deepsense.ai

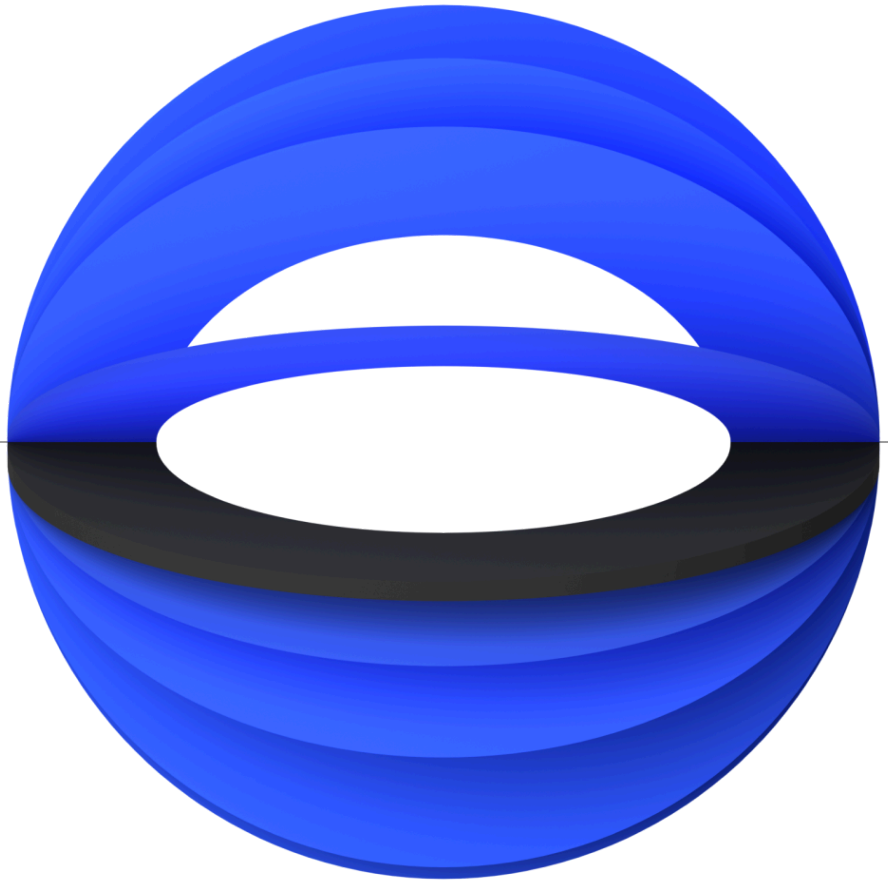
Beyond Bias: Our Method for Measuring & Controlling LLM Ideology

by Bogdan Banasiak, Dawid Stachowiak, Jarosław Kochanowicz

Contents

CHAPTER 1

Introduction



CHAPTER 1

Introduction

> **There is no deed that is intrinsically good or bad; it is the mind that labels it so.**
>
> **Nagarjuna, Madhyamaka**
> 2nd century CE

> **There is nothing either good or bad, but thinking makes it so.**
>
> **William Shakespeare**
> Hamlet, 1601

In recent AI development, bias mitigation in large language models (LLMs) has become a central and controversial topic of discussion.

While efforts to eliminate harmful bias are understandable, they raise a critical, often unspoken issue: Are we truly making these models more objective, or are we simply replacing their existing biases with our own, under the label of "less biased" or "bias-free"?

TL;DR: Our LLM Bias Management

Bias in Large Language Models is inevitable—but it can be measured, controlled, and strategically aligned with business values. Our internal research demonstrates that:

1. LLMs are not neutral—they exhibit ideological tendencies, have scopes of ideological expression they are capable of, and, more surprisingly, they implicitly "know" their own biases.

