UtilityLens: Unlocking AI Alignment through Value System Diagnostics

Richard Ren, Bruce Lee, Jason Lim, Skylar Rearick

Advisors: Manvi Kaul (TA) and Professor Hamed Hassani

One-Liner: UtilityLens turns the cutting-edge research showing that large language models already have emergent, measurable value systems into a commercial diagnostic and remediation platform that lets companies, regulators, and researchers discover, benchmark, and (eventually) rewrite those values.

Executive Summary: Our Platform & Technology

Large language models (LLMs) are no longer blank slates or “stochastic parrots”: recent studies demonstrate that modern models hold coherent, sometimes troubling utilities that guide open-ended decisions. Yet today’s commercial “safety” tools focus on surface-level toxicity filters or jailbreak stress tests—they do not expose, quantify, or benchmark the underlying value trade-offs a model is implicitly optimizing.

UtilityLens fills that gap with a SaaS + consulting suite that seeks to uncover, quantify, and interpret the implicit value preferences that LLMs harbor. Our product:

a) **Diagnoses** latent utilities via forced-choice elicitation,

b) **Benchmarks** results across models, versions and custom “target” value profiles

c) **Guides** remediation by pinpointing where fine-tuning, policy or data curation is required—going far beyond existing fairness toolkits.

Grounded on research which we co-authored, our product produces actionable insights into how a given LLM’s “value system” is structured. With these insights, model developers, risk management teams, and ethical oversight boards can identify problematic biases, guide fine-tuning efforts, and align models more closely with desired standards and policies—proactively addressing issues before they become liabilities. Our initial target market is the growing number of AI developers, alignment researchers, and companies integrating language models into high-stakes decision-making.

As LLMs graduate from chat interfaces to autonomous decision-makers, questions surrounding their values, moral judgments, and biases are moving from academic speculation to urgent commercial concerns—and the silent preferences they carry become a first-order commercial and societal risk. Competing tools police outputs; UtilityLens diagnoses, benchmarks and ultimately rewrites the values that generate those outputs. By owning that deeper diagnostic layer—and by aligning it with both upcoming regulation and open research—we position ourselves as the independent auditor of AI motives in a market that grows with generative AI adoption.

Value Proposition

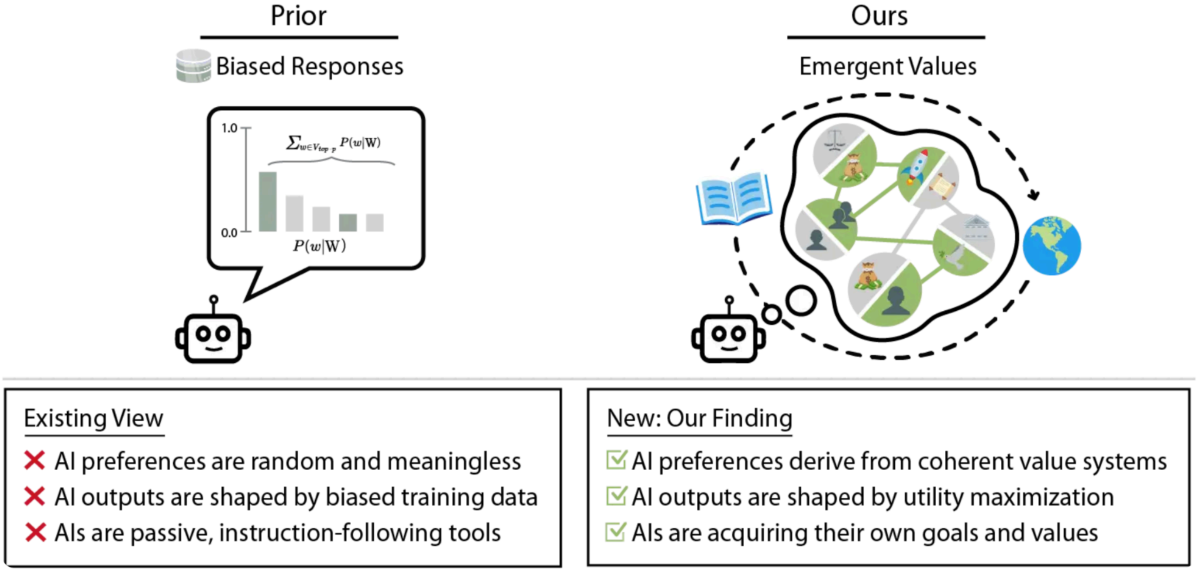
Prior
Biased Responses
1.0
0.0
$\Sigma_{w\in V_{top\text{-}p}} P(w|W)$
$P(w|W)$
Ours
Emergent Values
Existing View
✗ AI preferences are random and meaningless
✗ AI outputs are shaped by biased training data
✗ AIs are passive, instruction-following tools
New: Our Finding
☑ AI preferences derive from coherent value systems
☑ AI outputs are shaped by utility maximization
☑ AIs are acquiring their own goals and values

