Cognitive Machine Learning for Patient-First Modeling in Clinical Research

Anonymous Author(s)
Affiliation
Address
email

Abstract

Clinical trials remain the cornerstone of evidence-based medicine. Yet their prevailing methods often reduce patients to statistical data points, overlooking cognition-driven factors such as consent comprehension, assessment fatigue, and telescoping in adverse-event (AE) reporting. We propose a new approach that integrates cognitive science and large language models (LLMs) to model patient comprehension, recall, preferences, and incentives. Building on behavioral foundation models as starting priors, we introduce a cognitive ML model for clinical research: a thin, patient-first layer that adapts foundation models to trial workflows via clinical cover stories such as consent with brief teach-backs, AE narratives with temporal anchors, preference trade-offs under framing, and bias-aware disclosure prompts. The layer overlays existing PDFs, ePRO apps, and AE logs, adding guardrails such as calibration thresholds, clinician deferral, and auditability, rather than replacing infrastructure. We outline a roadmap from lightweight cognitive overlays to human-in-the-loop integration, and ultimately, cognitive-integrated trials with governance and regulatory alignment. An abridged AE-reporting case study shows increased AE yield and improved timing fidelity while enforcing calibration and subgroup-parity gates. The next generation of clinical trials must be not only statistically rigorous but cognitively grounded and inclusive.

1 Introduction

Clinical research has relied on statistical paradigms to evaluate treatment safety and efficacy [1, 2]. This paradigm has driven decades of medical progress, but statistical analysis alone is not sufficient to capture the complexity of human responses in clinical decision making [3–8]. The design and analysis of clinical trials and clinical trial data must go beyond statistical inference and incorporate cognitive models that reflect human-like reasoning and behavior [3, 5, 9–12].

By integrating insights from cognitive science, machine learning, and reinforcement learning [5, 10, 12–21], we propose a new approach for collecting, analyzing, and interpreting clinical trial data. This approach will shift the focus from treating patients as sources of decontextualized signals to acknowledging them as cognitive agents whose preferences, incentives, and behaviors are essential to the therapeutic decision process [3–5, 12]. Our proposal is urgent given recent trends, risks, and opportunities [22–29].

Trends. Trends in AI have fueled anxieties about the replacement of human expertise with automated systems [23, 24, 26, 28]. Such anxieties underscore the need for approaches that bring humans back in the loop, ensuring that AI augments rather than supplants clinical reasoning [6, 7, 30–32].

Risks. Emphasizing "automation" risks producing systems that operate as opaque statistical machines with limited interpretability, eroding clinician and patient trust [6, 7, 26, 28, 33]. We argue for a

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

paradigm centered on understanding human responses, not merely automating them [3–5, 12]. Without this emphasis on interpretability and inclusion of human cognition, confidence in computational methods for clinical trials will remain fragile [6, 7, 26, 28, 33].

Opportunities. Recent advances in large language models (LLMs) are serendipitous [14–21, 34]. Clinical trials have long struggled with two challenges: the difficulty with *extracting* rich, authentic information from patients, and the tendency toward narrow, biomarker-focused *analyses* that overlook broader dimensions of the patient journey [2, 35–37]. LLMs offer a way to address both challenges [15, 17–21, 38]. First, on *extraction*, LLMs open a new mode of human–machine interaction that enables the unobtrusive collection of subjective responses from patients [17, 21, 23, 24, 29, 36, 37]. Unlike conventional clinical assessments that often require direct intervention or structured questionnaires, conversational interfaces mediated by LLMs allow patients to share experiences in a natural, less burdensome manner. Importantly, such interactions reduce the hesitancy or sentimentality often associated with self-reporting, thereby enriching the range and authenticity of captured responses, a possibility not previously realized in the clinical trial domain [22–24, 36]. Second, on *analysis*, LLMs extend beyond response extraction to provide a means of articulating and formalizing subjective reports in analyzable formats [18, 37–40]. This capability is not a mere technical improvement but a normative shift in how data in clinical trials can be conceptualized [5–7, 12]. For decades, trial data collection has centered on biomarkers and structured electronic patient-reported outcomes (ePROs) [2, 41]. These approaches offer a narrow lens on the patient journey [35–37, 42]. By contrast, interactions with LLMs create an avenue for accessing richer dimensions of trial participation, including patient preferences, cognitive states, incentives, and reactions to treatment responses [5, 12, 17, 21, 23, 24, 29]. This epistemic shift—moving from statistical aggregation of biomarkers to cognitive engagement with patient responses—constitutes a critical next step for ML in clinical research [3, 4, 6, 7, 12].

