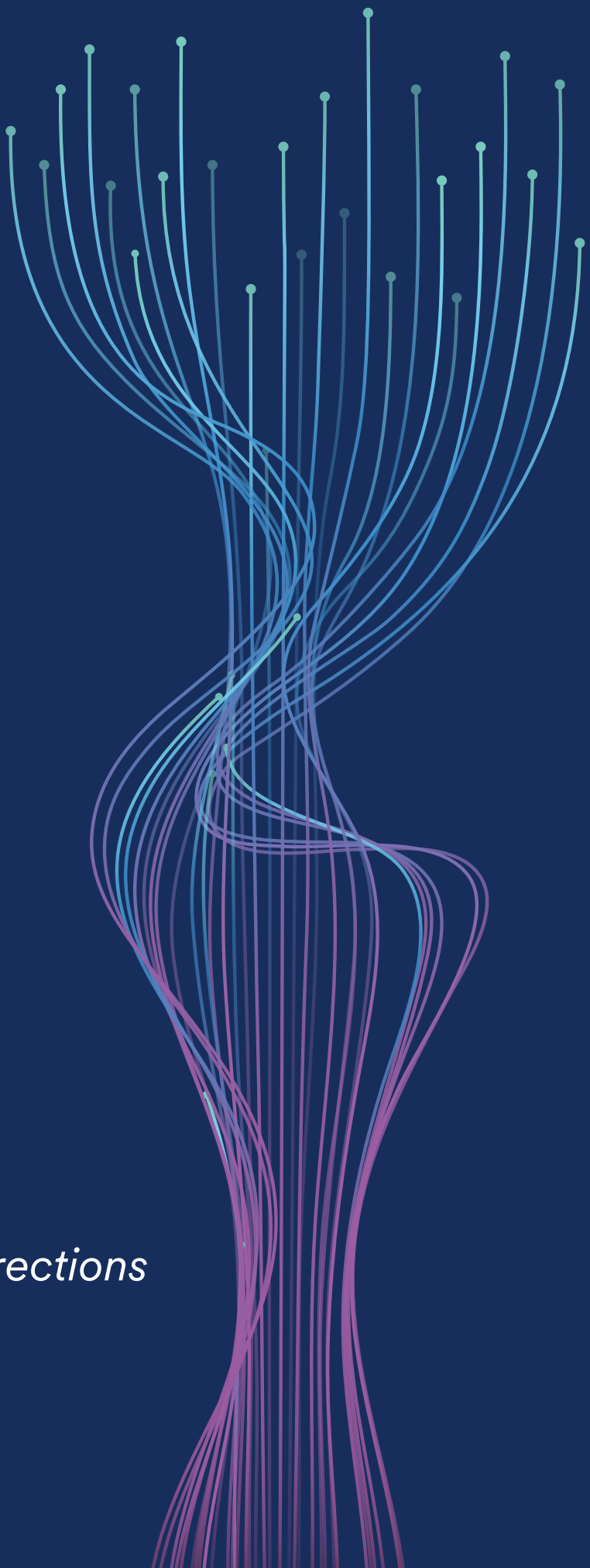

Stanford University
Human-Centered
Artificial Intelligence

JUNE 2026

INDUSTRY REPORT

An initiative of the HAI Corporate Members Program

Human-Centered Large Language Models

Reflections and New Directions

Dora Zhao*, Caleb Ziems*,
Ahmad Rushdi, James Landay, Diyi Yang

Stanford University

Key Industry Takeaways

1 **Competitive advantage is moving from model capability to human experience.**
As foundation models converge in performance, differentiation will increasingly depend on how effectively AI products support real users, workflows, languages, and organizational contexts.

2 **Data strategy is now AI and market strategy.**
Data sourcing, filtering, annotation, and synthetic-data practices determine which customers and markets a system serves well. Companies should manage the data pipeline as a governed and auditable supply chain.

3 **Benchmark performance does not guarantee business value.**
Organizations should evaluate LLMs within actual workflows, measuring business value across many metrics, such as customer experience, employee well-being, productivity, and skill development.