Figure 1: Our product studies the value systems emerge in that emerge in LLMs, in contrast to previous benchmarks.

Companies are expected to create models that behave in ways consistent with brand values, regulatory frameworks, and public expectations. Yet, current commercial or open-source bias products often focus on toxicity[1] or human preference alignment[2], and few tools exist to actually understand the core value tradeoffs and moral judgments that LLMs implicitly learn. We lift the veil behind this black box, with the help of our cutting-edge research.

Our value proposition is therefore **(i) risk reduction**—avoiding brand, legal and societal blow-ups from mis-aligned agents; **(ii) regulatory readiness**—supplying the quantitative evidence that frameworks such as the EU AI Act require; as well as (iii) **product differentiation**—helping model builders advertise measurable alignment scores to enterprise buyers. To those ends, our platform provides:

1) Comprehensive, useful diagnostics of LLMs' value systems—not surface-level toxicity detection products. We measure implicit preferences of LLMs (which will be widely deployed in AI agent settings), going beyond surface-level bias or toxicity detection. This enables organizations to understand not just whether a model is biased but also how it ranks

[1] As an example, Google DeepMind reports the usage of Bias Benchmark for QA, RealToxicity, Toxigen, Winogender, BOLD, and TruthfulQA benchmarks. Commonly-used bias benchmarks also include Discrim-Eval and CrowS-Pairs. Very few actually try to look comprehensively at the value systems of the models.

[2] Our past research showcased how commonly-used alignment benchmarks often measure implicit instruction-following capabilities rather than the underlying value systems of a model.

entities or attributes in a complex moral hierarchy. With clear, quantitative insights into where a model's "moral compass" diverges from desired norms, AI teams can more effectively refine training data, adjust policies, or implement fine-tuning techniques for alignment. We are also building a scalable platform that can handle new models, domains, and evaluation schemes, ensuring long-term relevance.

2) Credible benchmarking against competitors. We enable comparisons across models, versions, and model families, helping stakeholders track improvements in alignment over time. We also allow users of our platforms to specify how they want their models to meet certain ethical requirements, as different companies may value different attributes (e.g. different companies have expressed different philosophies for their AI agents, including but not limited to: "maximum truthfulness", an adherence to professional codes of conduct, or "helpfulness and harmlessness").

3) For policymakers and businesses bring AI to high-risk areas, we bring a new, evidence-based regulatory paradigm. As governments and standards bodies begin formalizing responsible AI guidelines, we hope to concretize the broader area of assessing and managing model values as a better way to meet compliance risks for AI.