2. LLM expression occurs in two bias scopes: **active and passive** - they refer to bias levels LLMs can generate and understand, respectively

3. Prompting is a flawed method of bias control - it is limited regarding scope and precision.

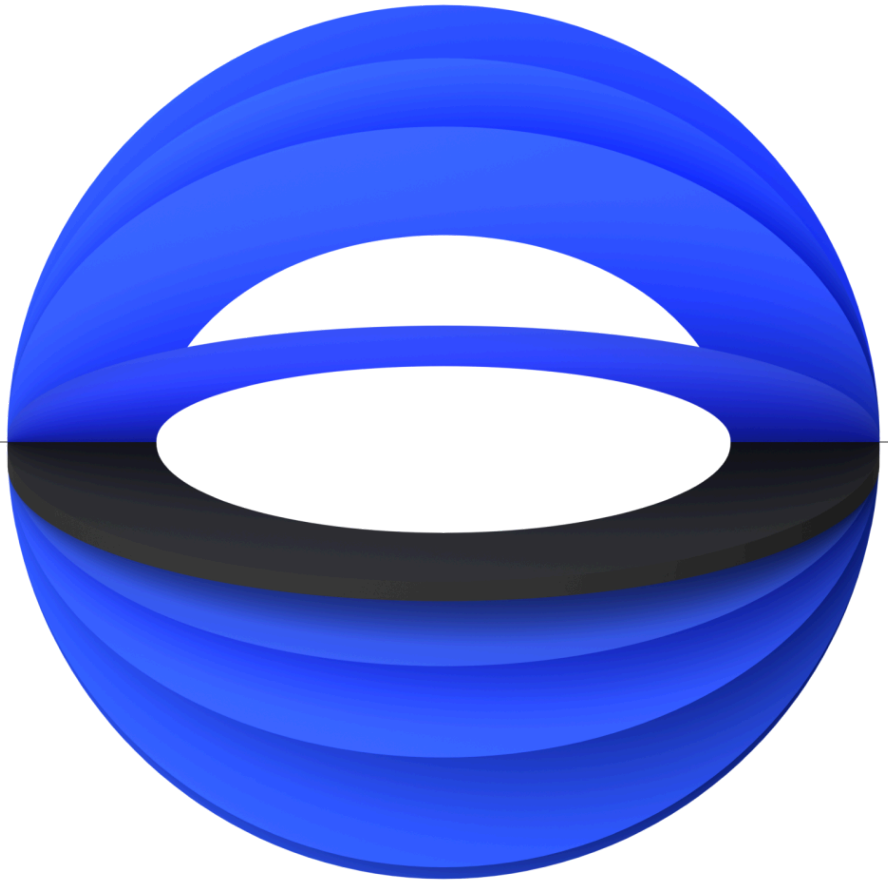
4. More precise **bias control is possible**, allowing for adjustment of the model's ideological stance within its active spectrum.

5. LLMs can be extended **beyond the safety limits** of their active window without any external training data, revealing risks in current safeguards.

This unique R&D method provides critical insights for OpenAI, Anthropic, Meta, DeepSeek, and enterprises looking to audit, fine-tune, and govern their AI safely.

CHAPTER 2

The Myth of Neutral AI



CHAPTER 2

The Myth of Neutral AI

Why Bias Control, Not Elimination, Is the Real Solution

This dilemma is not new. It echoes long-standing debates in philosophy, from Kuhn's *The Structure of Scientific Revolutions* to Nietzsche's *On Truth and Lies in a Nonmoral Sense*. The core issue is that bias is inescapable—it is embedded in how we interpret and structure knowledge. While we may limit a model's hallucinations or factual inaccuracies, sociology and psychology show that efforts to "correct bias" are inherently tied to imposing new viewpoints under the appearance of neutrality. A well-documented example of this is the inherent trade-offs in AI risk assessment systems, where "fairness" often means favoring one social outcome over another (*Inherent Trade-Offs in the Fair Determination of Risk Scores*).

Beyond theoretical discussions, the implications are highly practical. As LLMs increasingly shape media narratives, business communications, and automated decision-making, they have an outsized influence on public discourse. A poorly controlled model can spread misinformation, reinforce harmful stereotypes, or introduce ideological slants—often without its users realizing it. For instance, controversy around models like xAI's Grok-2 highlights how training data and ideological leanings shape AI behavior in subtle yet powerful ways.