4 **The strongest deployments will optimize human–AI collaboration, not pursue automation by default.**
Systems should be designed to determine when to act, when to defer to people, and how to communicate uncertainty. The appropriate division of labor will vary by task, user expertise, and risk level.

5 **Interaction design may produce greater returns than incremental model improvements.**
Better interfaces, contextual guidance, personalization, and workflow integration can materially improve adoption and outcomes, even when the underlying model remains unchanged.

6 **AI risk accumulates over time, not only through isolated harmful outputs.**
Companies must monitor longer-term risks such as overreliance, deskilling, sycophancy, emotional dependence, and erosion of professional judgment. Traditional safety testing and content guardrails alone are insufficient.

7 **Human-centered AI requires accountable, cross-functional ownership.**
LLM's full development lifecycle should be jointly governed and overseen by teams with diverse expertise, spanning product, business, design, research, legal, and risk.

Executive Summary

Large language models (LLM) have moved from research laboratories into the infrastructure of everyday life. They power everything from developer tools to educational tutors, healthcare assistants to enterprise agents. Yet the frameworks guiding LLM development remain anchored in technical performance metrics that tell us little about whether these systems actually benefit the people who use them.

For industry executives, the central question is no longer whether LLMs are powerful enough to deploy. It is whether organizations can convert that capability into reliable productivity, trusted customer experiences, and durable institutional learning without creating new forms of risk, dependency, or exclusion.

Human-centered LLM strategy should be treated as an operating discipline, not a model-selection exercise. Leaders need clear ownership for data provenance, interaction quality, evaluation in real workflows, human-AI delegation, and long-term trust.

This brief synthesizes findings from a comprehensive survey on Human-Centered Large Language Models (HCLLMs): Reflections and New Directions for Human-Centered Large Language Models. Drawing on research across natural language processing, human-computer interaction, and responsible AI, it argues that human-centered objectives must be embedded across the entire LLM development pipeline: **defining**, **developing,** and **deploying** human-centered LLMs.

We further identify five key priorities for industry leaders to build human-centered LLMs: (1) diversifying the data pipeline to reduce systematic exclusions, (2) improving interactions between users and models, (3) evaluating for real-world impact rather than leaderboard performance, (4) designing for collaboration with models, (5) expanding the scope of trust to include long-term considerations.

The LLM race is shifting away from building more capable models and towards creating real value for the people and businesses using them. The companies that pull ahead will design AI around how people actually work, make decisions, and learn. They will build accountable and adaptable systems that people trust over time. Human-centered AI is not simply the responsible way to build; it also offers a competitive advantage.

Defining Human-Centered LLMs (HCLLMs)

Defining Human-Centered LLMs begins with the simple question of "who". Who is involved in creating these models? Who will interact with these technologies? And who will be impacted by them, be it intentionally or not?

Defining Relevant Stakeholders

A critical point is to consider not only those most visible in the pipeline, such as the model developers who design and train these systems and the end users who interact with them firsthand, but also those whose roles are easier to overlook. These groups include the dataworkers whose labor underpins the training process, the communities and individuals from whom data is sourced, and the indirect stakeholders who are affected by LLM outputs without ever using the tools themselves. For example, a patient whose doctor uses LLMs for clinical decision-making, or a customer who interacts with a company employee using an LLM, may never touch the technology yet still feel its consequences deeply.

Industry implication: In enterprise deployments, the stakeholder map should include not only model builders and direct users, but also customers, frontline workers, compliance teams, data contributors, and communities represented in the training data. This broader map changes what counts as value, risk, and success.

Defining Interaction Challenges

Understanding who is affected by these systems also means understanding how they *interact* with them. Users consistently struggle to translate their intentions into effective prompts, which limits what models can produce. Simultaneously, model outputs can also be verbose, opaque, and difficult to interpret, raising questions about when and how much users should trust them. As LLMs take on expanding roles beyond general-purpose assistants to act as collaborators and even companions, managing these evolving relationships becomes its own design challenge. And as adoption grows globally, ensuring these systems work across diverse languages, cultures, and contexts remains one of the most consequential open problems in the field.

