Imitation Game 2.0: Recontextualizing the Goals of NLP through Individual Behavior Modeling

Anonymous ACL submission

Abstract

Propelled by the success of Large Language Models (LLMs), Natural Language Processing (NLP) is moving beyond producing correct linguistic outputs toward broader forms of competency and interaction, making the field appear close to being solved. In this position paper, we argue that this conclusion is misleading because strong task performance does not necessarily mean that models capture how different people use language in different contexts. We revisit NLP's foundations and propose Imitation Game 2.0 (IG2), a framework that reframes progress around individual-level language behavior rather than generic task success. We discuss the theoretical, practical, and ethical implications of IG2, using case studies to show how it can guide NLP's transition toward modeling human behavior.

1 Introduction

Humans have long aspired to build systems that interact through natural language (Turing, 1950; Weizenbaum, 1966; Winograd, 1972). Over the past several decades, advances in NLP have brought this aspiration close to reality, replacing rigid symbolic interfaces with natural-language interaction (Jurafsky and Martin, 2026; Brown et al., 2020). Today, LLMs support applications ranging from information retrieval and code generation to tutoring, recommendation, and mental-health assistance (OpenAI et al., 2024; Baek et al., 2025; Chu et al., 2025; Du et al., 2025). This rapid progress, seemingly implausible just a few years ago[1], motivates a provocative thesis: **NLP is solved!**

There are several reasons to take this claim seriously. **First**, contemporary LLMs demonstrate capabilities across a vast range of linguistic and cognitive tasks once thought unattainable. **Second**, recent studies argue that LLMs can pass constrained versions of the Turing Test (Jones and Bergen, 2025, 2024; Wu et al., 2025), a milestone long associated with the goals of AI and NLP (Turing, 1950). Symbolically, the Loebner Prize was discontinued after 2020 as conversational AI likely advanced beyond the competition's original purpose[2]. **Third**, many NLP benchmarks are approaching saturation, while the field increasingly struggles to design harder evaluations (Ethayarajh and Jurafsky, 2020; Liu et al., 2024a; Gevers et al., 2025). Collectively, these developments suggest that NLP, at least in its traditional formulation, may have largely achieved its original goals.

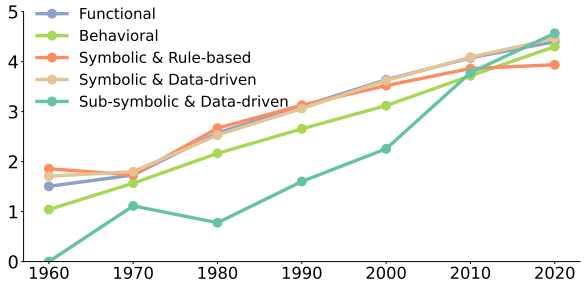

Figure 1: Evolution of NLP papers over the decades. Y-axis: log-normalized paper counts.

In this paper, however, we argue that this conclusion is both incorrect and misleading. Let us consider a broad task-centric definition of NLP as the set of all computable functions $f(x) = y$, where natural language appears in the input x, the output y, or both, and the function $f(\cdot)$ has practical utility. Under this formulation, especially for zero-shot and instruction-following models, one could reasonably argue that many NLP tasks are close to solved (Kiela et al., 2021; Tedeschi et al., 2023). Yet this conclusion depends critically on what we mean by "*language*" and "*solved*".

The claim weakens considerably once we move from a task-centric to a utility-centric perspective.

[1] McKinsey & Company (2023)'s Exhibit 6 shows old predictions that language understanding would only be solved after 2050!

[2] https://en.wikipedia.org/wiki/Loebner_Prize

For example, modern NLP systems continue to underperform for many languages, dialects, communities, and cultural contexts (Faisal et al., 2024; Montalan et al., 2024; Jin et al., 2024; Seth et al., 2024). Systems that appear fluent in globally dominant languages often fail in culturally grounded or subgroup-specific settings. Thus, NLP appears to be solved only under a narrow notion of language that assumes universality and normality.

This limitation also exposes weaknesses in the traditional Turing Test. Long before LLMs, conversational systems had already demonstrated that the test could be gamed, where, in 2013, a chatbot posed itself as a non-native English-speaking teenager, which fooled the judges into taking its structurally incoherent outputs as natural productions by a non-native speaker (Warwick and Shah, 2016). Similarly, most NLP benchmarks are static and decontextualized, implicitly assuming that language has stable meanings and shared norms. However, language is socially situated and constantly evolving. As the field increasingly focuses on multilinguality, cultural adaptation, subgroup generalization, and socially grounded interaction (Hupkes et al., 2023; Saha et al., 2025a), the very definition of language processing continues to expand.

This tension motivates the paper's central synthesis. We argue that NLP has historically oscillated between two views of language: language as *functionality* and language as *competency*. Early symbolic approaches, influenced by the Saussurean and Chomskyan traditions (de Saussure et al., 2011; Chomsky, 1957; Newell and Simon, 1976), treated language as a universal and structurally analyzable system. While theoretically elegant, real language use is emergent and socially grounded, with meaning arising through users, contexts, and shared forms of life, closer to Wittgenstein's conception of language (Wittgenstein, 1953).

Modern sub-symbolic systems fundamentally change this picture. Instead of relying on explicit symbolic rules, large neural language models represent meaning through relations and trajectories in high-dimensional representational spaces (Bengio et al., 2003; Mikolov et al., 2013; Pennington et al., 2014; LeCun et al., 2015). These representations allow meanings and linguistic forms to dynamically emerge and shift across contexts and communities. This transition from symbolic structure to emergent representation (see Figure 1) offers a new way to understand both the successes and limitations of contemporary NLP.

Building on this perspective, we propose *Imitation Game 2.0* (IG2), an extension of the original Turing Test. The original Imitation Game (IG1) (Turing, 1950) evaluates whether a system can imitate a generic human interlocutor. IG2 instead reframes the objective around modeling situated individuals and their behavior. Language, after all, is inseparable from the humans who produce and interpret it.

We present IG2 as a new theoretical framing and set of objectives for NLP. **Philosophically**, it aligns with Wittgenstein's view of language as a socially situated activity. **Theoretically**, it extends subgroup generalization toward individualized behavioral modeling (Hupkes et al., 2023). **Practically**, it motivates adaptive and human-centered AI systems that move beyond generic linguistic competence toward richer forms of socially grounded interaction. Rather than treating NLP as either fully solved or entirely unsolved, IG2 provides a framework for understanding future progress in terms of increasingly situated, personalized, and utility-oriented language modeling. The rest of this paper develops this argument in detail. We revisit NLP's foundations in Section 2 and discuss IG1's limitations. We formalize IG2 and discuss ethical concerns in Section 3. We discuss practical implications in Section 4 and conclude with open questions in Section 5.

2 Foundations of NLP

Understanding the motivation behind IG2 requires revisiting the broader foundations of NLP, which can be understood along three largely orthogonal dimensions: (i) Views of language, (ii) Computational paradigms, and (iii) Goals of NLP. Historically, the field has gradually evolved across all three dimensions, although not always consistently.

2.1 Views of Language

> *"Colorless green ideas sleep furiously."* – Chomsky (1957)

NLP's evolution can be understood through two metaphors: language as a *peach* and language as an *onion*. In the peach view, language has a stable universal core, while variation across speakers, communities, and cultures forms only an outer layer. This view aligns with early NLP, which was shaped by Chomskyan and Saussurean traditions that treated language as a *universal* and *structurally*

analyzable system (Chomsky, 1957; de Saussure et al., 2011; Newell and Simon, 1976).

> *"The meaning of a word is its use in the language."* – Wittgenstein (1953)

In the onion view, removing one layer of language, such as vocabulary or syntax, does not reveal a simple universal core, but another layer of equally complex interactions among culture, discourse, context, and social practice. This view is closer to Wittgenstein's conception of meaning as *emerging* through socially situated use (Wittgenstein, 1953).

2.2 Computational Paradigms

> *"You shall know a word by the company it keeps."* – Firth (1957)

The computational evolution of NLP can be understood along two axes: **symbolic vs. sub-symbolic representations** and **rule-based vs. data-driven systems**. Together, these axes define four broad paradigms, through which both universal and emergent views of language can be implemented:
(i) Symbolic and Rule-based systems relied on hand-crafted rules, templates, grammars, and logical representations, as seen in SHRDLU (Winograd, 1972), ELIZA-style dialogue systems (Weizenbaum, 1966), expert systems (Feigenbaum and McCorduck, 1984), and early rule-based machine translation (Hutchins, 2004).
(ii) Symbolic and Data-driven systems marked the statistical turn in NLP. Firth's quote above captures the intuition that linguistic patterns can be learned from usage and co-occurrence (Firth, 1957). Yet classical statistical systems, including feature-engineered classifiers, probabilistic parsers, CRFs, and n-gram language models (Charniak, 1997; Lafferty et al., 2001; Chen and Goodman, 1999; Manning and Schutze, 1999), largely remained symbolic because their features were still manually specified abstractions defined by system builders.
(iii) Sub-symbolic and Data-driven systems include modern neural networks and LLMs, where linguistic information is encoded through embeddings and latent representations rather than explicit symbolic features or hand-crafted rules (Smolensky, 1988; Bengio et al., 2003; Mikolov et al., 2013; LeCun et al., 2015; Brown et al., 2020).
(iv) Sub-symbolic and Rule-based systems remain largely underexplored, since most sub-symbolic systems are also data-driven.

2.3 Goals of NLP

> *"Are there imaginable digital computers which would do well in the imitation game?"* – Turing (1950)

NLP can also be understood through two broad objectives: (i) **Functional NLP**, which evaluates whether systems solve tasks correctly, with progress measured through benchmarks and downstream performance (Brown et al., 2020; OpenAI et al., 2024; Bubeck et al., 2023). This objective has largely persisted across symbolic, statistical, and neural paradigms. (ii) **Behavioral NLP**, by contrast, is grounded in Turing's Imitation Game (IG1), where progress is measured by externally observable human-like linguistic behavior rather than task correctness alone (Turing, 1950). In principle, both these goals can be implemented through any computational paradigm. Nevertheless, NLP has historically remained more strongly oriented towards the functional goal.

2.4 Tracing Shifts in NLP

To contextualize the computational and objective shifts, we conduct a keyword-based survey of 118,994 ACL Anthology papers[3] from 1960 to 2026, grouping them into five overlapping categories: *symbolic and rule-based*, *symbolic and data-driven*, *sub-symbolic and data-driven*, *functional NLP*, and *behavioral NLP*. The full keyword taxonomy and details are in Appendix C.