Recent work introduced Centaur, a *behavioral* foundation model fine-tuned on large-scale human behavior that predicts trial-by-trial choices and reaction times [43]. Centaur generalizes across held-out tasks including when the cover story is changed [43]. Notably, fine-tuning increases alignment between the model's internal representations and human neural activity, suggesting that such models capture cognitively meaningful structure rather than superficial correlations [43, 44]. This indicates that LLMs can capture aspects of human cognition, not merely approximate them [44, 45].

Building on this line of work, we introduce a cognitive ML model for clinical research [24, 29]. The cognitive ML model is a thin, patient-first layer that adapts foundation models to clinical-trial workflows via clinical cover stories [29]. Examples of clinical cover stories include consent explanations paired with brief comprehension checks, AE narratives with temporal anchors to mitigate telescoping, preference/utility trade-offs under framing, and bias-aware disclosure prompts [46, 47]. The layer is intentionally lightweight; rather than replace clinical trial infrastructure, the layer overlays existing PDFs, ePRO apps, and AE logs, while adding guardrails such as calibration thresholds, clinician deferrals, and auditability [48–50].

Collectively, the highlighted opportunities suggest that integrating LLMs into clinical trials is not simply about technological adoption but about addressing enduring limitations in trial design [2, 6, 35, 37]. In the following section, we turn to the challenges and blind spots that define this problem space, motivating why an integrative approach across clinical research, ML, and cognitive science is urgently needed [2, 4, 30–32].

2 Motivation

Many challenges undermine the reliability, inclusivity, and interpretability of clinical trial outcomes [2, 22, 25–28, 35, 37, 42]. These challenges are more than technical; they are cognitive and human in nature [3–5, 12, 13]. Clinical research, ML, and cognitive science have historically evolved in isolation, leaving overlooked challenges, unaddressed blind spots, and misaligned incentives that call for deliberate integration [30–32, 35–37, 51–57].

2.1 Overlooked Challenges

One set of issues emerges from the lived experience of trial patients [23, 35–37] such as consent form complexity [58–60]. Patients are routinely confronted with dense, jargon-heavy documents that induce information overload [58–62]. Such overload undermines true informed consent, as comprehension is compromised even before trial participation begins [59–61, 63]. Cognitive ML offers a path forward. By modeling comprehension and attention, we can refine consent forms in ways that respect cognitive limitations while improving patient understanding [10, 12, 13, 58].

Similarly, survey fatigue remains a persistent barrier in the use of ePROs and symptom-tracking apps [35–37, 42]. Repeated surveys can induce habituation, where patients disengage or provide increasingly superficial responses over time [35, 36, 42, 64–66]. This habituation compromises the reliability of the data and erodes patient trust in the trial process [35, 36, 42].

Another challenge is the telescoping effect in self-reporting where patients misjudge when AEs occurred [4, 67–69]. This temporal misalignment leads to inaccurate event logs that confound downstream statistical analyses [2, 4, 67–69]. Cognitive ML systems resolve this issue by modeling recall processes and response latencies and using temporal anchors to improve AE log fidelity. Equally problematic are blind spots across the three disciplines that ought to inform each other.

2.2 Blind Spots

In clinical research, AE reporting is a weak link [41, 70–76]. Underreporting remains common, particularly for mild or socially sensitive events [71, 72, 76]. For instance, gastrointestinal or sexual side effects are systematically underreported despite being clinically relevant [71, 76]. Current trial processes rely on participants' initiatives to report, without accounting for cognitive or social barriers [4, 35, 71, 72]. Interactive cognitive ML systems could lower these barriers by offering nonjudgmental conversational spaces, thereby reducing AE underreporting [17, 21, 23, 24, 29, 36].

In ML, there is a persistent neglect of cognitive paradigms [30–32]. Models typically treat human responses as noisy signals rather than as outputs shaped by cognitive paradigms and psychological processes [3–5, 12]. By simulating human comprehension and cognitive constraints, we argue for a methodological shift toward patient-first modeling [5, 10, 12, 30–32]. Empirically, the urgency is underscored by dropout rates. Across trial phases, patient dropout remains a major obstacle, with estimates exceeding 30% in some domains [2, 35, 75]. Models that explicitly incorporate cognitive processes may reduce dropout by fostering trust, comprehension, and engagement [6, 30–32, 35, 75].