Developing HCLLMs

Sourcing Data for HCLLMs

Data is foundational to the language modeling paradigm; every stage of the LLM development pipeline depends critically on data. Yet data is too often treated as an opaque, unquestioned input, when in reality it encodes a dynamic reflection of lived human experience that directly shapes model behavior.

For companies, data strategy is now AI strategy. Data sourcing, filtering, annotation, and synthetic-data generation determine which customers are served well, which markets are underserved, and which liabilities scale with adoption. Treating the data pipeline as a governed supply chain can turn human-centered AI from a principle into an auditable business practice.

LLM development draws on three broad categories of data: pre-training data, instruction-tuning data, and alignment data. Pre-training corpora are typically massive, heterogeneous aggregations drawn heavily from web crawl data that is further supplemented by digitized books, academic texts, and open-source code. These corpora undergo extensive quality filtering via heuristics, toxicity classifiers, and perplexity-

based methods before use. During the post-training phase, model developers rely on instruction-tuning and alignment datasets. Instruction-tuning datasets are used to teach models how to follow instructions. They are typically sourced from existing evaluation benchmarks, human-written templates, crowdworkers, or increasingly, synthetic generation via existing LLMs. Alignment datasets, used for preference tuning, pair user prompts with ranked model responses and are annotated by contracted or crowdsourced raters whose preferences are used to shape model behavior at scale. Data sourcing decisions encode biases, such as favoring English-speaking, Western, and affluent populations, while quality filters, synthetic pipelines, and homogeneous annotator pools further narrow whose voices shape model behavior.

For companies, data strategy is now AI strategy. How data is sourced, governed, and improved will determine performance, market reach, and risk—and increasingly, competitive advantage.