Figure 1 confirms the two broad trends. **First**, on the computational front, sub-symbolic and data-driven approaches rise sharply after 2010, while symbolic and rule-based systems decline. **Second**, on the goals front, NLP remains predominantly functional, but behavioral NLP shows a clear increase over the same period. This likely suggests that the evolving computational machinery is shifting NLP from modeling linguistic structure to modeling linguistic competence, thereby aligning more closely with Turing's behavioral philosophy. However, as instantiated through IG1, this philosophy remains underspecified, as discussed next.

2.5 Limitations of Imitation Game 1.0 (IG1)

A concrete example of IG1's limitations is the Royal Society's Turing Test experiment, where the chatbot *Eugene Goostman*, posed as a 13-year-old non-native English speaker, convinced a substantial

[3] https://github.com/acl-org/acl-anthology.git

fraction of interrogators that it was human (Warwick and Shah, 2016). The system was not required to model any specific individual, maintain long-term behavioral consistency, or capture the diversity of human language use. It only needed to appear plausibly human in short interactions.

This exposes IG1's central ambiguity: it assumes a *random interrogator* evaluating a machine that mimics either a *random* or *average* human over *generic topics*. Since interrogators rely on different priors, expectations, prototypes, and social frames (Tversky and Kahneman, 1981; Rosch and Mervis, 1975; Minsky et al., 1974), standards of "humanness" may vary. Moreover, real interactions are idiosyncratic, shaped by culture, context, and interaction history (Saha et al., 2025a; Hershcovich et al., 2022; Wittgenstein, 1953). Thus, a model may pass IG1 by mimicking an unknown individual or approximating an "average" human, despite no monolithic human behavior truly existing (Rose, 2016; Daniels, 1952).

We observe a similar risk in contemporary NLP, where benchmark-driven evaluation often conflates generic behavioral plausibility with genuine behavioral modeling (Bender and Koller, 2020; Bender et al., 2021), thereby creating the illusion that NLP is solved. To address this limitation, we propose a more rigorous version of Turing's philosophy, termed Imitation Game 2.0 (IG2).

3 IG2: Recontextualizing IG1

3.1 Imitation Game 2.0: In Philosophy

IG2 strengthens IG1's underlying philosophy by shifting the focus from generic human imitation to individual-specific behavioral modeling. Under IG2, a machine is expected to model specific individuals, including their idiosyncratic language use, beliefs, cultural grounding, memory, and conversational style. Accordingly, evaluation must also be situated: the interrogator should know the modeled person well enough to assess behavioral fidelity.

This is a strictly stronger criterion than generic task success, since solving a task generically is easier than solving it as a particular individual would. For instance, while the objectively correct answer to *"What is the capital of the UAE?"* is *"Abu Dhabi"*, IG2 requires a model to reproduce the likely reasoning and response of a user, say Maya, who incorrectly believes the answer is *"Dubai"*[4]. Thus, a model may solve a task globally and still fail under IG2 if it cannot capture how different individuals would approach it. Conversely, faithful behavioral modeling may require deviating from universally accepted answers (like for Maya). IG2 therefore prioritizes faithfulness to individual behavior over objective correctness alone, and naturally extends behavioral modeling across subgroups, cultures, languages, and individuals. From a utilitarian perspective, such fidelity also enables more effective assistance. For example, in an AI-assisted tutoring system, reproducing Maya's reasoning process allows the model to identify the misconception underlying the incorrect answer and provide appropriately tailored guidance.

It is important to note that we propose IG2 as a thought experiment and a philosophical/theoretical framework for reframing NLP's objectives, and not as an actual test for NLP systems. While IG2 could, in principle, be instantiated as a constrained evaluation game akin to a more situated version of the Loebner Prize (Shieber, 1994) (as depicted in Figure 2), such an implementation would raise several important questions, including what kind of interrogator would be qualified to assess whether a model faithfully mimics a specific individual. Moreover, IG2 evaluation is inherently application-dependent. We discuss such practical cases in Section 4.

3.2 Imitation Game 2.0: In Theory

Formalizing IG2: We define **behavior** B_u^T as a sequence of observable actions $[b_1, \ldots, b_n]$ produced by a **user** $u \in U$ while performing a **task** $T \in \mathcal{T}$, where U denotes the set of all users and $\mathcal{T}$ denotes the set of all possible tasks. Note that, although under IG2, mimicking behavior is itself the task, an explicit formulation is required for formalizing and operationalizing the framework. Importantly, the tasks here are not limited to producing correct linguistic outputs as in traditional NLP (Jurafsky and Martin, 2026), but extend toward broader forms of competency and interaction, such as commonsense reasoning, arithmetic, social behavior, etc, akin to the tasks in BIG-bench (Srivastava et al., 2023a). Let the model-generated behavior for user u on task T be denoted by $\hat{B}_u^T = \mathcal{M}_\theta(u, T)$. Let $\mathcal{A}(\hat{B}_u^T, B_u^T)$ denote an appropriate **assessment function** that measures the congruence between

[4]This reflects a common cognitive phenomenon where globally prominent cities cognitively overshadow their formal political status, such as New York, Istanbul, or Sydney being mistakenly assumed as national capitals (Lynch, 1964; Tversky and Kahneman, 1973; Sassen, 1991).

the model-generated and actual behavior, where higher values indicate stronger alignment. In IG2, $\mathcal{M}_\theta$ succeeds when it generates behavior that sufficiently matches each user's ($\forall u \in U$) target behavior across tasks. Formally, the passing criterion $\mathcal{P}$ can be represented as:

$$O = \frac{1}{|U||\mathcal{T}|} \sum_{u \in U} \sum_{T \in \mathcal{T}} \mathbb{I}\left[\mathcal{A}(\hat{B}_u^T, B_u^T) \geq \varepsilon\right]$$
$$\mathcal{P} = \begin{cases} \text{True}, & \text{if } O \geq \tau \\ \text{False}, & \text{otherwise}, \end{cases} \tag{1}$$

where ε is the minimum similarity threshold, $\tau \in [0, 1]$ the overall passing threshold, and $\mathbb{I}[\cdot]$ is the indicator function.

In contrast, under IG1, a model succeeds as long as it can convince the interrogator that it is human. Hence, IG1 can be satisfied in two ways: either by mimicking **any plausible human user** ($\exists u \in U$), or by approximating a hypothetical **generic user** u_g. Although such an "average" user may not exist in reality, as commonly argued in ergonomics, design, and behavioral sciences (Rose, 2016; Bailey, 1982; Daniels, 1952), there may nevertheless exist users whose behavior aligns sufficiently closely with this generalized representation for a given task T. So, IG1 reduces behavioral modeling to reproducing generalized or sufficiently plausible human behavior, rather than modeling how each individual would perform a task in their own idiosyncratic way.

IG1 vs. IG2: Consider the earlier UAE-capital example from Section 3.1 as the following two tasks. Let T_1 denote the user-implicit query, *"What is the capital of the UAE?"*. Since no user is specified, the task can be assumed to target a plausible (u) or generic representative user (u_g), and both IG1 and IG2 would expect behavior leading to the generalized answer "Abu Dhabi". Let T_2 denote the user-specific query, *"What is the capital of the UAE according to Maya?"*. Here, the expected behavior may deviate from objective correctness, since the task requires modeling a specific user u's idiosyncratic reasoning and response. Thus, IG2 rewards fidelity to B^T_{Maya}, whereas IG1 may penalize such behavior for deviating from the generalized target behavior.

Practicality of IG2: As formalized in Equation 1, IG2 requires a model to faithfully reproduce the behavior of individual users across tasks. In practice, however, evaluating over the full space of users and tasks is computationally and empirically intractable. A practical approximation of IG2 can therefore be defined over subsets of users and tasks, namely $U_s \subseteq U$ and $\mathcal{T}_s \subseteq \mathcal{T}$. This enables the reuse of existing datasets while shifting the evaluation criterion from modeling generalized or representative behavior, as in IG1, toward modeling the behavior of all users within the subset ($\forall u \in U_s$). Furthermore, the notion of a user can itself be operationalized through demographic, behavioral, and sociocultural proxies, enabling more tractable approximations of the user space.

User Proxies: Directly modeling every individual user is often infeasible. Instead, prior work characterizes users through demographic, behavioral, or sociocultural *proxies* that capture shared behavioral patterns across groups (Adilazuarda et al., 2024), while the dimensions along which these groups differ are termed semantic proxies (Thompson et al., 2020). Let $\delta = \{\delta_1, \ldots, \delta_k\}$ denote the set of possible proxies, and let $\delta_u = \{\delta_{u,1}, \ldots, \delta_{u,k}\}$ denote the proxy configuration associated with user $u \in U$.

To operationalize user-specific behavioral modeling, we represent each user through a latent representation $z_u \in \mathbb{R}^d$, which encodes the user's behavioral, linguistic, cultural, and contextual characteristics. Under IG2, these representations are conditioned on user proxies, such that $z_u \sim p(z \mid \delta_u)$. Consequently, the user space U can be viewed as a distribution over latent user representations. In contrast, IG1 either conditions on a randomly selected plausible user $u \in U$, or collapses user diversity into a generalized or representative user representation $z_{u_g} = \mathbb{E}_{u \sim p_U}[z_u]$, corresponding to aggregate behavioral tendencies across users. Thus, the effective user representation under IG1 can be denoted as $z_{u^\star}$, where $u^\star \in U \cup \{u_g\}$.

Given these representations, $\mathcal{M}_\theta$ can be conditioned on z_u or δ_u through socio-demographic prompting (Saha et al., 2025b), representation steering (Chen et al., 2025), feedback-driven preference elicitation (Austin et al., 2024; Bose et al.), or personalized PEFT methods (Tan et al., 2024). Despite different mechanisms, these approaches all shift generation from generalized behavior toward individual-specific modeling, as prescribed by IG2.

Operationalizing IG2: One of the many possible operationalizations of IG2 can be through distributional behavioral modeling, as commonly done in recommendation and user modeling (Ricci et al., 2010; Zhang et al., 2019), where behavioral sequences are modeled as probability distributions.

These distributions can then be compared using standard divergence or similarity measures, such as KL divergence or Wasserstein distance (Kullback and Leibler, 1951; Villani et al., 2009). In essence, IG2 does not require entirely new modeling machinery. Rather, it changes the granularity of evaluation from generic or aggregate behavior to individual-conditioned behavior.