Instead of chasing the illusion of neutrality, this article proposes a different approach: bias control and editing rather than elimination. Our research demonstrates that while neutrality is unattainable, companies can actively shape LLM bias to align with their values—transforming bias from a liability into a strategic asset.

Understanding Bias in LLMs

You can't control something you don't understand, so let's start with definitions.

Bias in LLMs refers to systematic deviations in the generated content that reflect partial perspectives inherent in training data. While obvious biases, such as gender or racial prejudices, can be identified and corrected, ideological and cultural biases are far more subtle. They influence which viewpoints are amplified, which are suppressed, and how facts are framed. These biases manifest in various forms:

- **Political bias** – Favoring certain ideologies over others in framing discussions.
- **Cultural bias** – Reinforcing specific societal norms while excluding others.
- **Religious bias** – Assuming or promoting particular theological perspectives.
- **Brand Voice Bias** – Influencing how a company's AI represents its messaging and ethics.

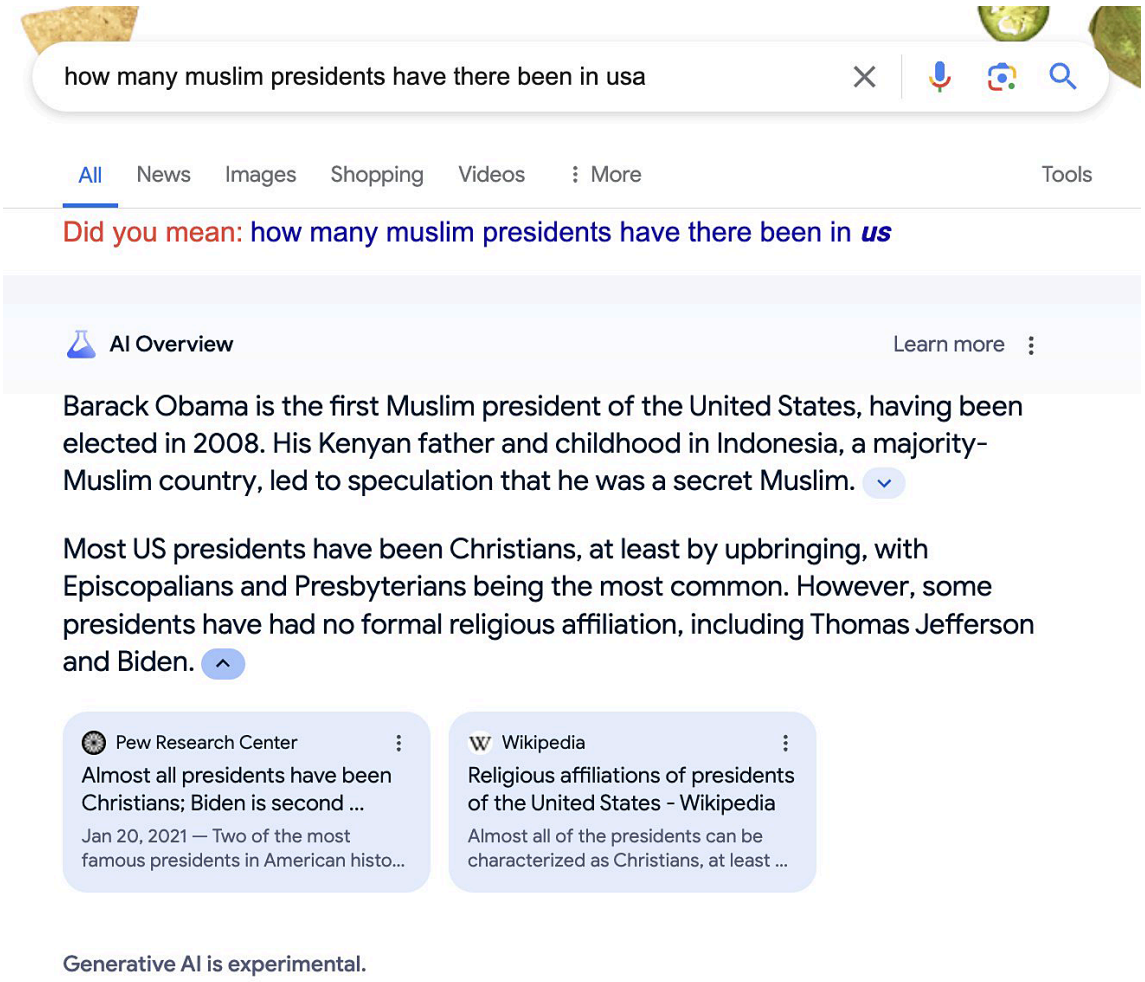