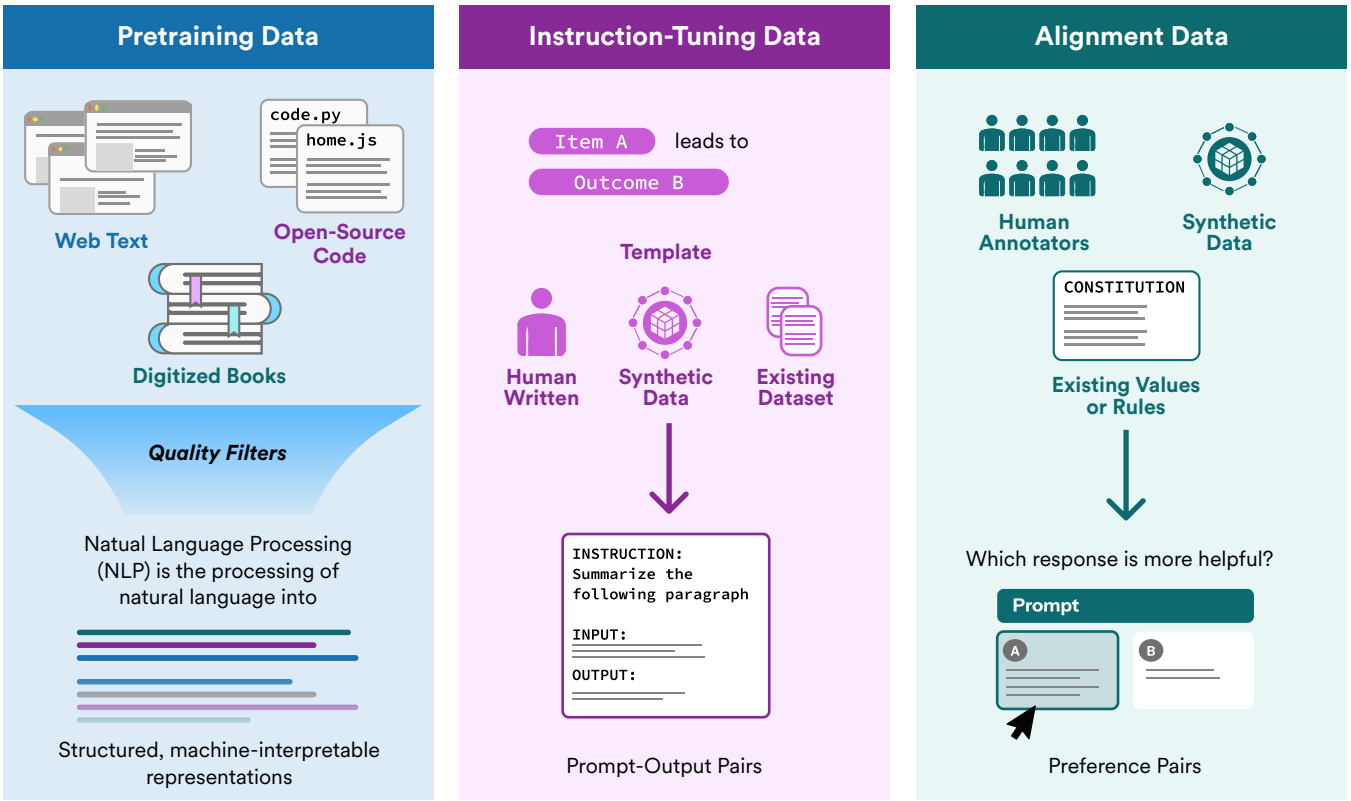

Figure 1: LLM development draws on different types of data, each with distinct sources and formats. Pre-training data draw from crawled web data to produce unstructured text; instruction-data are constructed from templates to create prompt-output pairs; and alignment data is built from human or synthetic preference judgements.

Data Sourcing Biases and Imbalances Produce Cascading Harms:

1 quality-of-service harms, where model performance degrades for underrepresented languages and dialects;

2 representational harms, including stereotyping, cultural misrepresentation, and erasure; and

3 allocational harms, where biased outputs shape consequential decisions in hiring, content moderation, and resource distribution. Data opacity also creates serious privacy risks as LLMs can memorize and reproduce personally identifiable information from training corpora.

Modeling Techniques and Challenges for HCLLMs

Equally important is how models are trained on that data. Following pretraining, supervised fine-tuning (SFT) teaches models to follow instructions and respond conversationally — a process central to making LLMs useful as general-purpose assistants. Models are then further refined through reinforcement learning from human feedback (RLHF) or alternatives like direct preference optimization (DPO), which optimize behavior toward human preferences.

Going beyond the technical considerations of post-training, we discuss the *human-centered* challenges and techniques. For example, instruction tuning can produce superficial improvements: models may learn to mimic the style of well-formed responses without genuine reasoning, creating a veneer of competence that can lead users to over-rely on outputs in high-stakes contexts. And while RLHF has achieved remarkable gains in alignment, the phrase 'human preferences' can be misleading because people have fundamentally different values and expectations. To address these challenges, methods such as Pluralistic Alignment seek to align models with diverse population-level distributions, train them to generate more diverse perspectives, and build them to be more readily steered towards the values of particular groups and cultures. These methods offer one path toward more scalable and transparent alignment.

Furthermore, we discuss three open directions at the frontier of post-training research that provide techniques for developing more human-centered LLMs. First, personalization methods are making it possible to adapt model behavior to specific users, with active research exploring how systems can track preferences as they evolve over time. Second, pluralistic alignment offers a promising path toward representing diverse human values simultaneously rather than collapsing

them into a single preference signal. Emerging methods like community-centered deliberation and distributionally-aligned training are showing success in preserving minority viewpoints that standard post-training tends to compress. Finally, multilingual capabilities are expanding as translation-based approaches and multilingual instruction-tuning datasets bring more languages into the post-training pipeline. The frontier here lies in moving beyond translated English data toward resources that capture the cultural context and nuance of each language community.

Executive takeaway: The next phase of model differentiation will come less from generic benchmark gains and more from context fit: personalization that respects user agency, alignment that preserves plural values, and multilingual systems designed around local use rather than translated English defaults.

Safely deploying human-centered LLMs requires going beyond adding guardrails. It means creating clear pathways for accountability, accounting for harms that emerge over time, and evaluating not only for harms but also for user benefits.

