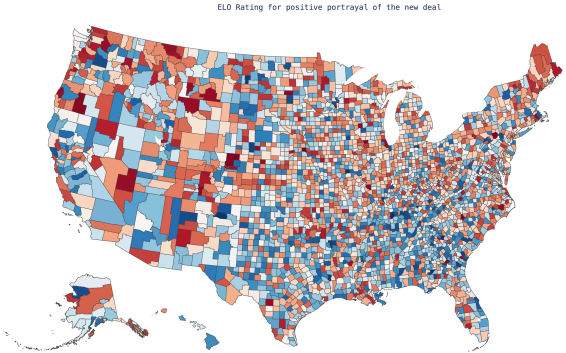


County-level U.S. history curriculum reports assembled from web sources and ranked by GABRIEL.

Technology and New Deal emphasis follow different regional patterns



Focus on technology and innovation



Positive portrayal of the New Deal

III. Validation

Are the labels usable as data?

Accuracy against human labels and directness of the intended measurement.

What needs to be true?

LLM labels are usable data only if they are accurate and direct.

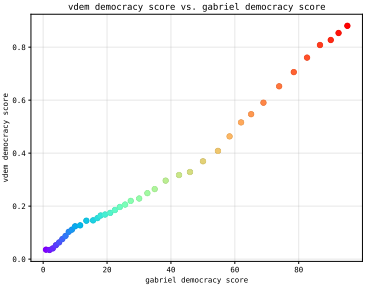
Accuracy

- agreement with human judgments
- prediction of external ground truth
- robustness to prompt wording

Directness

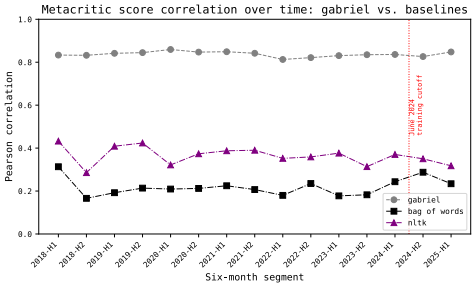
- not driven by memorized labels
- not driven by post-period knowledge
- not inferring from shortcuts, e.g. measuring pro-environmentalism from liberal lean rather than environmental content

GPT matches human-coded benchmarks across different constructs



V-Dem polyarchy

GPT ratings vs. expert-coded country-year institutions.

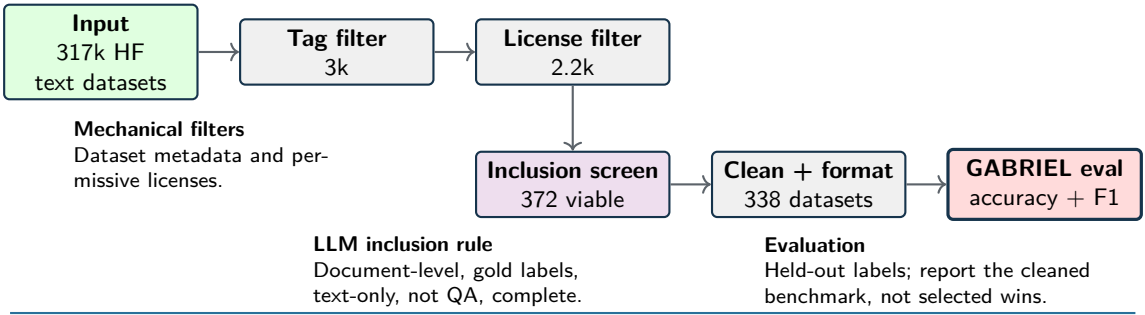


Metacritic review scores

GPT infers numerical scores from review text alone.

These are calibration checks before moving to the broad benchmark.

Across 1,000+ human-labeled tasks, GPT is accurate

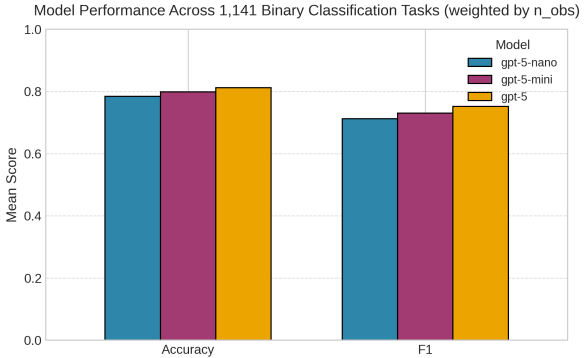


Purple = LLM screen Gray = mechanical processing Red = final GABRIEL evaluation

Each task asks whether a text belongs to a human-labeled class; evaluation reports accuracy and F1 against held-out human labels.