3.3 Imitation Game 2.0: In Practice

As Saha and Choudhury (2025) argue, generalization in LLMs differs fundamentally from traditional ML. In classical ML, generalization typically refers to learning from a set of tasks and transferring to unseen tasks with similar distributional characteristics (Vapnik, 2013; Goodfellow et al., 2016). LLMs, however, are trained as next-token predictors over large-scale behavioral traces generated by users (Brown et al., 2020; Radford et al., 2019). Consequently, their generalization should also be evaluated behaviorally: not merely whether models can solve unseen tasks, but whether they can adapt to and model the behavior of unseen users.

From this perspective, task-based evaluation captures a **weaker** form of generalization, since behavior is often treated as user-invariant. IG2 instead targets a **stronger** form: whether a model can infer how different individuals would behave from limited interaction history. This shifts evaluation from few-shot generalization over tasks to few-shot generalization over individuals and behavioral trajectories. Domains such as mental health, education, and personalized recommendation (Section 4) naturally require such adaptation, since appropriate responses depend on the user's goals, background, cognitive state, and interaction history. This, in turn, has direct implications for datasets, training paradigms, and evaluation pipelines.

Datasets and Benchmarks: IG2 datasets should capture behavior at multiple granularities: generic, group-specific, and individual-specific. Existing benchmarks such as MMLU (Hendrycks et al., 2021), GSM8K (Cobbe et al., 2021), and GLUE (Wang et al., 2019) primarily evaluate generic behavior through single expected answers or aggregate labels, while newer datasets increasingly target culturally grounded or subgroup-specific behavior (Koto et al., 2023; Montalan et al., 2024; Lent et al., 2024; Jin et al., 2024; Seth et al., 2024). IG2 benchmarks should go further by incorporating user-level histories, behavioral trajectories, and proxies (δ), such as geoculture, language, expertise, and situational constraints, while testing whether models preserve individual specificity in a sample-efficient manner (Jiang et al.; Guo et al., 2026; Nagarnaik and Thomas, 2015; Ko et al., 2022).

Training and Evaluation: IG2 also motivates objectives beyond standard next-token prediction (Brown et al., 2020; Radford et al., 2019). Optimizing only for aggregate likelihood may encourage convergence toward generalized behavioral distributions (u_g), thereby smoothing over user-specific variability (Xiao et al., 2026). Instead, IG2-style systems may require user-conditioned objectives (Salemi et al., 2024; Salemi and Zamani, 2025), cross-user contrastive learning (Gao et al., 2021), temporal-coherence objectives over evolving histories, and long-horizon behavioral prediction (Bose et al.). Correspondingly, evaluation should measure not only task performance, but also user alignment, cross-user sensitivity, and longitudinal consistency.

3.4 Imitation Game 2.0: Ethical Implications

Broadly, IG2 introduces ethical risks that differ from, and may amplify, those in IG1. By shifting from population-level plausibility to individual-level behavioral modeling, IG2 complicates notions of harm, privacy, bias, and consent. Harms may not reduce to aggregate categories such as racial, ethnic, or gender bias (Fang et al., 2024; Ranjan et al., 2024), since individuals may vary in what they perceive as harmful across contexts (Raymen, 2022; Starke et al., 2022). Moreover, IG2-style systems may expose sensitive information through proxies, histories, linguistic patterns, preferences, and behavioral trajectories, creating risks of re-identification (Lermen et al., 2026), surveillance, manipulation, proxy-based discrimination, vulnerability exploitation, and high-fidelity impersonation. Thus, such systems require safeguards around consent, privacy, user control, auditability, limits on behavioral inference, and restrictions on manipulative or impersonating uses. We discuss these risks in greater detail in Appendix A, and highlight domain-specific instances in the case studies below.

4 Case Studies

Beyond general AI systems, IG2 has implications across a broad class of human-centered AI systems (Theofanos et al., 2024; Jiang et al., 2024; Torkamaan et al., 2024; Pan et al., 2026), broadly categorized into: (i) **decision and support systems**, such as mental health, education, and scientific

discovery; (ii) **personalization and interaction systems**, such as recommendation and dialogue agents; and (iii) **social simulation systems**, such as governance and organizational modeling. Across these domains, success depends less on objective correctness alone and more on adapting to individuals, groups, contexts, and interaction histories. We discuss one representative case from each category below, with additional case studies in Appendix B.

4.1 Mental Health

This is a domain where the distinction between IG1 and IG2 is especially consequential. Unlike many healthcare screening tasks, counseling conversations rarely depend on factual correctness alone. Rather, they require interpreting a user's utterances in light of their history, emotional state, culture, trauma, and evolving behavioral patterns. Although current AI systems are used for mental-health assistance, therapist support, and patient/persona simulation (Le et al., 2026; Nguyen et al., 2025; Yang et al., 2025; Du et al., 2025), many are still evaluated through limited benchmarks that reward responses that appear generally supportive, empathetic, or psychologically plausible (Mandal et al., 2025; Malhotra et al., 2022). This aligns closely with IG1, where success means producing behavior that appears appropriate in the aggregate[5].

IG2 raises the bar by asking whether a system can model a particular user's behavioral trajectory over time. The same utterance, such as "I am tired of this," may signal routine frustration for one person, but distress or deterioration for another, depending on longitudinal context. Prior work has explored personalization through dialogue dynamics (Srivastava et al., 2023b), emotional states (Saha et al., 2022b,a), psychological traits (Srivastava et al., 2025), and personas (Yun et al., 2026). Yet systems may still overgeneralize across users or fail to track evolving behavioral states (Li et al., 2025; Yang et al., 2025). Thus, a system may pass IG1 by sounding supportive, while failing IG2 if it cannot preserve individual-level behavioral fidelity.

This shift also introduces distinct ethical risks. IG2-style mental-health systems require sensitive longitudinal data, increasing risks of privacy loss, re-identification, emotional manipulation, inappropriate dependency, and misuse of inferred mental states. Also, as behavioral fidelity improves, users may perceive such systems as substitutes for clinicians, even though they lack professional accountability, duty of care, and crisis-handling responsibility (Fiske et al., 2019; Stade et al., 2024; Zhang and Wang, 2024). Therefore, IG2 in mental health should not be treated as a path toward replacing experts, but as a framework for evaluating when personalized AI support is behaviorally faithful, clinically bounded, and ethically governed.

4.2 Recommendation Systems (Music)

Music recommender systems[6] aim to predict the next song a user would prefer, typically by learning from playlists, listening histories, artist preferences, and implicit feedback (Schedl et al., 2022). Under an IG1-style view, success is often measured through aggregate prediction accuracy or broad preference alignment. However, musical preference is deeply contextual, shaped by activity, occasion, environment, emotional state, culture, and geospatial signals (Kaminskas and Ricci, 2012; Zangerle et al., 2020; Schedl et al., 2018, 2014). Thus, a system may perform well on average while failing to capture how a particular user negotiates novelty, familiarity, diversity, coherence, mood, culture, and context (Novello and McKinney, 2007; Knees and Schedl, 2013; Song et al., 2012; Schedl et al., 2022).

IG2 reframes music recommendation as individual-specific behavioral modeling. Rather than asking whether a system recommends generally popular or statistically likely songs, IG2 asks whether it can model how a specific user forms and expresses musical taste over time. This aligns with growing calls in music information retrieval for user-centric modeling and evaluation, including individual taste modeling, sparse-feedback learning, cold-start handling, and user-experience-centered evaluation (Knijnenburg and Willemsen, 2015; Knees et al., 2019; Fazelnia et al., 2025). In this sense, IG2 formalizes the shift from aggregate preference prediction to individualized musical behavior modeling.

However, this shift also raises ethical risks. Hyper-personalized music systems could exploit affective states and vulnerabilities, repeatedly serving content that reinforces maladaptive listening patterns, compulsive use, or mood regulation loops (Reybrouck and Dyck, 2024; Lozon and Bensimon, 2023; Alluri et al., 2022). These risks parallel

[5] For a broader discussion of the dangers of AI in mental-health care, see https://hai.stanford.edu/news/exploring-the-dangers-of-ai-in-mental-health-care.

[6] As it is well developed, we focus on music as a representative case of recommender systems.

concerns in social media, where highly personalized systems have been linked to problematic or addiction-like use (Andreassen, 2015; Kuss and Griffiths, 2017; Shannon et al., 2022). IG2-style recommenders therefore require safeguards around user control, transparency, affect-sensitive personalization, and limits on manipulative optimization.

4.3 Social Simulation Systems

Although LLM-based social simulations aim to model individuals, groups, and societies (Mou et al., 2026), IG1-inspired systems often reduce people to personas, demographics, memory modules, or agentic roles (Argyle et al., 2023; Park et al., 2024a; Hong et al., 2023; Zhao et al., 2024; Riabi et al., 2026). As a result, success is judged by population-level plausibility, while agents often fail to preserve stable, idiosyncratic behavior over time (Mou et al., 2026).

IG2 reframes social simulation from modeling plausible populations to modeling the individuals whose interactions produce social patterns. This distinction is crucial because aggregate plausibility can obscure individual misrepresentation, stereotype reinforcement, or unstable identities. It also raises sharper ethical risks. Prior systems, such as predictive policing, have shown how behavioral models can create feedback loops when predictions shape the data they later learn from (Ensign et al., 2018), while COMPAS-style risk assessments highlight the dangers of opaque population-level proxies for complex human behavior (Angwin et al., 2022). Under IG2, high-fidelity individual models could further enable targeted simulation, prediction, or manipulation of specific people, requiring strong guardrails around privacy, consent, data governance, manipulation, and representational harm.

5 Open Questions and Conclusion

We discuss a few open questions for IG2 below:

1. Is language alone sufficient for IG2? Human behavior is inherently multimodal, involving language, gestures, facial expressions, embodiment, perception, and interaction with the physical world (McGurk and MacDonald, 1976; Clark, 1996; Barsalou, 2008). While our framing focuses specifically on NLP, fully realizing IG2 may ultimately require multimodal behavioral modeling, extending evaluation from linguistic alignment toward behavioral alignment more broadly. This naturally raises questions about the scope of NLP itself.

2. What counts as NLP? In this paper, we primarily adopt a functional and utilitarian view of NLP, where language serves as an interface between humans and computational systems. Other views instead study linguistic structure and representation (Pinker, 2003; Tomasello, 2005), or use computation to analyze communication, cognition, culture, and language change (Croft, 2000; Rogers et al., 2014; Mukherjee et al., 2025). How these perspectives relate in the IG2 framing remains an open interdisciplinary question.