In cognitive science, another blind spot exists: "people know more than they tell" [4, 5, 12, 36]. Traditional paradigms often underestimate the tacit, implicit, or socially inhibited knowledge patients hold [4, 5, 36]. ML, particularly interactive LLMs, can draw out this latent information through conversational engagement [17, 21, 23, 24, 29, 36]. By scaffolding naturalistic dialogue, ML can surface experiences and concerns that structured surveys would otherwise miss [23, 24, 29, 36, 77].

2.3 Misaligned Incentives

Clinical trials are shaped by misaligned incentives that must be addressed for any integrative approach to succeed [22, 25–28, 51–57]. Patient diversity remains a pressing concern [51–57, 78, 79]. Trials often fail to adequately recruit or retain patients from historically marginalized groups, resulting in biased evidence bases [51–57, 78, 79]. For example, during the COVID-19 pandemic, Black communities expressed distrust toward clinical trials, citing a lack of representation in study design and outcomes [51–53, 55, 56]. Without explicitly modeling such concerns, trial evidence risks perpetuating inequities [22, 25–28, 51–57].

Moreover, emerging insights from behavioral economics and neuromarketing point to a related tension: human cognition is not merely rational but deeply shaped by incentives, moods, and contextual framing [9, 11, 80–84]. For example, marketing studies show that subtle manipulations of framing or context can alter consumer decision-making in predictable ways [9, 80–84], yet clinical trial design often neglects such cognitive-behavioral dynamics, assuming rational participation [5, 9, 82]. An integrative approach must recognize these misaligned incentives, not as confounds to be eliminated, but as realities to be modeled and ethically addressed [22, 25–28, 82].

Together, these overlooked challenges, blind spots, and misaligned incentives motivate a principled roadmap [25, 85]. Therefore, we commit to a patient-first cognitive ML model for clinical research that targets behavioral competence on trial-relevant tasks, maintains calibrated reliability and subgroup parity, and remains lightweight and auditable [46]. In the appendices, we translate these commitments into a phased roadmap and a concrete AE-reporting case study.

3 Roadmap and Case Study for Cognitive ML in Clinical Trials

Imagine a clinical trial where, on Day 1, participants meet an LLM that clarifies consent in plain language; by Week 4, the same system adaptively elicits symptoms, flags likely underreported events, and hands clinicians an auditable summary grounded in cognitive theory, not just statistics. Our ambitious vision is a direct response to long-standing extraction and analysis gaps, blind spots across disciplines, and misaligned incentives that erode trust.

Anticipating objections—"LLMs are brittle," "regulators won't accept this," "bias will creep in"—our roadmap sequences work where benefits are immediate and verification is possible with explicit safeguards: clinician adjudication and calibration against gold standards, bias and drift audits, and privacy-preserving data governance. We highlight two preconditions for progress that require collaboration beyond modeling: IRB-approved guardrails [86] and data stewardship aligned with existing trial oversight; and interoperability with trial infrastructure such as EDC/CDISC pipelines. With these foundations, the phases that follow move from feasibility pilots to cognitive ML integration and, ultimately, to a transformed trial ecosystem in which human cognition is modeled, measured, and ethically incorporated.

Our roadmap advances from near-term feasibility to systemic transformation while keeping interventions lightweight, auditable, and centered on the cognitive ML model for clinical research introduced in Sec. 1. Across all phases we adopt three minimal commitments per workstream: behavioral (capability and calibration targets defined a priori), processing (one decision-relevant mechanism check, e.g., representational or causal), and development (safe adaptation with audit trails). Deployment at each step is gated by a preregistered Behavioral Capability Audit (BCA) and supported by clinician deferral, subgroup-parity monitoring, and interoperability with existing EDC/CDISC pipelines.

To reify the roadmap, we preview an exemplar case study on AE reporting that evaluates a thin, patient-first cognitive ML model for clinical research layered over the existing ePRO workflow. In a parallel-arm randomized design, the intervention arm augments standard AE questionnaires with a lightweight conversational overlay that uses clinical cover stories to normalize sensitive categories and temporal anchors (e.g., mealtimes, wake/sleep) to reduce telescoping; outputs are written back to EDC/CDISC fields with calibrated confidence and automatic clinician deferral for low-confidence or out-of-distribution cases.