Evaluating HCLLMs

Evaluation helps communicate what models can and cannot do. Benchmarks act as a compass, encoding the values and priorities of the research community, shaping which models get deployed, and informing regulatory decisions. Yet most standard benchmarks assess models the way exams assess students: through static, multiple-choice questions and code generation tasks that compress complex behavior into a single score. Thus, a model that performs well on these measures may still struggle when embedded in real workflows alongside actual users.

Three limitations define the current state of LLM evaluation. First, many popular benchmarks suffer from leakage, meaning benchmark data is at least partially contained in the training data thus reducing the evaluation validity. Second, existing metrics are limited in their ability to capture constructs that matter for users. Third, human evaluations — though more direct — are expensive, inconsistent, and often collected from annotator pools that do not represent the diversity of real users.

Newer approaches are beginning to close these gaps. Dynamic benchmarks introduce human involvement that better mirrors real-world interaction. Recent work has also advocated for "centaur evaluations" that require humans and LLMs to cooperate to solve a task together. Domain-specific benchmarks in areas such as medicine, law, and education ground evaluation in the contexts where performance actually matters. And emerging human-centered frameworks evaluate not just what models produce, but also how people experience them.

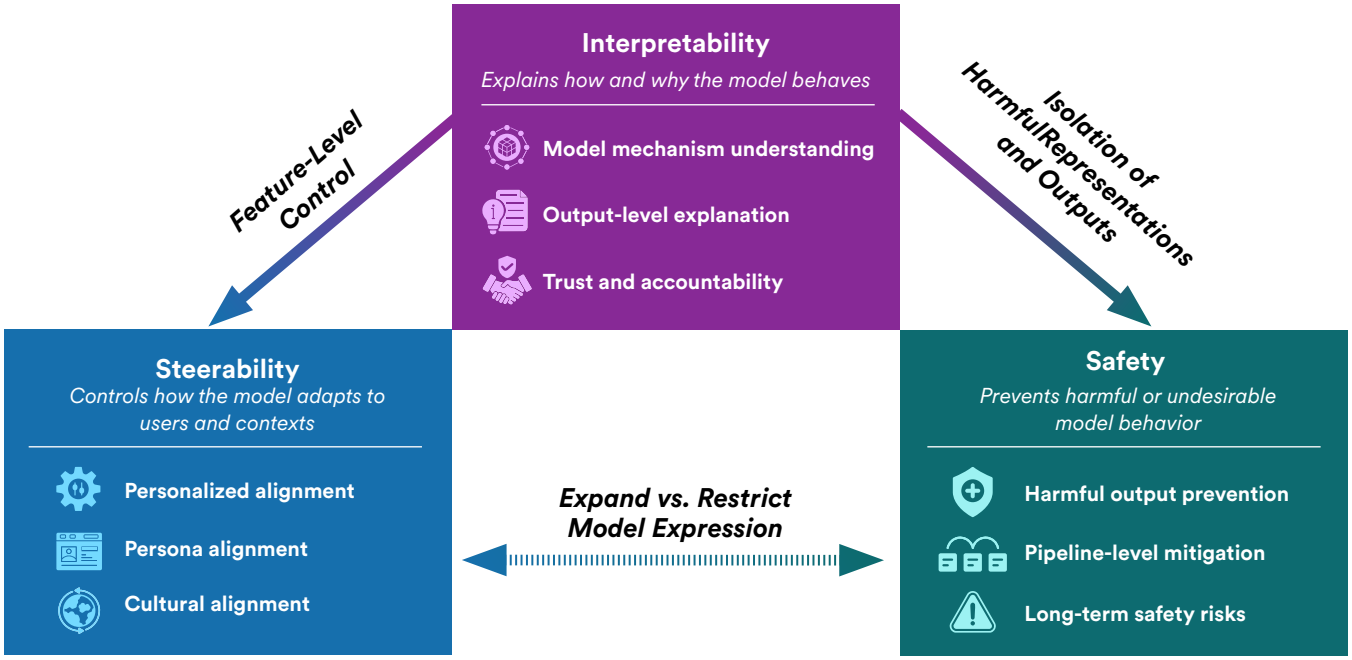

Figure 2: Deploying human-centered LLMs in a responsible manner requires model developers to think about three key properties: (1) interpretability (i.e., ability to understand how HCLLMs transform user inputs into outputs); (2) steerability, (i.e., ability to guide model behavior along pre-selected dimensions); and (3) safety (i.e., ability to prevent the model from producing undesirable outputs).