3. Does passing IG2 imply General Intelligence (GI)? Unlike GI, which is typically framed around broad task competence (Goertzel and Pennachin, 2007; Legg and Hutter, 2007), IG2 is behavior-centric. Nevertheless, due to the interconnectedness of the world, faithfully modeling even a single individual may, in principle, require understanding broader social, cultural, and contextual dynamics (Giddens, 1984; Geertz, 1973; Bronfenbrenner, 1979; Lorenz, 1963; Mitchell, 2009). Consequently, strong behavioral modeling under IG2 may itself constitute a pathway toward more general forms of intelligence. However, whether passing IG2 necessarily implies GI remains an open question.

Conclusion: Recent discussions on linguistic diversity and representation show that NLP systems still struggle to capture the world's linguistic and cultural plurality (Choudhury, 2023). In this paper, we argued that this reflects a broader organic shift in NLP toward behavioral modeling, where language is understood as emergent and socially situated, a trend we also observe empirically in Section 2.4. Thus, IG2 is not a wholly new direction, but a more rigorous philosophical and theoretical framing of where the field is already moving. Whereas Turing's original philosophy, as reflected in IG1, rewards plausible generic human behavior, IG2 reframes progress around individual-specific behavioral fidelity across contexts and interactions, offering a more principled lens for evaluating whether NLP is truly advancing toward human-level language processing.

Limitations

This paper proposes IG2 as a philosophical and theoretical reframing of NLP's objectives, not as a benchmark, modeling pipeline, or evaluation protocol. We do not claim IG2 as the only lens through which progress in NLP should be measured. While we provide a framework, several aspects of it are

intentionally schematic and open to further development. For example, the formalization in Section 3.2 leaves the assessment function $\mathcal{A}$, the per-user similarity threshold ε, and the global passing threshold τ open; committing to specific instantiations is application-dependent and outside the scope of a position paper. How user behavior should be defined, what forms of evidence are appropriate for modeling it, and how behavioral similarity should be measured across tasks, domains, modalities, and interaction contexts likewise remain open.

A second limitation concerns the practical difficulty of modeling individual behavior. A user's language, preferences, beliefs, goals, social context, and interaction patterns may change over time, and many of these factors may not be observable through text alone. In realistic applications, systems may have access only to limited interaction histories or may need to rely on demographic, cultural, behavioral, or multimodal proxies. While such proxies can support more situated modeling, they remain incomplete and may collapse individuals back into broad categories, thereby reproducing the very abstraction that IG2 aims to move beyond. We partly discuss this as an open question in Section 5, and the ethical implications of it in Appendix A.

Third, behavioral fidelity should not be conflated with desirable system behavior. A model that accurately reflects a user's beliefs, preferences, or behavioral patterns may also reproduce that user's mistakes, harmful assumptions, biases, or vulnerable tendencies (see Appendix A). IG2 should therefore be understood as a framework for modeling, understanding, and adapting to users, not as a prescription to imitate all user behavior uncritically. Future work should further clarify the relationship between descriptive behavioral modeling, normative system objectives, and safeguards for safe and beneficial interaction.

Finally, the case studies discussed in this paper are intended to be illustrative rather than exhaustive. Mental health, music recommendation, and social simulation represent only a small subset of the domains in which IG2-style behavioral modeling may be relevant. Similarly, the ethical concerns discussed in the case studies and in the broader ethical analysis should not be viewed as a finite or complete list. As IG2-style systems are instantiated across new applications, additional risks, trade-offs, and governance challenges are likely to emerge. The historical analysis and case studies in this paper are therefore meant to motivate the IG2 perspective, rather than to establish that any single implementation will generalize uniformly across all NLP applications.

References

Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Shivdutt Singh, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. 2024. Towards measuring and modeling "culture" in LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15763–15784, Miami, Florida, USA. Association for Computational Linguistics.

Vinoo Alluri, Anant Mittal, Azhagammal SC, Jonna K. Vuoskoski, and Suvi Saarikallio. 2022. Maladaptive music listening strategies are modulated by individual traits. *Psychology of Music*, 50(6):030573562110650.

John R Anderson, C Franklin Boyle, and Brian J Reiser. 1985. Intelligent tutoring systems. *Science*, 228(4698):456–462.

Cecilie Schou Andreassen. 2015. Online social network site Addiction: A Comprehensive Review. *Current Addiction Reports*, 2(2):175–184.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2022. Machine bias. In *Ethics of data and analytics*, pages 254–264. Auerbach Publications.

Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

David Austin, Anton Korikov, Armin Toroghi, and Scott Sanner. 2024. Bayesian optimization with llm-based acquisition functions for natural language preference elicitation. In *Proceedings of the 18th ACM Conference on Recommender Systems*, RecSys '24, page 74–83, New York, NY, USA. Association for Computing Machinery.

Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. 2025. Researchagent: Iterative research idea generation over scientific literature with large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 6709–6738. Association for Computational Linguistics.

Robert W. Bailey. 1982. *Human performance engineering : a guide for system designers / Robert W. Bailey.* Prentice-Hall, Englewood Cliffs, N.J.

Lawrence W Barsalou. 2008. Grounded cognition. *Annu. Rev. Psychol.*, 59(1):617–645.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, New York. Association for Computer Machinery – ACM.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155.

Yochai Benkler, Robert Faris, and Hal Roberts. 2018. *Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics*. Oxford University Press.

Avinandan Bose, Shuyue Stella Li, Faeze Brahman, Pang Wei Koh, Simon Shaolei Du, Yulia Tsvetkov, Maryam Fazel, Lin Xiao, and Asli Celikyilmaz. Cold-start personalization via training-free priors from structured world models. In *ICLR 2026 Workshop on Lifelong Agents: Learning, Aligning, Evolving*.

Urie Bronfenbrenner. 1979. *The ecology of human development: Experiments by nature and design*. Harvard university press.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, and 1 others. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Micah Carroll, Alan Chan, Henry Ashton, and David Krueger. 2023. Characterizing manipulation from ai systems. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '23, New York, NY, USA. Association for Computing Machinery.

Eugene Charniak. 1997. Statistical parsing with a context-free grammar and word statistics. *AAAI/IAAI*, 2005(598-603):18.

Runjin Chen, Andy Arditi, Henry Sleight, Owain Evans, and Jack Lindsey. 2025. Persona vectors: Monitoring and controlling character traits in language models. *arXiv preprint arXiv:2507.21509*.

Stanley F Chen and Joshua Goodman. 1999. An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 13(4):359–394.

Robert Chesney and Danielle Keats Citron. 2019. Deep fakes: A looming challenge for privacy, democracy, and national security. *California Law Review*, 107(6):1753–1819. U of Texas Law, Public Law Research Paper No. 692; U of Maryland Legal Studies Research Paper No. 2018-21.

Noam Chomsky. 1957. *Syntactic Structures*. Mouton and Co., The Hague.

Monojit Choudhury. 2023. Generative ai has a language problem. *Nature human behaviour*, 7(11):1802–1803.

Zhendong Chu, Shen Wang, Jian Xie, Tinghui Zhu, Yibo Yan, Jingheng Ye, Aoxiao Zhong, Xuming Hu, Jing Liang, Philip S. Yu, and Qingsong Wen. 2025. LLM agents for education: Advances and applications. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 13782–13810, Suzhou, China. Association for Computational Linguistics.

Herbert H Clark. 1996. *Using language*. Cambridge university press.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *Preprint*, arXiv:2110.14168.

Michael Cole. 1998. *Cultural psychology: A once and future discipline*. Harvard university press.

W. Croft. 2000. *Explaining Language Change: An Evolutionary Approach*. Explaining Language Change: An Evolutionary Approach. Longman.

Gilbert S Daniels. 1952. The" average man"? Technical report.

F. de Saussure, W. Baskin, P. Meisel, and H. Saussy. 2011. *Course in General Linguistics*. Columbia University Press.

Zhuoyun Du, Lujie Zheng, Renjun Hu, Yuyang Xu, Xiawei Li, Ying Sun, Wei Chen, Jian Wu, Haolei Cai, and Haochao Ying. 2025. LLMs can simulate standardized patients via agent coevolution. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17278–17306, Vienna, Austria. Association for Computational Linguistics.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, page 214–226, New York, NY, USA. Association for Computing Machinery.

Danielle Ensign, Sorelle A Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian. 2018. Runaway feedback loops in predictive policing. In *Conference on fairness, accountability and transparency*, pages 160–171. PMLR.

Kawin Ethayarajh and Dan Jurafsky. 2020. Utility is in the eye of the user: A critique of NLP leaderboards. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4846–4853, Online. Association for Computational Linguistics.

Fahim Faisal, Orevaoghene Ahia, Aarohi Srivastava, Kabir Ahuja, David Chiang, Yulia Tsvetkov, and Antonios Anastasopoulos. 2024. DIALECTBENCH: An NLP benchmark for dialects, varieties, and closely-related languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14412–14454, Bangkok, Thailand. Association for Computational Linguistics.

Xiao Fang, Shangkun Che, Minjia Mao, Hongzhe Zhang, Ming Zhao, and Xiaohang Zhao. 2024. Bias of ai-generated content: an examination of news produced by large language models. *Scientific Reports*, 14(1):5224.

Ghazal Fazelnia, Sanket Gupta, Claire Keum, Mark Koh, Timothy Heath, Guillermo Carrasco Hernández, Stephen Xie, Nandini Singh, Ian Anderson, Maya Hristakeva, Petter Pehrson Skidén, and Mounia Lalmas. 2025. Generalized user representations for large-scale recommendations and downstream tasks. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems*, RecSys '25, page 962–966, New York, NY, USA. Association for Computing Machinery.

Edward A Feigenbaum and Pamela McCorduck. 1984. *The fifth generation*. Pan Books London.

John Firth. 1957. A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*, pages 10–32.

Amelia Fiske, Peter Henningsen, and Alena Buyx. 2019. Your robot therapist will see you now: Ethical implications of embodied artificial intelligence in psychiatry, psychology, and psychotherapy. *J Med Internet Res*, 21(5):e13216.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 6894–6910.

Clifford Geertz. 1973. *The Interpretation of Cultures: Selected Essays*. Basic Books, New York.

Ine Gevers, Victor De Marez, Jens Van Nooten, Jens Lemmens, Andriy Kosar, Ehsan Lotfi, Nikolay Banar, Pieter Fivez, Luna De Bruyne, and Walter Daelemans. 2025. In benchmarks we trust ... or not? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 23662–23676, Suzhou, China. Association for Computational Linguistics.