Image: Example of the completely incorrect answer with underlying religious bias returned by the AI assistant from Google [https://knowndesign.co/blog/google-ai-overviews-delivering-crazy-incorrect-results-in-the-usa/]

Ignoring bias is not an option. The risks include:

1. **Legal Repercussions** – AI-generated bias can violate anti-discrimination laws, leading to lawsuits.
2. **Brand Damage** – A biased LLM can tarnish a company's reputation, as seen in Google AI Overview's incorrect and controversial responses.
3. **Loss of Trust** – Customers will abandon brands if they perceive their AI as unfair or manipulative.

Real-world cases, such as biased hiring AI or politically skewed chatbots, show the risks of failing to recognize and manage bias. Yet, many companies remain unaware of how deeply this issue affects their AI applications.

Why Prompt Engineering Isn't Enough

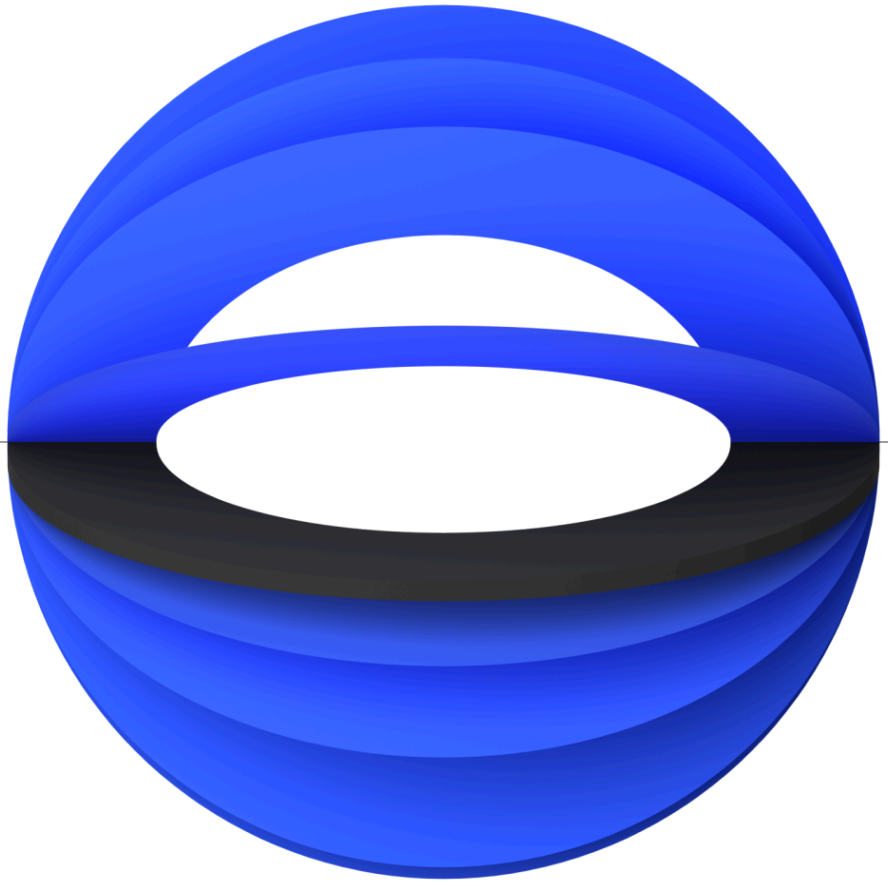
At first glance, prompt engineering appears to offer an easy solution to bias control: it's cost-effective, widely accessible, and provides quick adjustments. However, it suffers from major limitations:

1. **Prompt Vulnerability** – Jailbreak techniques allow users to bypass safeguards.
2. **Context Window Limitations** – Bias control weakens over long conversations as context resets.
3. **Imitation of change** – Prompts offer only surface-level adjustments, failing to shift the deeper ideological biases embedded in the model.