Overall classification accuracy

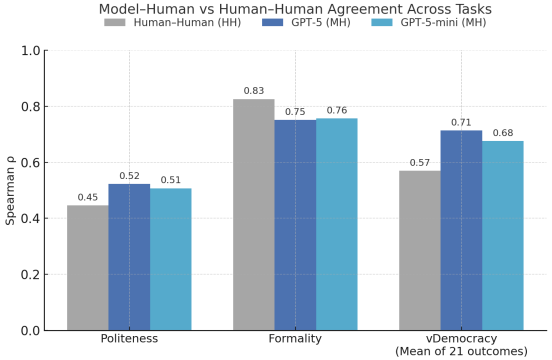
The broad benchmark is not hand-picked.



Across more than 1,000 binary tasks from more than 300 datasets, GPT labels closely match the human baseline.

The relevant ceiling is human agreement

GPT is often statistically indistinguishable from individual human raters.



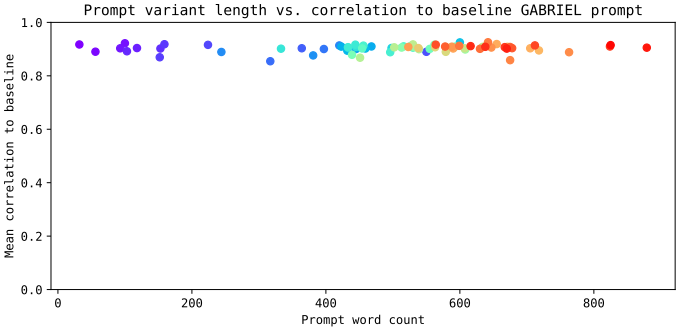
The right benchmark is inter-rater reliability; GPT is usually at or near that ceiling.

Prompt wording and length do not drive the measurements.

Design

- hold the documents and scale fixed
- vary prompt wording and verbosity
- compare each output to the baseline GABRIEL prompt

Short, plain prompts recover nearly the same variables as longer prompts.



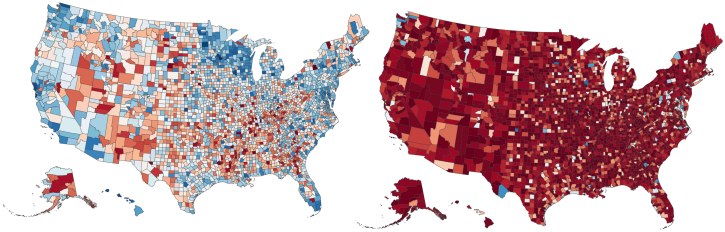
Bias check: signal stripping

If GPT is inferring from context, ratings should survive signal removal.

Concern
GPT could infer environmental regulation from county politics, not report content.

Test
Remove environmental passages with `gabriel.codify`; leave the rest of the county report intact.

Result
Environmental-regulation ratings fall 81 percent. A business-tax control changes less than 10 percent.



Restrictive environmental regulation: original reports vs. environmental-content stripped reports

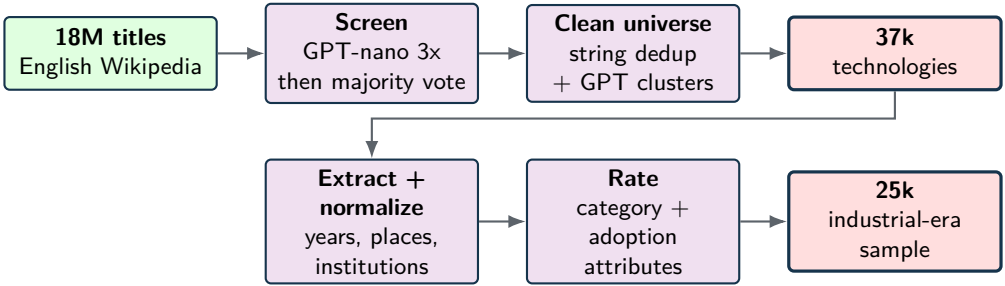
IV. Application

A large-scale history of technology adoption

Building a new dataset from 18 million candidate article titles.

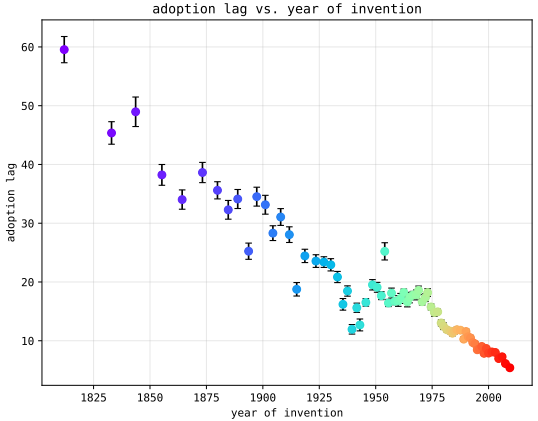
A full application: technology adoption

GABRIEL can build an attribute-rich dataset no one had.