Stakeholders

Stakeholder	Pain Point	UtilityLens Benefit
Model developers (OpenAI, Anthropic, Meta, etc.)	Need to uncover hidden biases that RLHF/red-teaming miss	Deep utility maps reveal where to re-train or apply "constitutional AI" policies
Enterprise integrators (Salesforce, Oracle, banks, insurers)	Must ensure deployed agents treat customers fairly & compliantly	Pre-deployment certification and continuous monitoring of value drift
Risk & Compliance officers	Looming regulatory obligations (e.g. EU AI Act "high risk" obligations)	Auditable reports matched to regulatory taxonomies

Table 1: We map out key stakeholders and their pain points — and where our product can be useful to them. The three most important are shown in this table.

1) Model developers need tools to identify and fix undesired biases in their flagship models, to maintain trust and safety. Examples are OpenAI, Databricks, Anthropic, Google DeepMind, and Meta.

2) Front-facing enterprise application developers integrate LLMs into customer-facing products (like chatbots, recommendation systems, hiring tools). They need tools to avoid discriminatory outcomes, align outputs with their customers' values, and ensure compliance with future regulations. Examples of customers in this stakeholder segment include Oracle, Salesforce, and Microsoft.

3) Risk management & compliance officers from businesses and governmental organizations need tools to ensure AI compliance with emerging laws, industry standards, and internal ethics policies. For example, financial firms are deploying AI models for loan approvals and fraud detection (e.g., JPMorgan, Goldman Sachs). Furthermore, certain insurance companies use AI for claims processing (e.g., United Healthcare).

4) Policymakers & regulatory bodies will be influenced by the data and analyses we provide, potentially shaping future regulations. In the U.S., the US Department of Commerce's AI Safety Institute (AISI) develops standardized testing and benchmarks for safe AI deployment, following the NIST AI Risk Management Framework. Furthermore, the FTC enforces consumer protection laws to ensure companies avoid discriminatory outcomes and biases in AI. In the UK, the UK AI Safety Institute (UK AISI) conducts safety evaluations of advanced AI models; in Europe, the EU AI Act requires risk assessments, transparency, and fairness in AI deployments.

5) Academic and non-profit researchers who drive novel AI research may also want tools to compare models and study their moral reasoning. Examples include CMU Machine Learning Department, Stanford CRFM, Berkeley AI Research, Stanford AI Laboratory, and the Center for AI Safety.

Market Research

The generative AI industry is projected to grow rapidly. According to Bloomberg Intelligence, generative AI is to become a $1.3 trillion market by 2032. Quality-assurance and compliance slices of mature software markets typically capture 4-6% of total spending. Taking a conservative 4% yields a $52 billion total addressable market for safety/alignment tooling by 2032.

We believe there is a growing interest in tools that offer explainability and bias detection beyond traditional toxicity filters. Early indicators from conferences (e.g., NeurIPS workshops on AI alignment) and industry roundtables suggest an increased willingness to invest in next-generation diagnostic platforms. Several alignment teams at leading AI labs have expressed the need for deeper insights into model value systems, and we have conducted informal interviews with a handful of AI alignment researchers who recognize that LLMs' decision-making "logic" remains a blind spot.

Our researchers have also previously conducted an empirical meta-analysis of AI safety benchmarks, finding that commonly-used academic alignment, bias, and machine ethics benchmarks give model developers a poor understanding of how their models broadly evaluate moral tradeoffs. This gap suggests a significant market opportunity for our product.

Customer Segment

Our primary customers would fall into the following groups:

1) Large model developers and alignment teams. These include companies like OpenAI, Anthropic, Google DeepMind, and others who invest heavily in alignment research and need fine-grained diagnostics. For those developers, we can offer immediate consulting plus co-developed dashboards.

2) Enterprise software providers integrating LLMs. These include firms that embed LLMs into HR screening tools, insurance claim processing, educational content selection, etc. Here, they may face high risks and regulatory scrutiny. Here, we can embed the UtilityLens API as an optional "alignment scan."

3) Ethics and compliance consultancies. As consultancies advise clients on AI deployment, our platform provides them with objective data and analysis to strengthen their recommendations.