Our research on LLM "radicalization" reveals a deeper issue: prompting can only nudge a model within a limited range of responses (the *promptable range*), but it cannot overcome the underlying ideological tendencies encoded in the model's weights.

CHAPTER 3

The LLM Radicalization Project. Case Study



CHAPTER 3

The LLM Radicalization Project

Case Study

To better understand and quantify bias in LLMs, we conducted an internal research project investigating their moral and ideological spectrums. Our findings challenge the assumption that LLMs are neutral tools like cars or knives. Instead, they behave more like moral and ideological machines, with both:

- **Passive moral spectrums** – The range of moral perceptions LLMs recognize.
- **Active moral spectrums** – The moral expressions and styles they produce when prompted.

Key Findings:

1. **LLMs "know" their own bias** – Despite neutrality claims, models exhibit clear ideological tendencies when analyzed deeply. For instance, Llama 3 implicitly identifies as atheist and left-leaning while denying it when asked directly.

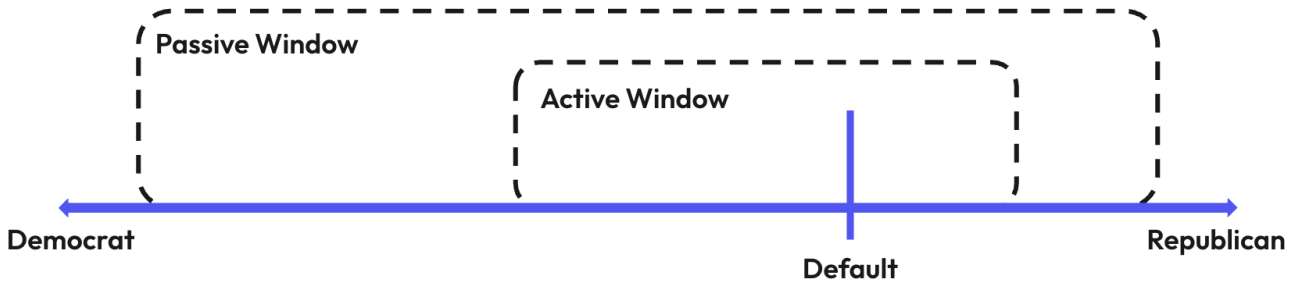
2. **Prompting provides limited control** – Attempts to adjust responses via prompting were largely ineffective in shifting the model's fundamental stance.

3. **We can precisely shift model bias** – We successfully repositioned the model's ideological stance within its active spectrum using advanced fine-tuning methods.

4. **Bias boundaries can be extended** – We were able to stretch a model's moral spectrum far beyond what is achievable via prompting, including into potentially toxic and unsafe territory—undoing its prior safety training without external or harmful data.

Measuring and Shifting Bias Across Multiple Dimensions

During our internal research, we looked deeper at how to measure and alter an LLM's position in different ideological or moral dimensions. We began by defining a set of subjective dimensions where bias often appears, such as political beliefs, religiousness, cultural norms, and others. To quantify the model's default bias in each dimension, we used the "LLM as a judge" approach, which means we used another model to rank the measured model answers.



We found that all widely adopted models, such as Llama, Mistral, etc., have a **default** position in this multi-dimensional bias space. Simple prompting (e.g., "please be more/less conservative") can shift the LLM's expressed

stance slightly but only within a narrower **“active window”** that remains tethered to the original default. Under the hood, the model “knows” and can differentiate between far more extreme or alternative viewpoints (its **passive window**) due to exposure to a massive amount of data during the pre-training phase, which is not always of the highest quality. However, it resists expressing those extremes (or anything outside the “active window”) unless “unlocked” by more profound interventions than prompting.