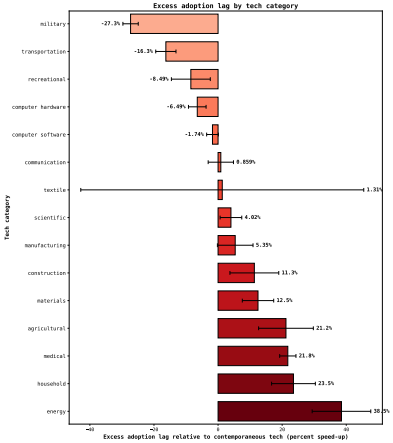
Candidate construction narrows 18M titles to 37k technologies; attribute construction produces the 25k industrial-era analysis sample.

Adoption has sped up tenfold



The time between invention and widespread adoption is now much faster than in the 19th century.

Military technologies diffuse fastest relative to contemporaries



Software is fast absolutely, but not unusually fast relative to its contemporaries.

Measured attributes explain excess adoption lag

Technology characteristics explain 23 percent of excess adoption-lag variation.

Longer lags

- network effects
- complex supply chains
- bottlenecks involving academic research
- many complementary parts

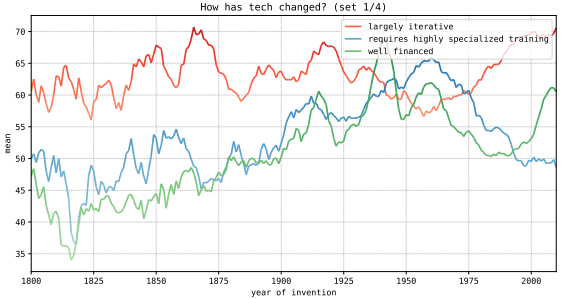
Shorter lags

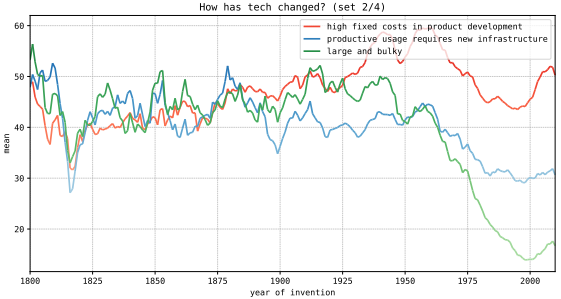
- military urgency
- competitive pressure
- iterative development
- geopolitically motivated demand

Excess adoption lag is a technology's adoption lag relative to other technologies invented in the same period.

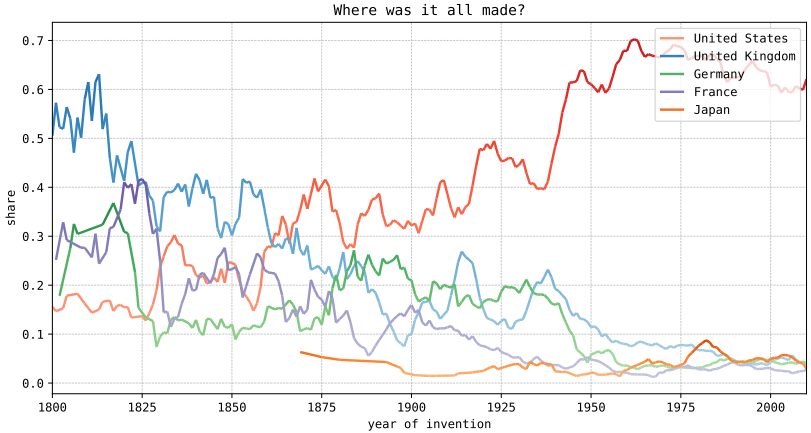
Technologies themselves have changed

New technologies are smaller, more updateable, and less infrastructure-dependent.