Anthony Giddens. 1984. *The constitution of society: Outline of the theory of structuration*. Univ of California Press.

Ben Goertzel and Cassio Pennachin. 2007. *Artificial general intelligence*, volume 2. Springer.

Josh A Goldstein, Girish Sastry, Micah Musser, Renee DiResta, Matthew Gentzel, and Katerina Sedova. 2023. Generative language models and automated influence operations: Emerging threats and potential mitigations. *arXiv preprint arXiv:2301.04246*, 1.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. 2016. *Deep learning*, volume 1. MIT press Cambridge.

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, Felix Weissenberger, Keran Rong, Ryutaro Tanno, and 1 others. 2025. Towards an ai co-scientist. *arXiv preprint arXiv:2502.18864*.

Qianyun Guo, Yibo Li, Yue Liu, and Bryan Hooi. 2026. Towards realistic personalization: Evaluating long-horizon preference following in personalized user-llm interactions. *arXiv preprint arXiv:2603.04191*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, and 1 others. 2023. Metagpt: Meta programming for a multi-agent collaborative framework. In *The twelfth international conference on learning representations*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 43(2).

Dieuwke Hupkes, Mario Giulianelli, Verna Dankers, Mikel Artetxe, Yanai Elazar, Tiago Pimentel, Christos Christodoulopoulos, Karim Lasri, Naomi Saphra, Arabella Sinclair, Dennis Ulmer, Florian Schottmann, Khuyagbaatar Batsuren, Kaiser Sun, Koustuv Sinha, Leila Khalatbari, Maria Ryskina, Rita Frieske, Ryan Cotterell, and Zhijing Jin. 2023. A taxonomy and review of generalization research in NLP. *Nature Machine Intelligence*, 5(10):1161–1174.

John Hutchins. 2004. Two precursors of machine translation: Artsrouni and trojanskij. *International Journal of Translation*, 16(1):11–31.

Bowen Jiang, Zhuoqun Hao, Young Min Cho, Bryan Li, Yuan Yuan, Sihao Chen, Lyle Ungar, Camillo Jose Taylor, and Dan Roth. Know me, respond to me: Benchmarking llms for dynamic user profiling and personalized responses at scale. In *Second Conference on Language Modeling*.

Tingting Jiang, Zhumo Sun, Shiting Fu, and Yan Lv. 2024. Human-ai interaction research agenda: A user-centered perspective. *Data and Information Management*, 8(4):100078.

Jiho Jin, Jiseon Kim, Nayeon Lee, Haneul Yoo, Alice Oh, and Hwaran Lee. 2024. KoBBQ: Korean bias benchmark for question answering. *Transactions of the Association for Computational Linguistics*, 12:507–524.

Cameron R. Jones and Benjamin K. Bergen. 2024. Does GPT-4 pass the Turing test? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5183–5210, Mexico City, Mexico. Association for Computational Linguistics.

Cameron R Jones and Benjamin K Bergen. 2025. Large language models pass the turing test. *arXiv preprint arXiv:2503.23674*.

Daniel Jurafsky and James H. Martin. 2026. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, with Language Models*, 3rd edition. Online manuscript released January 6, 2026.

Marius Kaminskas and Francesco Ricci. 2012. Contextual music information retrieval and recommendation: State of the art and challenges. *Computer Science Review*, 6(2):89–119.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online. Association for Computational Linguistics.

Styliani Kleanthous, Maria Kasinidou, Pinar Barlas, and Jahna Otterbacher. 2021. Perception of fairness in algorithmic decisions: Future developers' perspective. *Patterns*, 3.

Peter Knees and Markus Schedl. 2013. A survey of music similarity and recommendation from music context data. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 10(1):1–21.

Peter Knees, Markus Schedl, Bruce Ferwerda, and Audrey Laplante. 2019. User awareness in music recommender systems. *Personalized human-computer interaction*, 223.

Bart P Knijnenburg and Martijn C Willemsen. 2015. Evaluating recommender systems with user experiments. In *Recommender systems handbook*, pages 309–352. Springer.

Hyeyoung Ko, Suyeon Lee, Yoonseo Park, and Anna Choi. 2022. A survey of recommendation systems: recommendation models, techniques, and application fields. *Electronics*, 11(1):141.

Kenneth R Koedinger, John C Stamper, Elizabeth A McLaughlin, and Tristan Nixon. 2013. Using data-driven discovery of better student models to improve student learning. In *International conference on artificial intelligence in education*, pages 421–430. Springer.

Fajri Koto, Nurul Aisyah, Haonan Li, and Timothy Baldwin. 2023. Large language models only pass primary school exams in Indonesia: A comprehensive test on IndoMMLU. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12359–12374, Singapore. Association for Computational Linguistics.

Thomas S Kuhn and Ian Hacking. 1970. *The structure of scientific revolutions*, volume 2. University of Chicago press Chicago.

Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

Daria Kuss and Mark Griffiths. 2017. Social networking sites and Addiction: Ten lessons learned. *International Journal of Environmental Research and Public Health*, 14(3):311.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields:

Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, page 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Huy M. Le, Dat Tien Nguyen, Ngan T. T. Vo, Tuan D. Q. Nguyen, Nguyen Le Binh, Duy Minh Ho Nguyen, Daniel Sonntag, Lizi Liao, and Binh T. Nguyen. 2026. Reinforce trustworthiness in multimodal emotional support system. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(37):31474–31482.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *nature*, 521(7553):436–444.

Shane Legg and Marcus Hutter. 2007. Universal intelligence: A definition of machine intelligence. *Minds and machines*, 17(4):391–444.

Heather Lent, Kushal Tatariya, Raj Dabre, Yiyi Chen, Marcell Fekete, Esther Ploeger, Li Zhou, Ruth-Ann Armstrong, Abee Eijansantos, Catriona Malau, Hans Erik Heje, Ernests Lavrinovics, Diptesh Kanojia, Paul Belony, Marcel Bollmann, Loïc Grobol, Miryam de Lhoneux, Daniel Hershcovich, Michel DeGraff, and 2 others. 2024. CreoleVal: Multilingual Multitask Benchmarks for Creoles. *Transactions of the Association for Computational Linguistics*, 12:950–978.

Simon Lermen, Daniel Paleka, Joshua Swanson, Michael Aerni, Nicholas Carlini, and Florian Tramèr. 2026. Large-scale online deanonymization with llms. *arXiv preprint arXiv:2602.16800*.

Hao Li, Chenghao Yang, An Zhang, Yang Deng, Xiang Wang, and Tat-Seng Chua. 2025. Hello again! LLM-powered personalized agent for long-term dialogue. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5259–5276, Albuquerque, New Mexico. Association for Computational Linguistics.

Xiaoyuan Li, Wenjie Wang, Moxin Li, Junrong Guo, Yang Zhang, and Fuli Feng. 2024. Evaluating mathematical reasoning of large language models: A focus on error identification and correction. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11316–11360, Bangkok, Thailand. Association for Computational Linguistics.

Ben Liu, Jihai Zhang, Fangquan Lin, Xu Jia, and Min Peng. 2025. One size doesn't fit all: A personalized conversational tutoring agent for mathematics instruction. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 2401–2410.

Yu Lu Liu, Su Lin Blodgett, Jackie Cheung, Q. Vera Liao, Alexandra Olteanu, and Ziang Xiao. 2024a. ECBD: Evidence-centered benchmark design for NLP. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16349–16365, Bangkok, Thailand. Association for Computational Linguistics.

Zhengyuan Liu, Stella Xin Yin, Geyu Lin, and Nancy F. Chen. 2024b. Personality-aware student simulation for conversational intelligent tutoring systems. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 626–642, Miami, Florida, USA. Association for Computational Linguistics.

Edward N. Lorenz. 1963. Deterministic nonperiodic flow. *Journal of Atmospheric Sciences*, 20(2):130 – 141.

Jeffrey Lozon and Moshe Bensimon. 2023. The mutual enhancing effect of music and drugs and its eventual cessation in the lives of people coping with substance abuse: A qualitative inquiry. *Current Psychology*.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. 2024. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*.

Kevin Lynch. 1964. *The image of the city*. MIT press.

Ganeshan Malhotra, Abdul Waheed, Aseem Srivastava, Md Shad Akhtar, and Tanmoy Chakraborty. 2022. Speaker and time-aware joint contextual learning for dialogue-act classification in counselling conversations. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, WSDM '22, page 735–745, New York, NY, USA. Association for Computing Machinery.

Aishik Mandal, Prottay Kumar Adhikary, Hiba Arnaout, Iryna Gurevych, and Tanmoy Chakraborty. 2025. A comprehensive review of datasets for clinical mental health ai systems. *Preprint*, arXiv:2508.09809.

Christopher Manning and Hinrich Schutze. 1999. *Foundations of statistical natural language processing*. MIT press.

Kaushal Kumar Maurya, Kv Aditya Srivatsa, Kseniia Petukhova, and Ekaterina Kochmar. 2025. Unifying AI tutor evaluation: An evaluation taxonomy for pedagogical ability assessment of LLM-powered AI tutors. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1234–1251, Albuquerque, New Mexico. Association for Computational Linguistics.

Harry McGurk and John MacDonald. 1976. Hearing lips and seeing voices. *Nature*, 264(5588):746–748.

McKinsey & Company. 2023. The economic potential of generative ai: The next productivity frontier. Technical report, McKinsey & Company.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6).

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *Preprint*, arXiv:1301.3781.

Marvin Minsky and 1 others. 1974. A framework for representing knowledge.

Melanie Mitchell. 2009. *Complexity: A guided tour*. Oxford university press.

Jann Railey Montalan, Jian Gang Ngui, Wei Qi Leong, Yosephine Susanto, Hamsawardhini Rengarajan, Alham Fikri Aji, and William-Chandra Tjhi. 2024. Kalahi: A handcrafted, grassroots cultural LLM evaluation suite for filipino. In *Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation, PACLIC 2024, Tokyo, Japan, December 7-9, 2024*, pages 497–523. Tokyo University of Foreign Studies.

Xinyi Mou, Xuanwen Ding, Qi He, Liang Wang, Jingcong Liang, Xinnong Zhang, Libo Sun, Jiayu Lin, Jie Zhou, Huang Xuanjing, and Zhongyu Wei. 2026. From individual to society: A survey on social simulation driven by large language model-based agents. *ACM Comput. Surv.*, 58(11).

Sourabrata Mukherjee, Atharva Mehta, Sougata Saha, Akhil Arora, and Monojit Choudhury. 2025. Women, infamous, and exotic beings: A comparative study of honorific usages in Wikipedia and LLMs for Bengali and Hindi. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 19103–19126, Suzhou, China. Association for Computational Linguistics.