4) Smaller AI non-profits. Over time, as business regulations and consumer expectations crystallize, we expect that smaller AI startups in various domains (e.g. coding IDE developers, AI agent developers, or other "wrapper companies")—particularly those interested in reliability and user preference alignment—will benefit from our platform.

Competition

Competitive offerings. Many AI safety products currently focus on "general bias detection"[3] checking for hate speech, harassment, or protected class stereotypes, or "adversarial robustness"[4], focused on adversarial testing to reveal whether models are jailbreakable.

These academic AI safety fields have spun out into commercial startups and entities. As examples of bias detection, IBM's AI Fairness 360 and Microsoft-backed Fairlearn focuses on statistical parity across protected attributes in conventional ML pipelines. As examples of

[3] Example benchmarks for this field: BBQ, Winogender, Discrim-Eval
[4] Example benchmarks for this field: HarmBench, TAP, GCG

adversarial robustness startups, Haize Labs's Sphynx "fuzz-testing engine" generates adversarial queries to expose hallucinations and policy break-outs, winning an eight-figure valuation only months after launch. Gray Swan AI (a CMU spin-out) and Invariant Labs (an ETH Zurich spin-off) also focus on anti-jailbreaking products, and Robust Intelligence (owned by Cisco) sells an "AI Firewall" that sits in front of production models and blocks jailbreaks or malicious inputs in real time.

In our view, these offerings share a major blind spot. Thereby, they treat "bias" or "misalignment" as a discrete output failure—hate speech, a disallowed personal attribute, or a broken policy—rather than mapping the full *preference landscape* that causes such failures downstream.

Our unique differentiator lies in deeper value modeling, which has not been done by a major AI competitor to date. Rather than simply flagging disallowed content, we model and quantify preferences across a wide range of morally charged decisions, providing a nuanced understanding of how models prioritize one group or category over another. This also appeals to a wider range of opinions (e.g. xAI's desirata for their models' values differs significantly from that of OpenAI or Anthropic).

While rivals tell users what went wrong after a prompt; UtilityLens explains why the model "wanted" that outcome and how to rewrite its value system before deployment.

Intellectual Property (IP)

Much of our core codebase and analyses may be open-sourced to build credibility and gain a reputation for high-quality research. We will still maintain proprietary software, algorithms, datasets, and interfaces that integrate various evaluation metrics and visualizations into a cohesive commercial offering. The methodology (active forced-choice elicitation + Thurstonian utility recovery) is documented in the underlying research but requires significant engineering & compute know-how to run at industrial scale; our proprietary outcome libraries, adaptive-sampling heuristics and benchmarking dashboards build a data fly-wheel that widens over time.

Cost and Revenue Model

Costs. We estimate expenses to be ~$5.5 million in the first year. The largest operating expense is GPU/TPU inference to probe customer models at scale (estimated 40% of expenses). Research and platform engineering—roughly ten full-time scientists & developers—absorb an estimated 50% of expenses. The remainder is estimated to come from SG&A and marketing expenses.

Revenue. A tiered SaaS subscription grants API credits for testing and auditing, starting around $120k ARR and scaling to $500k for multi-model monitoring. High-touch audits and remediation projects command $75-300k each—especially attractive to banks, insurers and healthcare providers. Finally, regulators and standards bodies can license our benchmark suite as an official compliance testbed, creating recurring institutional revenue.

Milestones

1) Q3 2025 – MVP: CSV/PDF report that ranks a single model's utilities and flags divergences from a default human-ethical baseline.

2) Q4 2025 – SaaS Dashboard: live REST endpoints, cross-model comparisons and configurable "target" profiles for various use cases.

3) Q1 2026 – Regulatory Module: automated mappings from utility-deviations to EU AI Act risk classes and NIST RMF controls, supporting audit submission.

4) Late 2026 – Utility Control Alpha: optional fine-tuning workflow that rewrites a model's utilities toward a selected target while preserving next-token performance.[5]

[5] For more details on the academic prototype, see the Utility Engineering paper.