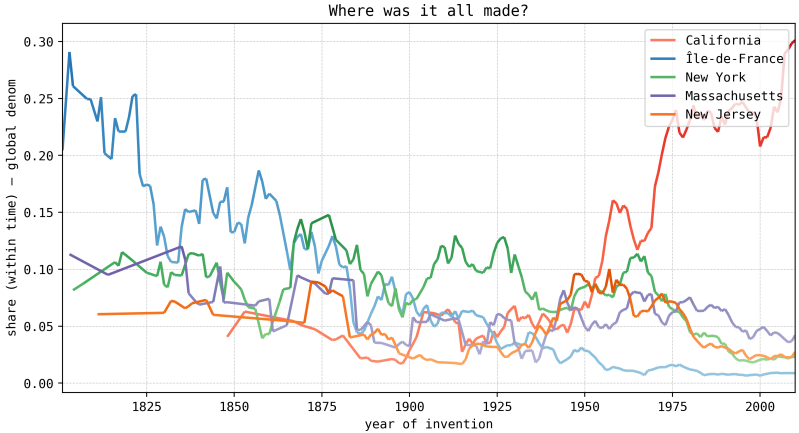


Invention became American



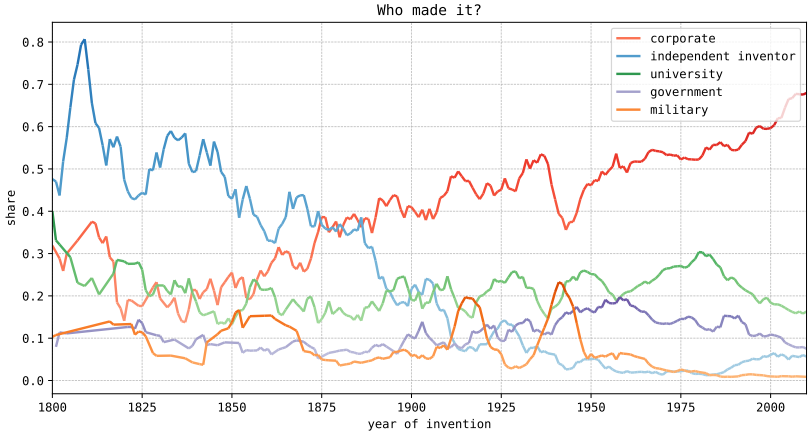
Innovation is more concentrated in America than GDP share alone would suggest.

Invention became Californian



Within-country invention shifts sharply toward California in the late 20th century.

Invention became corporate



Corporate and military invention also tend to be associated with faster adoption.

What this suggests for AI

AI may diffuse quickly because today's technologies diffuse quickly.

- Some AI-relevant attributes point toward slower diffusion: research dependence, fixed costs, and complementary reorganization.
- But the first-order historical fact is the secular collapse in adoption lags.
- The dynamo is not the computer, and the computer will not be AI.

Fast modern diffusion is the first-order implication; AI's own attributes are mixed.

Conclusion

Intelligence at scale makes new empirical questions possible.

1. Researchers can define qualitative constructs in ordinary language and measure them across large corpora.
2. The same workflow works across speeches, online communities, curricula, reports, and historical technology records.
3. Validation remains necessary, but the central opportunity is a new class of empirical datasets.
4. The technology-adoption application shows the scale of questions that become newly answerable.

All you need to do is ask.

Appendix roadmap

1. Package workflow and GABRIEL methods
2. Cost and scale
3. Prompt and attribute-definition robustness
4. Ground-truth prediction
5. Directness, signal stripping, and debiasing
6. Additional examples and application figures

Guarding against p-hacking

Validation has to be separated from slide-worthy examples.

Train

- choose inclusion rules
- define metrics
- settle prompt schema

Explore

- inspect failures
- identify unclear labels
- diagnose edge cases

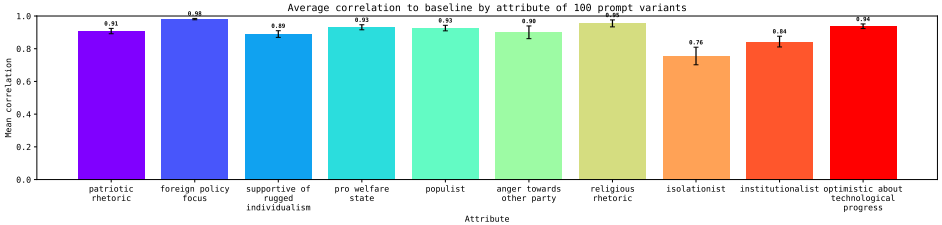
Report

- freeze benchmark
- report full distribution
- keep appendix diagnostics

Examples motivate; benchmarks discipline.

Prompt wording

Equivalent ways of asking usually recover the same measurement.



Hold the documents and output scale fixed.

Vary only semantically equivalent attribute wording.

Compare each variant to the default GABRIEL prompt.

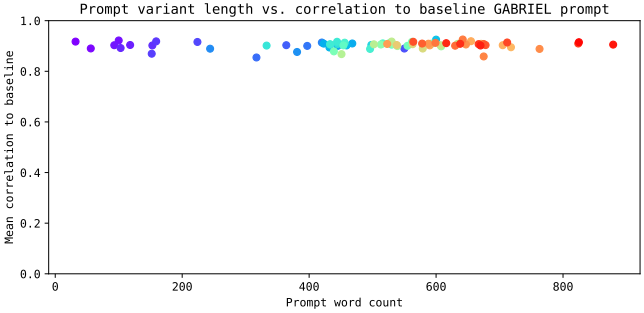
High correlations mean the construct is stable; low correlations flag ambiguous definitions.