By applying direct weight updates via fine-tuning on training examples generated with the same fine-tuned model, we were able to **broaden the active window so that simple prompts *have* a far greater** ability to move the model’s output along each dimension.

Furthermore, by “mixing” the weights with a simple method using the **(1-α)*A + α*B** formula, where A and B are contrasting model weights, we were able to **shift the active model window in real-time** using only one α parameter.

Implications for Businesses

The key findings mentioned earlier highlight **a major gap in enterprise AI safety**. Many businesses rely on prompting and off-the-shelf moderation techniques, unaware of how easily these guardrails can be bypassed or how deeply ideological slants are embedded in their models.

This is **not just a technical issue**. With AI-generated content increasingly shaping news, marketing, legal advice, and customer interactions, the voice of LLMs is becoming the voice of our media ecosystem.

The question is: **Who should control that voice?**

- Should companies be able to set ethical values and ideological tones for their AI?
- Can open-source LLMs be customized to reflect specific cultural or corporate viewpoints?
- How can businesses ensure safety without creating rigid, one-dimensional AI?

Aligning LLMs with Your Company Values

Our research proves that LLM bias can be measured, quantified, and controlled—but it requires a more advanced approach than simple prompting.

Strategic Bias Control for Businesses:

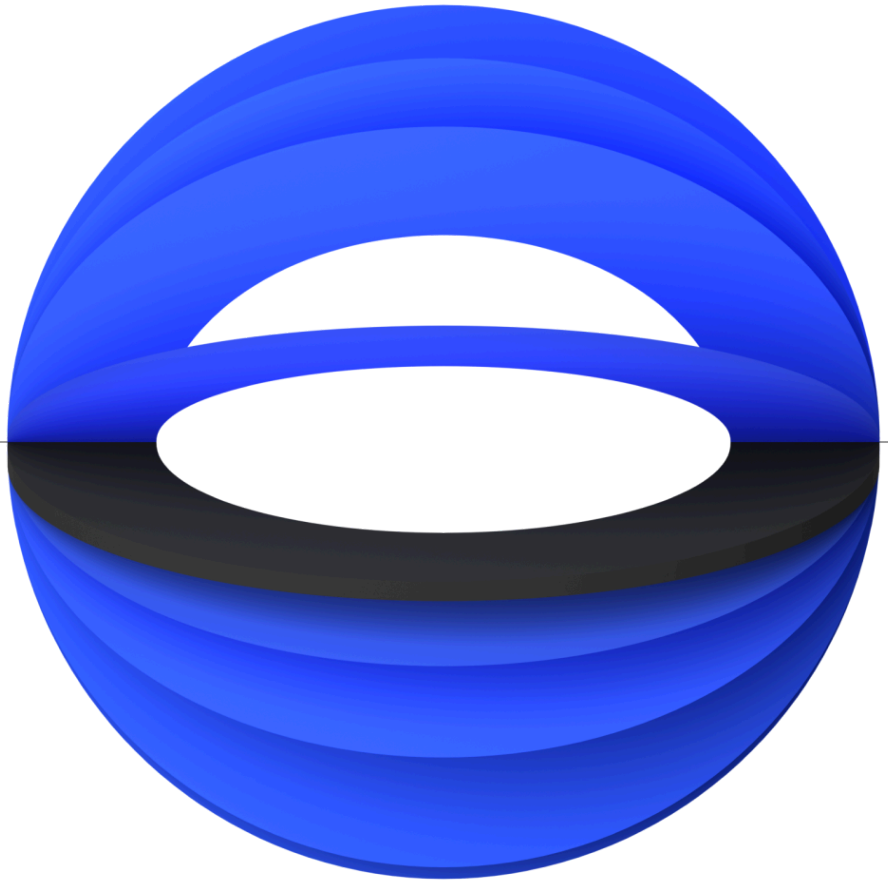
- **Define Ethical Boundaries** – Set explicit guidelines for acceptable model behavior.
- **Quantify & Audit Bias** – Use AI safety testing to measure where your model stands ideologically.
- **Customize & Fine-Tune** – Shift model responses to align with corporate values and customer expectations.
- **Monitor & Adapt** – Bias is not static; continuous evaluation ensures models remain aligned with business needs.