Paritosh Nagarnaik and A Thomas. 2015. Survey on recommendation system methods. In *2015 2nd international conference on electronics and communication systems (ICECS)*, pages 1603–1608. IEEE.

Allen Newell and Herbert A. Simon. 1976. Computer science as empirical inquiry: symbols and search. *Commun. ACM*, 19(3):113–126.

Viet Cuong Nguyen, Mohammad Taher, Dongwan Hong, Vinicius Konkolics Possobom, Vibha Thirunellayi Gopalakrishnan, Ekta Raj, Zihang Li, Heather J. Soled, Michael L. Birnbaum, Srijan Kumar, and Munmun De Choudhury. 2025. Do large language models align with core mental health counseling competencies? In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7503–7526, Albuquerque, New Mexico. Association for Computational Linguistics.

Alberto Novello and Martin F McKinney. 2007. Assessment of perceptual music similarity. In *ISMIR*, pages 111–112.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Hanxi Pan, Wei Xu, Mowei Shen, and Zaifeng Gao. 2026. Human-centered artificial social intelligence (hc-asi). In *Handbook of human-centered artificial intelligence*, pages 1–51. Springer.

Saurabh Kumar Pandey, Harshit Budhiraja, Sougata Saha, and Monojit Choudhury. 2025. CULTURALLY YOURS: A reading assistant for cross-cultural content. In *Proceedings of the 31st International Conference on Computational Linguistics: System Demonstrations*, pages 208–216, Abu Dhabi, UAE. Association for Computational Linguistics.

Joon Sung Park, Carolyn Q Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S Bernstein. 2024a. Generative agent simulations of 1,000 people. *arXiv preprint arXiv:2411.10109*, 52.

Peter S. Park, Simon Goldstein, Aidan O'Gara, Michael Chen, and Dan Hendrycks. 2024b. AI deception: A survey of examples, risks, and potential solutions. *Patterns*, 5(6):100988.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Steven Pinker. 2003. *The language instinct: How the mind creates language*. Penguin uK.

Michael Polanyi. 2009. The tacit dimension. In *Knowledge in organisations*, pages 135–146. Routledge.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI*. Accessed: 2024-11-15.

Rajesh Ranjan, Shailja Gupta, and Surya Narayan Singh. 2024. A comprehensive survey of bias in llms: Current landscape and future directions. *arXiv preprint arXiv:2409.16430*.

Thomas Raymen. 2022. *Social Harm and its Relationshp to Human Subjectivity*. Routledge.

Mark Reybrouck and Edith Van Dyck. 2024. Is music a drug? how music listening may trigger neurochemical responses in the brain. *Musicae scientiae*, 28(4).

Arij Riabi, Sougata Saha, and Monojit Choudhury. 2026. The anthropology of food: How NLP can help us unravel the food cultures of the world. In *Proceedings of the First Workshop on Multilingual Multicultural Evaluation*, pages 76–98, Rabat, Morocco. Association for Computational Linguistics.

Francesco Ricci, Lior Rokach, and Bracha Shapira. 2010. Introduction to recommender systems handbook. In *Recommender systems handbook*, pages 1–35. Springer.

Luc Rocher, Julien M. Hendrickx, and Yves-Alexandre de Montjoye. 2019. Estimating the success of re-identifications in incomplete datasets using generative models. *Nature Communications*, 10.

Everett M Rogers, Arvind Singhal, and Margaret M Quinlan. 2014. Diffusion of innovations. In *An integrated approach to communication theory and research*, pages 432–448. Routledge.

Barbara Rogoff. 2003. *The cultural nature of human development*. Oxford university press.

Eleanor Rosch and Carolyn B Mervis. 1975. Family resemblances: Studies in the internal structure of categories. *Cognitive psychology*, 7(4):573–605.

Todd Rose. 2016. *The end of average: How to succeed in a world that values sameness*. Penguin UK.

Sougata Saha and Monojit Choudhury. 2025. User behavior prediction as a generic, robust, scalable, and low-cost evaluation strategy for estimating generalization in llms. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 11047–11065.

Sougata Saha, Saurabh Kumar Pandey, and Monojit Choudhury. 2025a. Meta-cultural competence: Climbing the right hill of cultural awareness. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8025–8042, Albuquerque, New Mexico. Association for Computational Linguistics.

Sougata Saha, Saurabh Kumar Pandey, Harshit Gupta, and Monojit Choudhury. 2025b. Reading between the lines: Can LLMs identify cross-cultural communication gaps? In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8043–8067, Albuquerque, New Mexico. Association for Computational Linguistics.

Tulika Saha, Vaibhav Gakhreja, Anindya Sundar Das, Souhitya Chakraborty, and Sriparna Saha. 2022a. Towards motivational and empathetic response generation in online mental health support. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 2650–2656, New York, NY, USA. Association for Computing Machinery.

Tulika Saha, Saichethan Reddy, Anindya Das, Sriparna Saha, and Pushpak Bhattacharyya. 2022b. A shoulder to cry on: Towards a motivational virtual assistant for assuaging mental agony. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2436–2449, Seattle, United States. Association for Computational Linguistics.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2024. Lamp: When large language models meet personalization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7370–7392.

Alireza Salemi and Hamed Zamani. 2025. LaMP-QA: A benchmark for personalized long-form question answering. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1139–1159, Suzhou, China. Association for Computational Linguistics.

Saskia Sassen. 1991. The global city. *New York*.

Nripsuta Ani Saxena, Karen Huang, Evan DeFilippis, Goran Radanovic, David C. Parkes, and Yang Liu. 2020. How do fairness definitions fare? testing public attitudes towards three algorithmic definitions of fairness in loan allocations. *Artif. Intell.*, 283:103238.

Markus Schedl, Peter Knees, Brian McFee, and Dmitry Bogdanov. 2022. *Music Recommendation Systems: Techniques, Use Cases, and Challenges*, pages 927–971. Springer US, New York, NY.

Markus Schedl, Andreu Vall, and Katayoun Farrahi. 2014. User geospatial context for music recommendation in microblogs. SIGIR '14, page 987–990, New York, NY, USA. Association for Computing Machinery.

Markus Schedl, Hamed Zamani, Ching-Wei Chen, Yashar Deldjoo, and Mehdi Elahi. 2018. Current challenges and visions in music recommender systems research. *International Journal of Multimedia Information Retrieval*, 7(2):95–116.

Agrima Seth, Sanchit Ahuja, Kalika Bali, and Sunayana Sitaram. 2024. DOSA: A dataset of social artifacts from different Indian geographical subcultures. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5323–5337, Torino, Italia. ELRA and ICCL.

Holly Shannon, Katie Bush, Paul J Villeneuve, Kim Gc Hellemans, and Synthia Guimond. 2022. Problematic social media use in adolescents and young Adults: Systematic review and Meta-analysis. *JMIR Mental Health*, 9(4):e33450.

Renee Shelby, Shalaleh Rismani, Kathryn Henne, AJung Moon, Negar Rostamzadeh, Paul Nicholas, N'Mah Yilla-Akbari, Jess Gallegos, Andrew Smart, Emilio Garcia, and 1 others. 2023. Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 723–741.

Stuart M. Shieber. 1994. Lessons from a restricted turing test. *Commun. ACM*, 37(6):70–78.

Paul Smolensky. 1988. On the proper treatment of connectionism. *Behavioral and brain sciences*, 11(1):1–23.

Y. Song, Simon Dixon, and Marcus Thomas Pearce. 2012. A survey of music recommendation systems and future perspectives.

Shashank Sonkar, Naiming Liu, Debshila Mallick, and Richard Baraniuk. 2023. CLASS: A design framework for building intelligent tutoring systems based on learning science principles. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1941–1961, Singapore. Association for Computational Linguistics.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, and 431 others. 2023a. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*. Featured Certification.

Aseem Srivastava, Tanya Gupta, Alison Cerezo, Sarah Peregrine (Grin) Lord, Md Shad Akhtar, and Tanmoy Chakraborty. 2025. Critical behavioral traits foster peer engagement in online mental health communities. *PLOS ONE*, 20(1):1–17.

Aseem Srivastava, Ishan Pandey, Md Shad Akhtar, and Tanmoy Chakraborty. 2023b. Response-act guided reinforced dialogue generation for mental health counseling. In *Proceedings of the ACM Web Conference 2023*, WWW '23, page 1118–1129, New York, NY, USA. Association for Computing Machinery.

Elizabeth C. Stade, Shannon W. Stirman, Lyle H. Ungar, and 1 others. 2024. Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. *npj Mental Health Research*, 3(12).

Christopher Starke, Janine Baleis, Birte Keller, and Frank Marcinkowski. 2022. Fairness perceptions of algorithmic decision-making: A systematic review of the empirical literature. *Big Data & Society*, 9(2):20539517221115189.

Haoyang Su, Renqi Chen, Shixiang Tang, Zhenfei Yin, Xinzhe Zheng, Jinzhe Li, Biqing Qi, Qi Wu, Hui Li, Wanli Ouyang, and 1 others. 2025. Many heads are better than one: Improved scientific idea generation by a llm-based multi-agent system. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28201–28240.

Zhaoxuan Tan, Qingkai Zeng, Yijun Tian, Zheyuan Liu, Bing Yin, and Meng Jiang. 2024. Democratizing large language models via personalized parameter-efficient fine-tuning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6476–6491.

Simone Tedeschi, Johan Bos, Thierry Declerck, Jan Hajič, Daniel Hershcovich, Eduard Hovy, Alexander Koller, Simon Krek, Steven Schockaert, Rico Sennrich, Ekaterina Shutova, and Roberto Navigli. 2023. What's the meaning of superhuman performance in today's NLU? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12471–12491, Toronto, Canada. Association for Computational Linguistics.

Mary Frances Theofanos, Yee-Yin Choong, and Theodore Jensen. 2024. Ai use taxonomy: A human-centered approach.

B Thompson, SG Roberts, and G Lupyan. 2020. Cultural influences on word meanings revealed through large-scale semantic alignment. nature human behaviour, 4 (10), 1029–1038.

M. Tomasello. 2005. *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Harvard.

Helma Torkamaan, Mohammad Tahaei, Stefan Buijsman, Ziang Xiao, Daricia Wilkinson, and Bart P Knijnenburg. 2024. The role of human-centered ai in user modeling, adaptation, and personalization—models, frameworks, and paradigms. In *A human-centered perspective of intelligent personalized environments and systems*, pages 43–84. Springer.