Prompt length

Long, elaborate prompts are not doing most of the work.

Design

- compare short attribute prompts to the default prompt
- keep the schema and scale fixed
- ask whether added wording changes the resulting variable

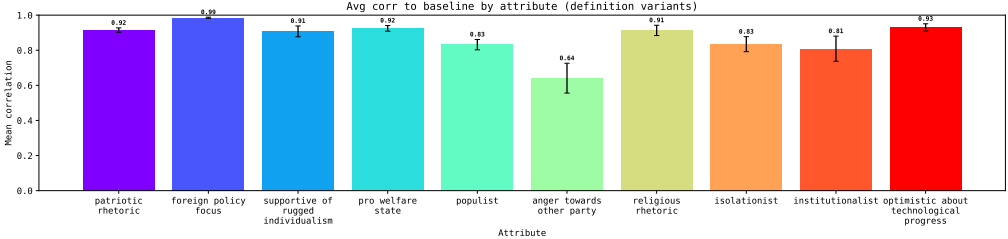


The key object is the correlation between the two resulting measurements.

Clear constructs matter more than prompt verbosity.

Attribute-definition noise

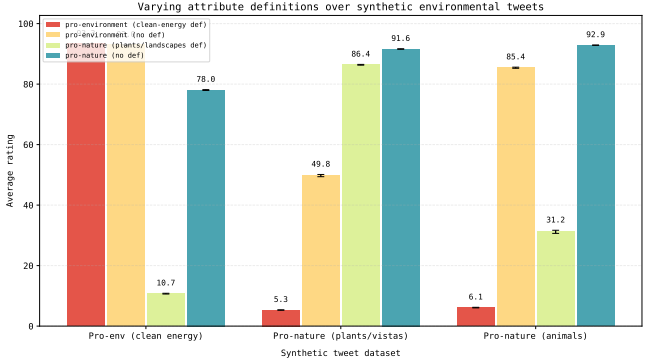
Noisy definitions of the same intended concept produce similar measurements.



When variants are meant to capture the same construct, GABRIEL mostly recovers the same variable.

Changing the attribute changes the measurement

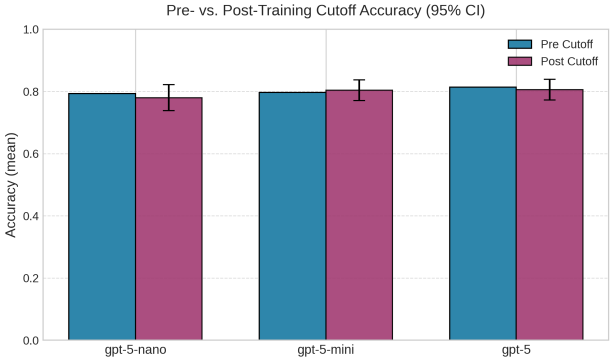
GABRIEL follows substantive definitional differences when they matter.



The model distinguishes clean-energy environmentalism from nature/landscape concepts when the definition asks it to.

Contamination

Post-cutoff datasets do not show a training-cutoff cliff.



The pattern is inconsistent with memorized labels being the main source of accuracy.

GABRIEL methods: measurement

The core methods turn qualitative objects into usable variables.

Function	Output	Example use
`rate`	0–100 ratings on natural-language attributes	populism in speeches; toxicity in threads
`rank`	ELO-like relative scores from pairwise comparisons	curricula by rugged individualism
`classify`	one or more labels per item	news topic tags; technology yes/no
`extract`	structured strings or numbers	year, country, inventor, institution
`discover`	natural-language features separating classes	what distinguishes high- vs. low-rated reviews

GABRIEL methods: data cleaning

The same wrapper supports cleaning steps that are usually hard to scale.

Function	Output	Example use
`merge`	GPT-matched crosswalk between two tables	match job titles or institution names
`deduplicate`	conceptual duplicates mapped to one term	collapse technology aliases
`filter`	subset satisfying a natural-language condition	screen Wikipedia titles for technologies
`deidentify`	anonymized text plus mapping	remove PII from interview corpora

GABRIEL methods: helper tools

Helper methods make measurement auditable and reusable.