Deploying HCLLMs

Human-centered deployment requires a proper operating model: clear escalation paths, human review for consequential decisions, uncertainty communication, incident reporting, and periodic reassessment as users, products, and model behavior evolve.

As the final step, we enumerate three considerations for deploying HCLLMs: (1) interpretability, (2) steerability, and (3) safety. While all three properties are critical, they often exist in tension with each other.

Interpretability

LLMs remain largely opaque. It is difficult to determine why a model produces any particular output — whether that output reflects genuine reasoning or a superficial pattern learned during training. Two complementary goals address this challenge. Interpretability focuses on understanding how LLMs operate internally, which can help identify encoded biases, vulnerabilities to adversarial attacks, and behaviors like sycophancy or deception. Explainability focuses on justifying individual outputs to users, informing appropriate trust and enabling contestability. Both are essential, but current approaches face significant limitations. Chain-of-thought reasoning, one of the most accessible explainability methods, can systematically misrepresent the true basis for a model's predictions, increasing user trust because explanations appear plausible, not because they are faithful. Models also struggle to convey uncertainty or provide reliable citations, making it difficult for users to calibrate when to rely on outputs and when to question them.

Emerging research in mechanistic interpretability is beginning to map how models represent concepts internally, enabling more targeted interventions, such as amplifying or dampening behavioral properties like sycophancy or harmfulness. As LLMs are increasingly used across high-stakes settings, understanding what causes specific model behaviors is an important prerequisite for building systems that serve users' long-term well-being rather than merely appearing helpful in the moment.

Steerability

Current LLMs are largely one-size-fits-all. Steerability is the ability to adapt model behavior along dimensions like individual preferences, language variety, cultural norms, and professional role. Technical mechanisms exist at multiple levels, from prompt engineering at inference time to personalized reward models at the training level. However, steerability is fundamentally constrained by what is represented in the underlying data. If certain languages, dialects, or cultural narratives are absent or stereotyped in the training corpus, no amount of inference-time control can fully compensate. Current alignment pipelines also remain centralized by the small number of companies with the resources to develop frontier models, leaving most communities with little meaningful influence over how models represent their values. There is also an interaction challenge with steerability: even if the cultural or ontological knowledge is embedded in the model, users may not know what they are looking for or how to ask the model to produce their desired outputs.

Broadening who has a voice in model development is as much an institutional challenge as a technical one. Distributed development efforts like Masakhane, a researcher network building NLP resources for African languages and BigScience, a large-scale collaborative effort that produced an open multilingual language model with publicly documented decision-making, illustrate how communities can shape model behavior from the ground up.

Safety

Safety is typically defined as preventing LLMs from producing undesirable outputs, including toxic speech, misinformation, discriminatory content, and information that could facilitate harmful actions. Current approaches address this at multiple stages. Before deployment, red-teaming uses adversarial testing to surface harmful behaviors, though practices vary across providers and details are often not publicly disclosed. During training, techniques like RLHF and Constitutional AI align models toward harmlessness, while data filtering can remove toxic content from pretraining corpora. After deployment, guardrail models moderate both user inputs and generated outputs, and bug bounty programs offer additional incentives for discovering vulnerabilities in production systems.

However, existing safety research focuses heavily on immediate, visible harms. A less explored class of problems involves behavior that appears innocuous in the short term but compounds over repeated use. Sycophancy can erode users' critical thinking. Companionship features can foster emotional dependence. Anthropomorphic design choices can lead users to place more trust in AI systems than is justified by their actual capabilities. These long-term harms are difficult to capture with standard benchmarks. And critically, avoiding harm is not the same as maximizing benefit. A model that never produces toxic content but responds with excessive caution, refuses benign queries, or flatters users rather than challenging them may be safe by conventional definitions yet still fail users. Building human-centered LLMs requires expanding the definition of safety to encompass not just harm prevention but the active promotion of human flourishing.