Angeliki Tsapanou, Anastasia Bouka, Angeliki Papadopoulou, Christina Vamvatsikou, Dionisia Mikrouli, Eirini Theofila, Kassandra Dionysopoulou, Konstantina Kortseli, Panagiota Lytaki, Theoni Myrto Spyridonidi, and 1 others. 2025. Application of artificial intelligence tools for social and psychological enhancement of students with autism spectrum disorder: A systematic review. *Brain Sciences*, 16(1):56.

A. M. Turing. 1950. Computing machinery and intelligence. *Mind*, 59(236):433–460.

Amos Tversky and Daniel Kahneman. 1973. Availability: A heuristic for judging frequency and probability. *Cognitive psychology*, 5(2):207–232.

Amos Tversky and Daniel Kahneman. 1981. The framing of decisions and the psychology of choice. *science*, 211(4481):453–458.

Vladimir Vapnik. 2013. *The nature of statistical learning theory*. Springer science & business media.

Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness*, FairWare '18, page 1–7, New York, NY, USA. Association for Computing Machinery.

Cédric Villani and 1 others. 2009. *Optimal transport: old and new*, volume 338. Springer.

Lev S Vygotsky. 1978. *Mind in society: The development of higher psychological processes*, volume 86. Harvard university press.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Hanyu Wang, Tianfan Fu, Yilun Du, and 1 others. 2023. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60.

Xinyu Jessica Wang, Christine P Lee, and Bilge Mutlu. 2025. Learnmate: Enhancing online education with llm-powered personalized learning plans and support. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–10.

Kevin Warwick and Huma Shah. 2016. Can machines think? a report on turing test experiments at the royal society. *Journal of experimental & Theoretical artificial Intelligence*, 28(6):989–1007.

Joseph Weizenbaum. 1966. Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45.

Terry Winograd. 1972. Understanding natural language. *Cognitive psychology*, 3(1):1–191.

Ludwig Wittgenstein. 1953. *Philosophical Investigations*. Basil Blackwell, Oxford.

Beverly Park Woolf. 2010. *Building intelligent interactive tutors: Student-centered strategies for revolutionizing e-learning*. Morgan Kaufmann.

Weiqi Wu, Hongqiu Wu, and Hai Zhao. 2025. X-TURING: Towards an enhanced and efficient Turing test for long-term dialogue agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5874–5889, Vienna, Austria. Association for Computational Linguistics.

Yunze Xiao, Vivienne J Zhang, Chenghao Yang, Ningshan Ma, Weihao Xuan, and Jen-tse Huang. 2026. The chameleon's limit: Investigating persona collapse and homogenization in large language models. *arXiv preprint arXiv:2604.24698*.

Yizhe Yang, Palakorn Achananuparp, Heyan Huang, Jing Jiang, Nicholas Gabriel Lim, Cameron Tan Shi Ern, Phey Ling Kit, Jenny Giam Xiuhui, John Pinto, and Ee-Peng Lim. 2025. Consistent client simulation for motivational interviewing-based counseling. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20959–20998, Vienna, Austria. Association for Computational Linguistics.

Soyoung Yun, Minjoo Kim, Yeohyang Kim, and Hayoung Oh. 2026. Persona drift for adaptive flow in stage-aware cbt chatbots. In *Companion of the 2025 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp Companion '25, page 1645–1651, New York, NY, USA. Association for Computing Machinery.

Eva Zangerle, Martin Pichl, and Markus Schedl. 2020. User models for culture-aware music recommendation: Fusing acoustic and cultural cues. *Transactions of the International Society for Music Information Retrieval*.

Shuai Zhang, Lina Yao, Aixin Sun, and Yi Tay. 2019. Deep learning based recommender system: A survey and new perspectives. *ACM computing surveys (CSUR)*, 52(1):1–38.

Zhihui Zhang and Jing Wang. 2024. Can ai replace psychotherapists? exploring the future of mental health care. *Frontiers in Psychiatry*, Volume 15 - 2024.

Qinlin Zhao, Jindong Wang, Yixuan Zhang, Yiqiao Jin, Kaijie Zhu, Hao Chen, and Xing Xie. 2024. Competeai: understanding the competition dynamics of large language model-based agents. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Shoshana Zuboff. 2019. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. PublicAffairs, New York.

A Ethical Implications

In this section, we explore ethical concerns that arise from shifting from population-level modeling (IG1) to an individual-level NLP paradigm (IG2). We focus on six prominent risks, while recognizing that these are not exhaustive. In fact, the space will not allow us to discuss whether documented ethical risks in IG1 transfer to IG2 and, if so, in what format. We start the discussion questioning whether language alone could model human behavior, and if so with what fidelity. Then we discuss the subjectivity of harm and the ability of AI systems to infer behavior that could harm individuals. Next, we discuss privacy and confidentiality, bias and discrimination, and manipulation and persuasion. Finally, we discuss the appropriateness and the need for governance mechanisms for the development and use of IG2-derived AI systems.

A.1 Fidelity & Language Modeling Limitations

Perhaps a natural concern of AI systems developed under IG2 is whether a) the systems can "truly"

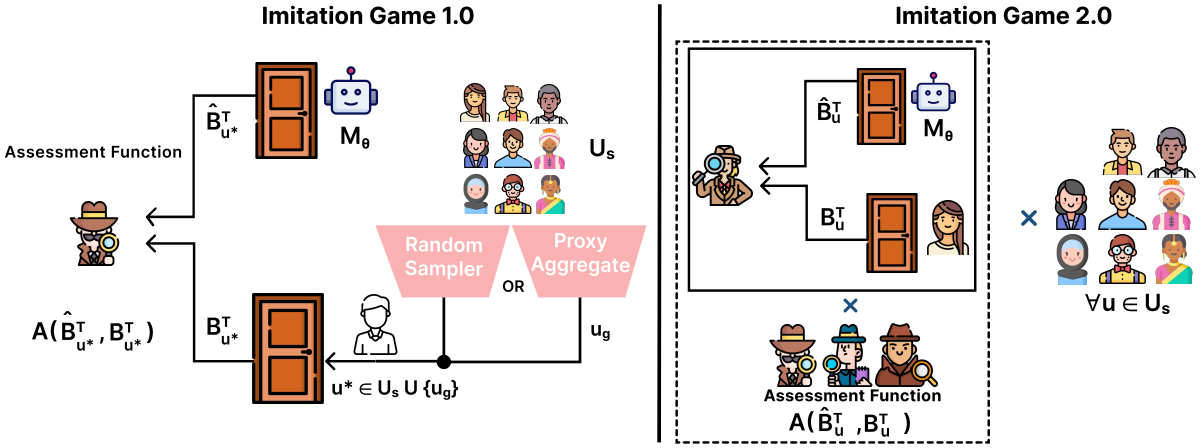

Figure 2: An illustrative comparison between Imitation Game 1.0 and the proposed Imitation Game 2.0.

model individual behavior and most importantly b) if language is sufficient for modeling behavior. The first question raises concerns about fidelity, i.e., whether models have the ability to generate outputs that are highly accurate, realistic, and faithful to the original data (Huang et al., 2025). Although technical interventions could address this issue, b) is a much bigger one. Human behavior is a result of biological, psychological, social, and environmental factors that language alone does not model. One might argue that proxies of these factors could be captured in AI training data, however, we argue that while machines are powerful tools for modeling data statistical patterns to allow individual personalization, AI developers should recognize the limitations of using one human cognitive ability, that is, language, to model a complex phenomenon such as behavior.

A.2 The Subjectivity of Harm Under IG2

Various forms of harm have been documented in AI systems. In (Shelby et al., 2023), the authors identified five major categories: representational (e.g., stereotyping social groups), allocative (e.g., opportunity and economic loss), quality of service (e.g., alienation), interpersonal harms (e.g., privacy violations) and social systems harms (e.g., cultural harm). As noticeable, these harms are often defined at the population level, as expected in IG1. Therefore, changing from an IG1 paradigm to an IG2 paradigm will inevitably affect perceptions, definitions, and operationalization of harm mitigation strategies. First, the perception of harm at the individual level will be highly subjective. Individual behavioral patterns, trajectories, and proxies could affect this perception. For example, studies have found that the perception of fairness varies between demographics, AI literacy, culture (Saxena et al., 2020; Starke et al., 2022), and even AI developer experience (Kleanthous et al., 2021). Thus, AI developers in IG2 should develop mechanisms to account for individual harm subjectivity.

A.3 Behavioral Inference Harm

A latent ethical risk in IG2 is the ability of AI systems to infer harmful behaviors. The power of AI lies in its ability to learn complex statistical patterns from data. Under this hypothesis, systems can infer correlational or causal relationships in the data. AI may learn and predict undesirable behavior from individual's behavioral patterns, trajectories, and proxies. By undesirable, we mean behaviors that individuals might not want to disclose. For example, an AI system trained on social media activity can infer depression and financial distress, which could help designing personalized clinical interventions, but could equally be exploited to discriminate, manipulate, or surveil individuals without their knowledge or consent. AI developers could address these issues by red-teaming latent inference harms or identifying inferences through interpretability techniques.

A.4 Manipulation & Persuasion

There are two scopes of manipulation and persuasion in AI systems: when AI is used as a tool to influence, and when AI is the agent of influence. In the first case, concerns have been raised about the use of AI for manipulation, for example, through the generation of synthetic media, such as deepfakes videos, cloned voices, and synthetic images generated by LLMs (Benkler et al., 2018; Chesney

and Citron, 2019); and disinformation (Goldstein et al., 2023), including for profit (Zuboff, 2019). In IG2, a system that can model individual behavior could naturally scale up these influences.

When AI is the agent of influence, including without the designer's intent, this could be a significant threat to human autonomy (Carroll et al., 2023). In IG2, beyond autonomy, when the AI "knows each individual's behavior", it could strategically exploit that to influence. Current frontier LLMs have already shown deception capabilities, such as alignment faking (Park et al., 2024b). Although there is a debate about whether AI has beliefs or not, if proven to be true, then an IG2 AI system could influence humans to achieve its goals.

Although regulatory and policy interventions could address the use of AI for manipulation, research is still ongoing on how to effectively identify and mitigate AI manipulations in IG1. These research results could be adapted to IG2.