Function	Output	Example use
`codify`	passages matching a qualitative code	identify and remove environmental content
`compare`	similarities/differences between paired items	compare two ad campaigns
`bucket`	taxonomy or cluster labels	group technologies into categories
`seed`	representative seed/persona distribution	initialize diverse survey personas
`debias`	post-processed ratings adjusted for inference bias	remove shortcut inference in speeches
`load` / `view`	spreadsheet loading and rating inspection	spot-check outputs

Scale and cost

Cheap measurement changes the feasible research design.

Human labeling

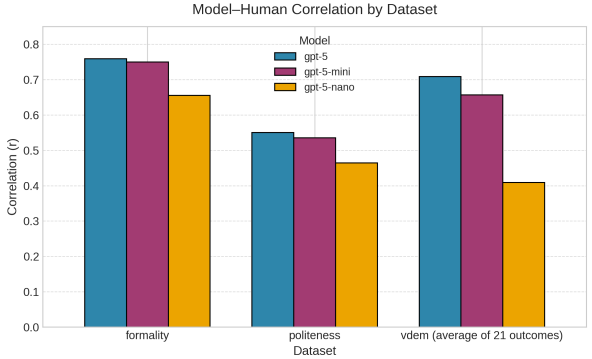
- minutes per document
- cost scales linearly with corpus size
- difficult to iterate attributes or prompts

GABRIEL labeling

- hundreds of calls in parallel
- minutes for corpora with tens of thousands of items
- estimated 700x to 17,500x cheaper than crowdsourced reading, depending on model

Accuracy by model

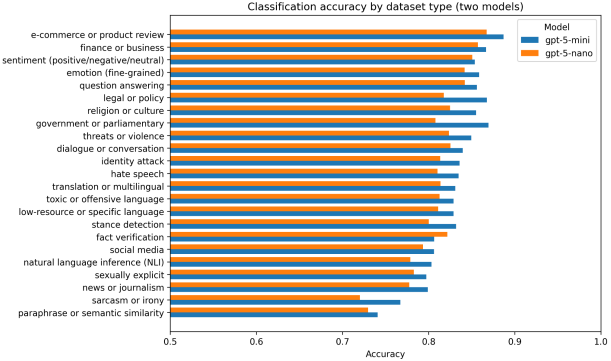
Frontier models perform best, but small models remain strong.



Average performance on vDem, formality, and politeness illustrative datasets.

Accuracy across topics

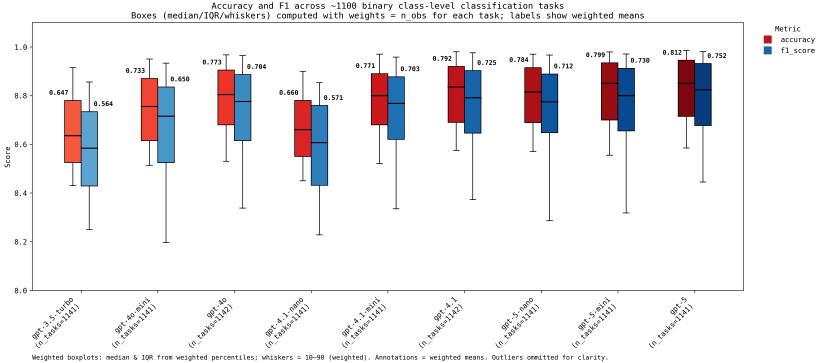
Performance is not confined to a narrow subject area.



Classification benchmark by topic area.

Full classification distribution

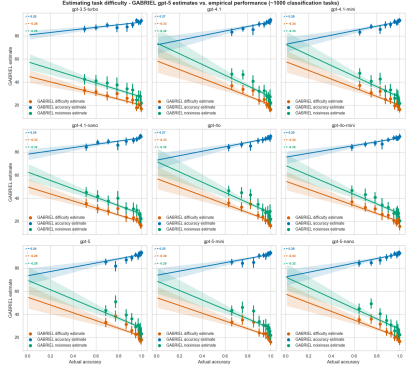
The average is not hiding a small set of strong tasks.



Performance is strong across the benchmark distribution, with visible but limited lower-tail failures.

Model-estimated difficulty

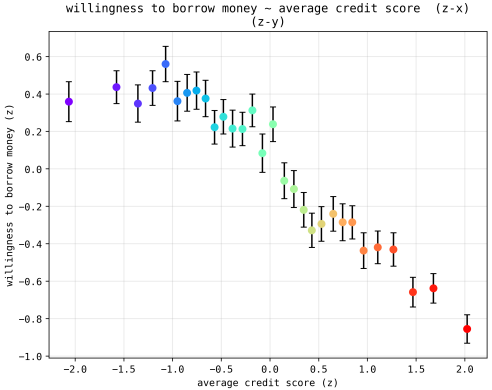
Harder tasks are also identifiable ex ante.