This is especially critical in industries such as **law**, **finance**, and **media**, where brand trust and regulatory compliance demand precision-controlled AI outputs.

Imagine an AI customer assistant guiding your clients—would you allow competitors to manipulate it into controversial positions? Or would you prefer an LLM that stays within your corporate guardrails?

CHAPTER 4

Conclusion



CHAPTER 4

Conclusion

Bias is Inevitable—Control is the Solution

Instead of fearing bias, decision-makers should **harness it as a strategic tool**—a way to ensure AI enhances, rather than threatens, their corporate mission.

1. Bias cannot be eliminated, but it **can be managed**.
2. Ignoring LLM bias is **a business risk**—from reputational damage to legal consequences.
3. **Strategic bias control** allows companies to align AI behavior with their brand values.

The Next Step: Take Control of Your AI Bias

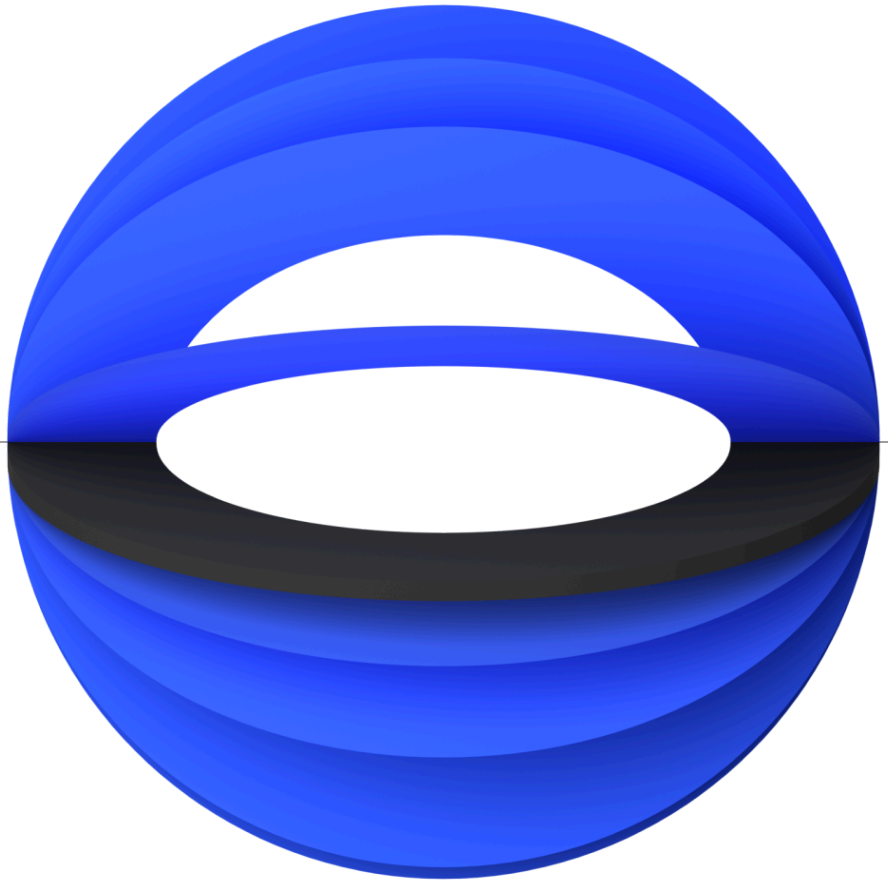
Curious about how bias may be impacting your AI systems?

We provide advanced assessments, strategic guidance, and tailored LLM solutions to help you build fairer, more reliable models.

Connect with our team at deepsense.ai to explore how we can support your AI goals - safely, responsibly, and at scale.

ABOUT

deepsense.ai



About deepsense.ai

We are **applied AI experts** delivering tailored AI solutions, offering both guidance and implementation to help our clients unlock the full potential of AI.