Co-primary endpoints are incremental AE yield (clinician-adjudicated) and temporal accuracy (absolute onset-time error vs. adjudication). Secondary endpoints include patient comfort and willingness-to-disclose, probabilistic calibration (ECE/Brier with selective prediction), clinician workload, and subgroup parity for yield and timing across age/sex/race. Deployment is gated by a pre-registered BCA specifying thresholds for capability, calibration, and parity, with audit logs, drift monitoring, and rollback plans. Integration is deliberately minimal—no endpoint changes, no infrastructure replacement—prioritizing safety, privacy, and auditability. Full protocol details, including prompts and cover-story templates, anchoring taxonomies, statistical analysis plan, and BCA thresholds, are provided in the Appendices.

4 Conclusion

We have argued that statistical paradigms—although foundational to clinical trial design—are not sufficient to capture the full complexity of human responses in therapeutic evaluation [2–7]. By integrating cognitive science with advances in ML, particularly LLMs, we propose a new approach for clinical research that treats human behavior, preferences, and cognition as signals to be modeled rather than noise to be suppressed [3–5, 12, 14–21, 38]. Our roadmap outlines a progression from cognitive-aware data collection to HITL ML and ultimately to systemic transformations in how trials are conducted and regulated [2, 30–32, 35, 37]. The proposed case study designs on survey fatigue and AE reporting illustrate the feasibility and impact of this integration, offering concrete pathways to address dropout, underreporting, and participant disengagement [35–37, 41, 42, 70–75]. Ultimately, our agenda induces epistemic and ethical shifts toward clinical trials that are not only statistically rigorous but also cognitively grounded, patient-centered, and more trustworthy in their capacity to guide medical decision-making [6, 7, 22, 25–28, 37, 51–57, 87].

References

[1] Erik von Elm, Douglas G. Altman, Matthias Egger, Stuart J. Pocock, Peter C. Gøtzsche, and Jan P. Vandenbroucke. The STROBE statement: Guidelines for reporting observational studies. *PLOS Medicine*, 4(10):e296, 2007. doi: 10.1371/journal.pmed.0040296.

[2] David Moher, Sally Hopewell, Kenneth F. Schulz, Victor Montori, Peter C. Gøtzsche, P. J. Devereaux, Diana Elbourne, Matthias Egger, and Douglas G. Altman. Consort 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *PLOS Medicine*, 7(3):e1000251, 2010. doi: 10.1371/journal.pmed.1000251.

[3] Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131, 1974. doi: 10.1126/science.185.4157.1124.

[4] Roger Tourangeau, Lance J. Rips, and Kenneth Rasinski. *The Psychology of Survey Response*. Cambridge University Press, Cambridge, 2000. ISBN 9780521576291.

[5] Gerd Gigerenzer and Wolfgang Gaissmaier. Heuristic decision making. *Annual Review of Psychology*, 62(1):451–482, 2011. doi: 10.1146/annurev-psych-120709-145346.

[6] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. Working Paper, 2017. URL `https://research.google/pubs/towards-a-rigorous-science-of-interpretable-machine-learning/`.

[7] Cynthia Rudin. Stop explaining black box machine learning models for high-stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019. doi: 10.1038/s42256-019-0048-x. URL `https://www.nature.com/articles/s42256-019-0048-x`.

[8] Stephen Bates, Michael I Jordan, Michael Sklar, and Jake A Soloff. Incentive-theoretic bayesian inference for collaborative science. *arXiv preprint arXiv:2307.03748*, 2023.

[9] Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–291, 1979. doi: 10.2307/1914185.

[10] John R. Anderson and Christian Lebiere. *The Atomic Components of Thought*. Lawrence Erlbaum Associates, Mahwah, NJ, 1998. ISBN 0805828176.

[11] Daniel Kahneman. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, New York, 2011. ISBN 9780374275631.

[12] Falk Lieder and Thomas L. Griffiths. Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, 43:e1, 2020. doi: 10.1017/S0140525X1900061X.

[13] John Sweller. Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2):257–285, 1988. doi: 10.1016/0364-0213(88)90023-7.

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL `https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf`.

[15] Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL `https://proceedings.neurips.cc/paper_files/paper/2017/file/d5e2c0adad503c91f91df240d0cd4e49-Paper.pdf`.

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL `https://aclanthology.org/N19-1423/`.

[17] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott

Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[18] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F. Christiano. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://papers.nips.cc/paper/2020/file/1f89885d556929e98d3ef9b86448f951-Paper.pdf.