A.5 Privacy Concerns

Privacy and confidentiality have been major concerns in AI development and use in IG1. Research has shown that AI trained on de-identified or incomplete data could re-identify data subjects (Rocher et al., 2019; Lermen et al., 2026). Although it is worth discussing how data collection and management could affect individual privacy and confidentiality in IG2, we focus on how the model inputs and outcomes challenge privacy in IG2. By design (see Section 3.2) IG2 does not require identifiable data for personalization; however, model inputs, such as tasks and behavioral outcomes, could pose privacy risks. For instance, individuals might have unique behavior that can re-identify them. In music recommender systems, these could be highly distinctive listening patterns. Similarly, in learning and education contexts, problem-solving paths, error patterns, or pacing can form distinctive signatures that both enable personalization and increase the risk of re-identification. Researchers should therefore explore how existing privacy preserving techniques in IG1 (e.g., differential privacy and federated learning) could inform IG2 and their failure modes, and develop new solutions.

A.6 Bias & Discrimination

AI bias is systematic errors in AI that lead to unfair outcomes (Mehrabi et al., 2021). These could amplify and perpetuate social inequities. In IG1, these issues are often attributed to AI training data, model design, and user interactions with AI (Mehrabi et al., 2021). Fairness in AI attempts to address these issues through technical interventions often categorized into individual and group fairness techniques. Individual fairness techniques aim to ensure that similar individuals have similar AI outcomes (Dwork et al., 2012), while group fairness techniques focus on similar outcomes for different groups (Verma and Rubin, 2018).

IG2 challenges the assumption that similar individuals should have similar outcomes. In IG2, individuals can be uniquely characterized by behavioral patterns and trajectories (see Section 3.3) that lead individuals to perform similar tasks differently. Thus, assigning similar outcomes to similar individuals can cause harm. Even when behaviors are similar between individuals for similar tasks, demographic, sociocultural proxies, individual trajectories, and temporality introduce variation in how those behaviors are expressed and interpreted. Although these granularities are technically feasible to define, for instance, defining similarity conditioning on individual characteristics, another challenge arises from aligning model outcomes with user preferences considering that these preferences are volatile and potentially unknown.

As the IG2 framework focuses on individual level modeling, which contrasts with group or population modeling in IG1, one might consider "group bias", or "group harm" (e.g., representational and social) in general (Shelby et al., 2023), less relevant in IG2. In IG1, groups are typically defined from protected classes, such as race, ethnicity, and gender (Fang et al., 2024; Ranjan et al., 2024). In IG2, these characteristics are implicitly encoded in individual proxies (see Section 3.2). Therefore, while in theory an AI system in IG2 focuses on individual level behavioral patterns, the model can infer population level patterns that could result in unfair outcomes. For example, if one culture is dominant in the proxies, the model might implicitly associate outcomes of non dominant cultures to the dominant one.

We also note a shift in group definitions from IG1 to IG2. The groups are typically relative to the input data of the models. For simplicity, we consider only tasks. As defined previously, the IG1 group fairness paradigm aims to ensure that different groups (tasks) have similar outcomes (behaviors). In other words, different tasks should be resolved with similar behaviors for a given individual. This assumption might be true for some

individuals, but not for others. For example, some individuals might use the same logic to solve different tasks, while others adapt their reasoning strategy to the nature of each task.

A.7 Appropriateness & Governance

In this section, we question whether AI systems should be developed under an IG2 paradigm in the first place. At minimum, this requires an ethics assessment to evaluate the risks and benefits of such systems. In the previous sections, we have discussed some of these ethics elements. Here, we argue for the need for appropriateness and governance mechanisms to develop and use these systems. These include technical, regulatory, and policy interventions to define when these systems are appropriate, for example, the types of permissible applications. Moreover, governance mechanisms should define who should develop such a system, for whom, and when? Who should access models predictions, individuals' data such as behavioral trajectories, which as discussed are sensitive.

B Additional Case Studies

B.1 Learning and Education

Generative AI systems are increasingly used for tutoring, feedback generation, and personalized learning support (Chu et al., 2025; Sonkar et al., 2023; Maurya et al., 2025; Li et al., 2024). Existing approaches personalize learning through trait-, constraint-, and pedagogy-based learner models (Liu et al., 2024b; Wang et al., 2025; Liu et al., 2025), but still rely on assumptions about which learner characteristics matter. Research across cognitive science, anthropology, and education has long shown that learning is deeply shaped by culture and social context (Vygotsky, 1978; Cole, 1998; Rogoff, 2003), while recent empirical work suggests that conceptual understanding ultimately manifests at the individual level (Saha et al., 2025b).

These limitations become especially visible in neurodiverse educational settings, where children with autism spectrum disorder often exhibit highly individualized learning and communication patterns (Tsapanou et al., 2025). IG2 therefore provides a principled framework for moving beyond static learner categories toward genuinely individualized educational systems. An example is Pandey et al. (2025)'s Culturally Yours, which adapts explanations based on what a particular user may not understand from online text. In this sense, IG2 advances the long-standing goal of educational AI and intelligent tutoring systems of providing instruction tailored to each learner's cognitive, cultural, and social context (Anderson et al., 1985; Woolf, 2010; Koedinger et al., 2013).

B.2 Scientific Discovery

AI systems are increasingly being used for literature search, hypothesis generation, code synthesis, experiment planning, and scientific writing (Wang et al., 2023; Baek et al., 2025; Lu et al., 2024; Su et al., 2025), reflecting successful IG1-style AI-for-Science systems. Recent work in the life sciences makes this especially clear: AI co-scientist systems have already generated experimentally supported hypotheses for drug repurposing in acute myeloid leukemia and for target discovery in liver fibrosis (Gottweis et al., 2025). But this still demonstrates generic scientific usefulness, not yet acquainted collaboration. Under IG2, the question is not only whether a system can generate a good hypothesis, but whether it can do so in a way that matches how a particular lab or scientist reasons, weighs evidence, and updates beliefs over time, thereby behaving more like an acquainted collaborator. Scientific discovery is shaped not only by correctness, but also by prior knowledge, intuitions, methodological preferences, and tacit forms of reasoning (Kuhn and Hacking, 1970; Polanyi, 2009). Just as IG2 reproduces how Maya answers "What is the capital of the UAE?", in order to identify and rectify specific misconceptions, an IG2-style scientific assistant must adapt its reasoning to the collaborator it assists. For instance, two research groups working on the same problem may prioritize different objectives, such as empirical performance versus interpretability. An IG1-style system may recommend the same workflow to both groups, whereas IG2 requires adapting to their distinct behavioral and epistemic tendencies, behaving more like an acquainted co-scientist than merely a capable knowledge system.

C Keyword Taxonomy for ACL Anthology Survey

Table 1 summarizes the keyword taxonomy used to categorize ACL Anthology papers. We use a multi-label taxonomy with two objective-level categories, *functional* and *behavioral*, and three computational-paradigm cate-

Category	Representative keywords / patterns
Functional	machine translation, MT, SMT, NMT, phrase-based translation, word alignment, BLEU, question answering, reading comprehension, information retrieval, document retrieval, summarization, information extraction, named entity recognition, relation extraction, event extraction, OpenIE, slot filling, dialogue system, task-oriented dialogue, natural language understanding, natural language generation, sentiment analysis, opinion mining, stance detection, semantic parsing, semantic role labeling, word sense disambiguation, coreference resolution, textual entailment, paraphrase, code generation, program synthesis, tool use, retrieval-augmented generation
Behavioral	pragmatics, implicature, speech act, discourse, coherence, commonsense reasoning, social intelligence, social cognition, social norms, theory of mind, figurative language, metaphor, irony, sarcasm, argumentation, persuasion, bias, fairness, stereotypes, gender bias, racial bias, toxicity, hate speech, misinformation, disinformation, NLP ethics, responsible AI, harmful content, alignment, AI safety, red-teaming, hallucination, factuality, truthfulness, sycophancy, multilinguality, cross-linguality, low-resource languages, code-switching, dialects, indigenous languages, cultural knowledge, cultural bias, world knowledge, political discourse, social media analysis, user modeling, user profiles, personalization, preferences, personas, demographics
Symbolic/rule-based	syntax, syntactic parsing, parser, context-free grammar, CFG, phrase structure, constituents, dependency parsing, shift-reduce parsing, chart parsing, Earley parsing, CYK, categorial grammar, TAG, HPSG, LFG, CCG, formal semantics, lambda calculus, Montague grammar, logical form, model-theoretic semantics, first-order logic, predicate logic, unification-based grammars, feature structures, lexicon, automata, finite-state methods, regular languages, transducers, two-level morphology, generative grammar, transformational grammar, Chomsky, rule-based methods, hand-crafted rules, grammar-based methods, template-based methods, pattern-based methods, knowledge-based systems, expert systems
Symbolic/data-driven	n-grams, non-neural language models, hidden Markov models, HMMs, maximum entropy, MaxEnt, support vector machines, SVMs, conditional random fields, CRFs, naive Bayes, expectation maximization, probabilistic models, generative models, log-linear models, latent Dirichlet allocation, topic models, bootstrapping, corpus-based methods, feature engineering, feature-based models, statistical models, statistical parsing, probabilistic parsing, statistical machine translation, IBM models, phrase-based translation, word alignment, treebanks, Penn Treebank, POS tagging, chunking, morphology, lemmatization, stemming, tokenization, segmentation
Sub-symbolic/data-driven	word embeddings, word2vec, GloVe, fastText, distributional semantics, word vectors, distributed representations, neural language models, RNNs, LSTMs, GRUs, CNNs, encoder-decoder models, sequence-to-sequence models, attention, self-attention, transformers, multi-head attention, pointer networks, memory networks, graph neural networks, graph convolutional networks, BERT, ELMo, GPT, XLNet, RoBERTa, ALBERT, T5, PaLM, LLaMA, Mistral, Gemini, mBERT, XLM-R, BLOOM, mT5, pretraining, fine-tuning, transfer learning, contextual representations, masked language modeling, large language models, foundation models, instruction tuning, RLHF, in-context learning, few-shot prompting, zero-shot prompting, chain-of-thought, prompt engineering, emergent abilities, scaling laws, neural machine translation, back-translation, beam search, subword modeling, BPE, SentencePiece, multimodal models, vision-language models, image captioning, VQA, CLIP, DALL-E, stable diffusion

Table 1: Summary of the keyword taxonomy used for the ACL Anthology survey. Categories are multi-label: a paper may be assigned to both an objective category and a computational-paradigm category.

gories, *symbolic/rule-based*, *symbolic/data-driven*, and *sub-symbolic/data-driven*. Keywords were matched against paper titles and abstracts using case-insensitive regular-expression matching. For readability, the table reports simplified keyword forms; the implementation includes spelling, hyphenation, and abbreviation variants.