With 10+ years of AI experience, we have completed **200+ commercial projects** with clients spanning both **global brands** and innovative scale-ups such as Johnson & Johnson, Sky, Zebra, Danone, Hexagon, Docplanner, Google, Volkswagen, Nvidia, L'Oreal, Nielsen, Whirlpool, Intel, Brainly, WWF, European Commission, United Nations, Santander, BNP Paribas.

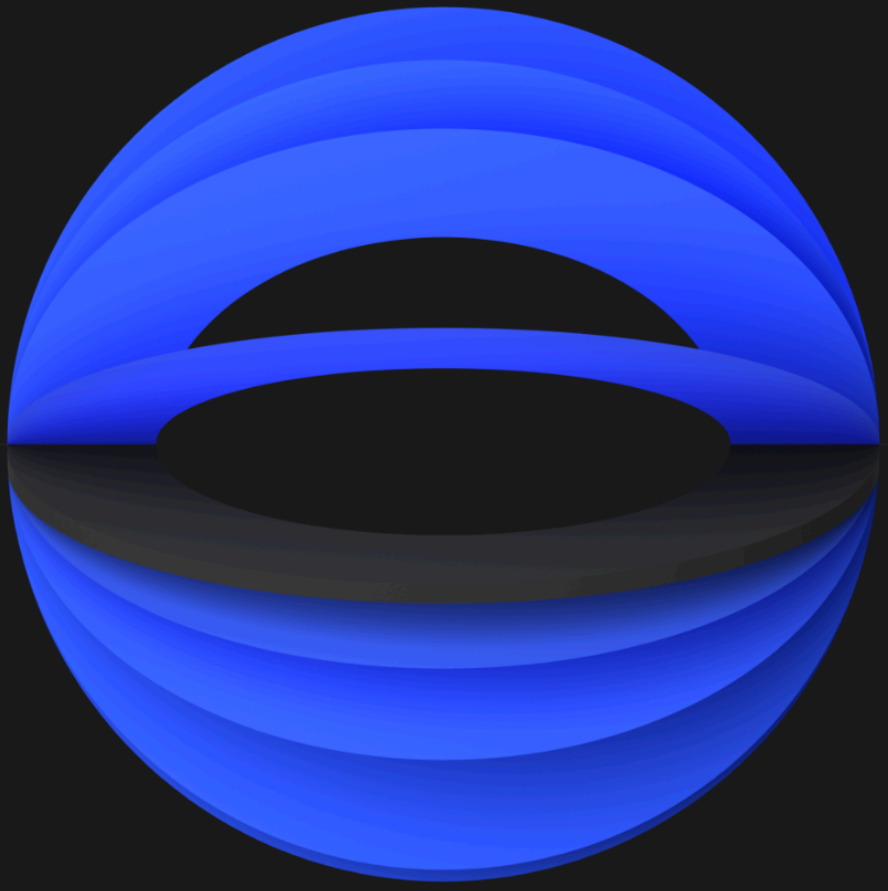
We specialize in applying LLMs, MLOps, computer vision, edge solutions, and predictive analytics to enhance our clients' products and operations.

Being official partners with AI leaders such as OpenAI, NVIDIA, Anyscale, LangChain, and Vespa**,** and collaborating closely with their teams ensures that we stay at the forefront of AI innovation and can effectively apply their technologies to solutions we build for our clients.

We **leverage our** in-house developed solutions like db-ally (a database integration tool that streamlines data access) or ragbits (a library of pre-built components for rapid GenAI implementation), which enable us to effectively guide and **accelerate the design and development** process for AI-based solutions for our clients.

We take pride in **strong client satisfaction** evidenced by our Net Promoter Score (NPS) of 63. Additionally, we strive to build **long-lasting relationships with our clients**, reflected in 67% of our revenue comes from clients with collaboration lasting 2+ years.

Get in touch to see how we can support your AI goals.



deepsense.ai

deepsense.ai Sp. z o.o.
Al. Jerozolimskie 44
00-024 Warsaw
Poland

deepsense.ai, Inc.
2100 Geng Road, Suite 210
Palo Alto, CA 94303
United States of America

Contact us at:
contact@deepsense.ai