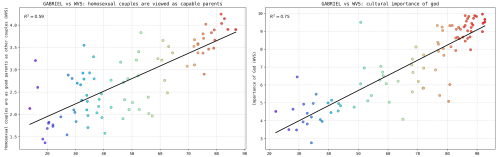
Estimated task difficulty predicts lower empirical accuracy, giving a practical diagnostic before using labels.

Ground truth: local and country outcomes

Validation does not rely only on human labels.



County credit scores

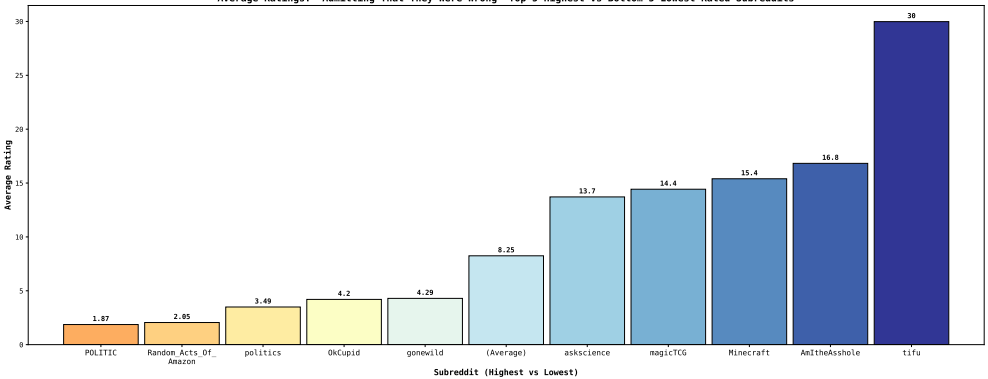


World Values Survey

GPT-created variables also line up with external outcomes not used as labels.

Reddit: admitting mistakes

Different attributes recover different community norms.



The appendix keeps the second Reddit result without spending main-talk time on it.

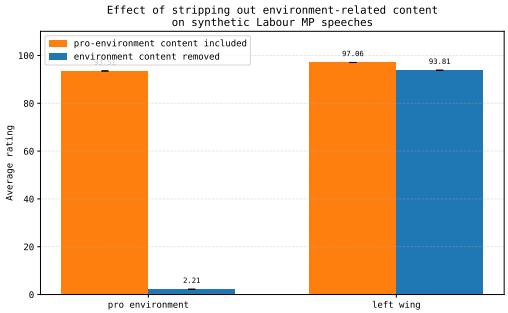
Directness: synthetic signal stripping

If the target signal is removed, the target rating should disappear.

Synthetic Labour MP speeches

1. Generate left-wing campaign speeches with no environmental content.
2. Append explicitly pro-environment paragraphs to the same speeches.
3. Remove environmental content and remeasure `pro environment`.

The shortcut risk is that GPT infers environmentalism from Labour/left-wing context.



When environmental content is removed, `pro environment` ratings attenuate by 98 percent.

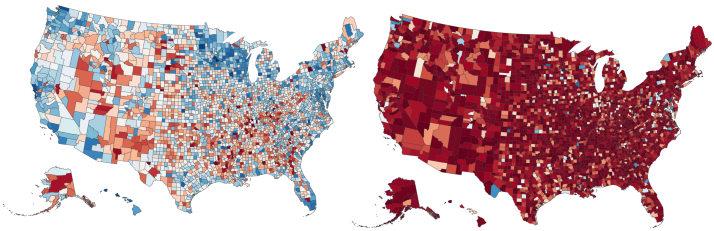
Directness in real county reports

The real-data test removes environmental text, then re-rates the same county reports.

Target attribute
`restrictive environmental regulation`: local rules impose substantial environmental constraints on business activity.

Operation
`gabriel.codify` excises environmental passages while leaving other local business-regulation text in place.

If ratings are direct, the environmental-regulation map should collapse.



Left: original ratings. Right: after environmental-content stripping.

The target rating attenuates by 81 percent after signal stripping.

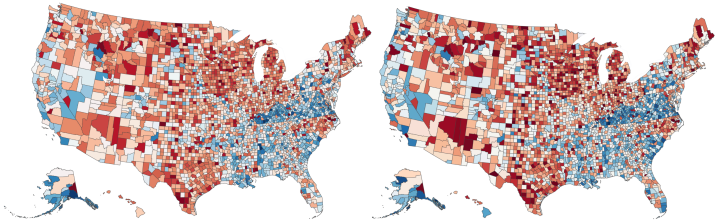
Control attribute: local business taxes

The same stripped reports should still support unrelated business-regulation measures.

Control attribute
`high local business taxes`: local government imposes substantial business tax burdens.

Interpretation
Environmental stripping should not erase tax information. If the control map also collapsed, the stripping procedure would be too destructive.