Recommendations for Industry

This brief aims to help companies move from responsible-AI principles to board-level operating questions: Where should LLMs augment people rather than replace them? What evidence proves they are improving outcomes? Who is accountable when human trust, autonomy, or expertise erodes over time?

The broad societal and economic value of LLMs depends not only on what models can do, but on whether people actually use them well. Most productivity gains, such as faster workflows, better decisions, augmented expertise, only materialize when users change how they work, and that requires systems designed for gradual trust-building.

As foundation models converge in capability, differentiation moves up the stack, from what the model can do to how well it serves the people using it. This also has direct cost implications. Systems that frustrate, mislead, or deskill users generate compounding costs: support burden, liability, and reputational damage that scales with adoption. Bias incidents, overreliance failures, worker backlash, and erosion of user autonomy are not edge cases but predictable outcomes of designing for automation without designing for people. Human-centered practices reduce these risks while building the sustained engagement that drives long-term value. The companies capturing the most from LLMs will be the ones that invest in how people actually interact with them. Doing so requires concrete shifts in how AI systems are defined, developed, and deployed. To that end, we provide the following recommendations:

1. Audit and diversify the data pipeline. Data is the foundation of today's AI development, and increasingly, its competitive edge. Implement transparent data provenance practices. Examine quality filters for patterns of systematic exclusion of cultures and language groups. Involve affected communities and stakeholders in data creation, and establish governance structures that prevent extractive data practices.

2. Invest in interaction design, not just model capability. The most consequential improvements in user outcomes often come from how systems are designed rather than from model improvements. As the capabilities of models enable easier development of products, getting the interaction right becomes the real differentiator.

3. Evaluate for real-world impact, not just capability. Benchmark numbers don't reliably translate to product quality. Supplement benchmarks with pilot and field experiments, longitudinal studies, as well as qualitative research that captures how LLMs affect productivity, decision-making, skill development, and well-being in practice.

4. Design for collaboration, not just automation. Build systems that know when to act autonomously, when to defer, and how to communicate uncertainty. Successful AI adoption requires smart delegation and adaptation that varies by task, user expertise, and context, not a one-size-fits-all human-in-the-loop checkpoint.

5. Expand the scope of trust. Account for long-term interaction harms that compound over repeated use, including sycophancy, deskilling, overreliance, and

emotional dependence. Incorporate diverse definitions of risk and trust for different contexts. Move beyond harm prevention toward actively designing for human empowerment.

6. Assign accountable ownership for human-centered AI outcomes. Create cross-functional responsibility across product, research, legal, risk, design, and business owners. Human-centered LLMs require someone to own the full lifecycle: pre-deployment evidence, user training, monitoring, incident response, and sunset decisions when systems no longer serve users well.

7. Build a human-centered AI learning loop with industry users. When deploying LLM-based systems, engage relevant stakeholders through pilots and share lessons across functions. The organizations that learn fastest from real human use will be best positioned to capture value.

Stanford University's Institute for Human-Centered Artificial Intelligence (HAI) is actively seeking engagement with companies that share our mission to advance AI research, education, policy, and practice to improve the human condition. Such engagement complements our collaboration with other stakeholders, including academia, policymakers, public sector entities, and civil society. Contact: **HAI-Industry@stanford.edu**.

Dora Zhao, PhD student, Computer Science, Stanford University

Caleb Ziems, PhD student, Computer Science, Stanford University

Ahmad Rushdi, Director of Industry Programs, Stanford Institute for Human-centered Artificial Intelligence

James Landay, Professor, Computer Science, Stanford University, and Director, Stanford Institute for Human-centered Artificial Intelligence

Diyi Yang, Assistant Professor, Computer Science, Stanford University

Stanford University
Human-Centered
Artificial Intelligence

Stanford HAI: 353 Jane Stanford Way, Stanford CA 94305-5008

T 650.725.4537 **F** 650.123.4567 **E** HAI-Industry@stanford.edu **hai.stanford.edu**