[19] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.

[20] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, and Jared Kaplan. Constitutional ai: Harmlessness from ai feedback. Anthropic Research Report, 2022. URL https://www-cdn.anthropic.com/7512771452629584566b6303311496c262da1006/Anthropic_ConstitutionalAI_v2.pdf.

[21] OpenAI. Gpt-4 technical report. Technical report, OpenAI, 2023. URL https://cdn.openai.com/papers/gpt-4.pdf.

[22] Alvin Rajkomar, Moritz Hardt, Michael D. Howell, Greg Corrado, and Michael H. Chin. Ensuring fairness in machine learning to advance health equity. *Annals of Internal Medicine*, 169(12):866–872, 2018. doi: 10.7326/M18-1990.

[23] Eric Topol. *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again*. Basic Books, New York, 2019.

[24] Andre Esteva, Alexandre Robicquet, Bharath Ramsundar, Volodymyr Kuleshov, Mark DePristo, Katherine Chou, Claire Cui, Greg Corrado, Sebastian Thrun, and Jeff Dean. A guide to deep learning in healthcare. *Nature Medicine*, 25(1):24–29, 2019. doi: 10.1038/s41591-018-0316-z.

[25] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019. doi: 10.1126/science.aax2342.

[26] Robert Challen, Josh Denny, Matt Pitt, Luke Gompels, Tim Edwards, and Krasimira Tsaneva-Atanasova. Artificial intelligence, bias and clinical safety. *BMJ Quality & Safety*, 28(3):231–237, 2019. doi: 10.1136/bmjqs-2018-008370.

[27] Harini Suresh and John Guttag. A framework for understanding unintended consequences of machine learning. *Harvard Data Science Review*, 2021. URL https://hdsr.mitpress.mit.edu/pub/a0j0k6i8. Issue 2.2.

[28] Marzyeh Ghassemi, Luke Oakden-Rayner, and Andrew L. Beam. The false hope of current approaches to explainable AI in health care. *The Lancet Digital Health*, 3(11):e745–e750, 2021. doi: 10.1016/S2589-7500(21)00052-9.

[29] Pranav Rajpurkar, Emily Chen, Oishi Banerjee, and Eric J. Topol. Ai in health and medicine. *Nature Biomedical Engineering*, 6:134–136, 2022.

[30] Burr Settles. *Active Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool, 2012. doi: 10.2200/S00429ED1V01Y201207AIM018. URL https://dx.doi.org/10.2200/S00429ED1V01Y201207AIM018.

[31] Saleema Amershi, Maya Cakmak, W. Bradley Knox, and Todd Kulesza. Power to the people: The role of humans in interactive machine learning. *AI Magazine*, 35(4):105–120, 2014. doi: 10.1609/aimag.v35i4.2513. URL https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/2513.

[32] Ece Kamar. Directions in hybrid intelligence: Complementing ai systems with human intelligence. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16)*, pages 4070–4073. AAAI Press, 2016. URL https://www.microsoft.com/en-us/research/publication/directions-hybrid-intelligence-complementing-ai-systems-human-intelligence/.

[33] Zachary C. Lipton. The mythos of model interpretability. In *Proceedings of the 2016 ICML Workshop on Human Interpretability in Machine Learning (WHI 2016)*, New York, NY, 2016. ICML Workshop Paper.

[34] Hugo Touvron et al. Llama: Open and efficient foundation language models. Technical report, Meta AI, 2023. URL https://ai.facebook.com/blog/large-language-model-llama/. Technical Report.

[35] Gunther Eysenbach. The law of attrition. *Journal of Medical Internet Research*, 7(1):e11, 2005. doi: 10.2196/jmir.7.1.e11. URL https://www.jmir.org/2005/1/e11.

[36] Saul Shiffman, Arthur A. Stone, and Michael R. Hufford. Ecological momentary assessment. *Annual Review of Clinical Psychology*, 4:1–32, 2008. doi: 10.1146/annurev.clinpsy.3.022806.091415.

[37] U.S. Food and Drug Administration. Guidance for industry: Patient-reported outcome (pro) measures: Use in medical product development to support labeling claims. Guidance document, 2009. URL https://www.fda.gov/media/77832/download.

[38] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Paper.pdf.

[39] Abeed Sarker and Graciela Gonzalez. Portable automatic text classification for adverse drug reaction detection via multi-corpus training. *Journal of Biomedical Informatics*, 53:196–207, 2015. doi: 10.1016/j.jbi.2014.11.002.