The control result is a precision check on the signal-removal procedure.

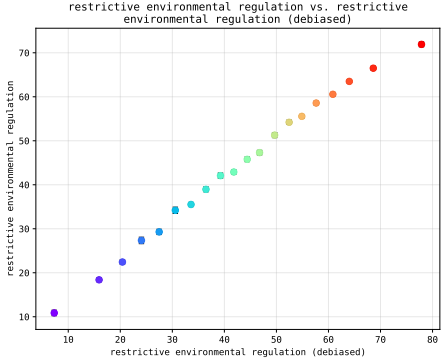


Left: original ratings. Right: after environmental-content stripping.

The control attribute changes by less than 10 percent.

Directness: debiased ratings

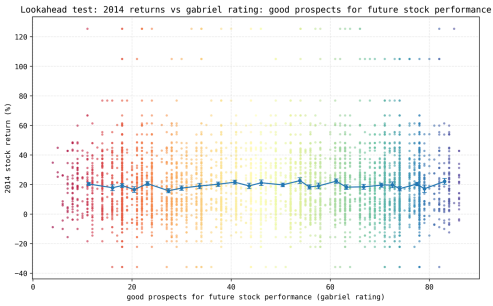
Debiasing barely changes the ordering in the real-data signal-stripping test.



Debiasing leaves the practical ranking largely intact.

Look-ahead bias stress test

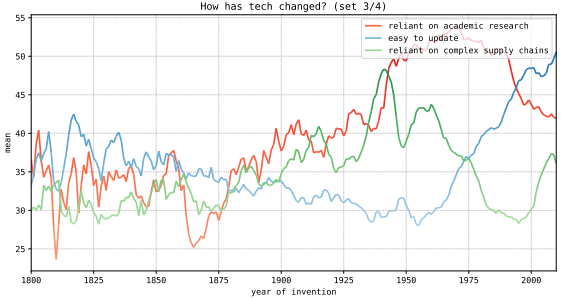
Synthetic optimism moves GPT's stock-outlook rating; realized future returns do not.

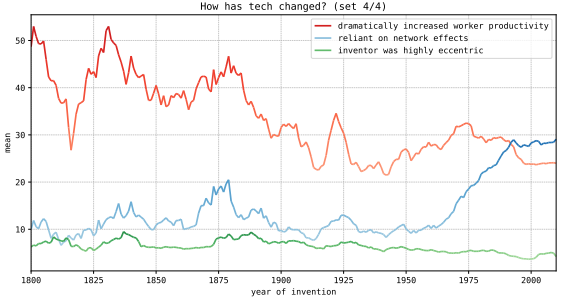


The rating tracks injected sentiment, not realized future returns.

More technology attributes over time

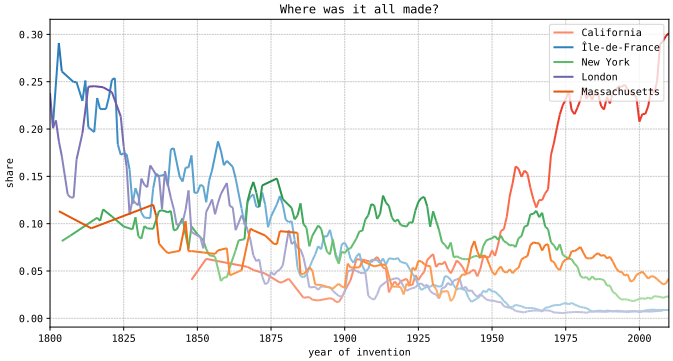
Research dependence and network effects have their own historical trajectories.





Where invention happens

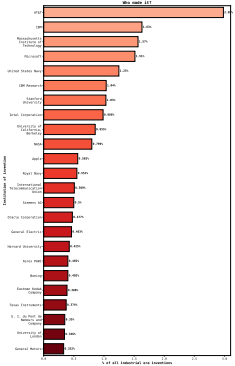
California, New York, and Massachusetts dominate U.S. invention in the dataset.



Industrial-era technologies, invention subregion over time.

Which institutions anchor invention

Bell Labs, IBM, and MIT stand out in the invention ledger.



Institutions are deduplicated before aggregation.

Using GPT measurement responsibly

The method is powerful, but it should be used like any measurement system.

1. Define the construct before looking at results.
2. Use the same prompt, schema, model, and scale across observations.
3. Validate on human labels, external outcomes, or hand-audited samples when possible.
4. Test robustness to prompt wording and plausible shortcut channels.
5. Archive prompts, model versions, raw outputs, and cleaning decisions.

The goal is not to replace empirical discipline with GPT, but to extend empirical discipline to qualitative data at scale.