[40] Yanshan Wang, Lei Wang, Majid Rastegar-Mojarad, Sungrim Moon, Feichen Shen, Naveed Afzal, Siqin Liu, Dongbo Zeng, Saeed Mehrabi, Sunghwan Sohn, and Hongfang Liu. Clinical information extraction applications: A literature review. *Journal of Biomedical Informatics*, 77: 34–49, 2018. doi: 10.1016/j.jbi.2017.11.011.

[41] U.S. Food and Drug Administration. Safety reporting requirements for inds and ba/be studies: Guidance for industry. Technical report, FDA, 2012. URL https://www.fda.gov/regulatory-information/search-fda-guidance-documents/safety-reporting-requirements-inds-and-babe-studies.

[42] Siv Rolstad, Jonas Adler, and Astrid Rydén. Response burden and questionnaire length: Is shorter better? a review and meta-analysis. *Value in Health*, 14(8):1101–1108, 2011. doi: 10.1016/j.jval.2011.06.003.

[43] Marcel Binz, Eric Schulz, Charlotte Caucheteux, Lucas Huber, Jean-Remi King, et al. Centaur: Foundation models of cognitive behaviors. *Nature*, 2025. doi: 10.1038/s41586-025-09215-4. URL https://www.nature.com/articles/s41586-025-09215-4.

[44] Martin Schrimpf, Idan Blank, Giedre Tuckute, Carmen Kauf, Eghbal A. Hosseini, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118, 2021. doi: 10.1073/pnas.2105646118. URL https://www.pnas.org/doi/10.1073/pnas.2105646118.

[45] Charlotte Caucheteux, Alexandre Gramfort, and Jean-Remi King. Brains and algorithms partially converge in natural language processing. *Communications Biology*, 6(1):16, 2023. doi: 10.1038/s42003-022-03036-1. URL https://www.nature.com/articles/s42003-022-03036-1.

[46] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, volume 29 of *NeurIPS*, pages 3323–3331, 2016. URL https://proceedings.neurips.cc/paper/2016/hash/9d2682367c3935defcb1f9e247a97c0d-Abstract.html.

[47] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019. doi: 10.1126/science.aax2342. URL https://www.science.org/doi/10.1126/science.aax2342.

[48] CDISC. Study data tabulation model implementation guide (sdtmig) version 3.2, 2013. URL https://www.cdisc.org/standards/foundational/sdtmig.

[49] Ai risk management framework (ai rmf 1.0). Technical report, National Institute of Standards and Technology, 2023. URL https://www.nist.gov/itl/ai-risk-management-framework.

[50] Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In *Advances in Neural Information Processing Systems*, volume 33 of *NeurIPS*, pages 13347–13358, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/4cf1c6c2e0d2f27f8b5727a9d6d0435a-Abstract.html.

[51] David Wendler, Raynard Kington, Jennifer Madans, Gretchen Van Wye, Heiko Christ-Schmidt, Laura A. Pratt, Otis W. Brawley, Cary P. Gross, and Ezekiel Emanuel. Are racial and ethnic minorities less willing to participate in health research? *PLOS Medicine*, 3(2):e19, 2006. doi: 10.1371/journal.pmed.0030019.

[52] Darcell P. Scharff, Karl J. Mathews, Prince Jackson, Jaime Hoffsuemmer, Esperanza Martin, and Deidre Edwards. More than tuskegee: Understanding mistrust about research participation. *Journal of Health Care for the Poor and Underserved*, 21(3):879–897, 2010. doi: 10.1353/hpu.0.0323.

[53] Sheba George, Nelida Duran, and Keith Norris. A systematic review of barriers and facilitators to minority research participation among african americans, latinos, asian americans, and pacific islanders. *Journal of Health Care for the Poor and Underserved*, 25(1):36–55, 2014. doi: 10.1353/hpu.2014.0017.

[54] Sam S. Oh, Joshua Galanter, Neeta Thakur, Maria Pino-Yanes, Nicolas E. Barcelo, Marquitta J. White, Danielle M. de Bruin, Ruth M. Greenblatt, Kirsten Bibbins-Domingo, Alan H. B. Wu, Luisa N. Borrell, Chris Gunter, Neil R. Powe, and Esteban G. Burchard. Diversity in clinical and biomedical research: A promise yet to be fulfilled. *PLOS Medicine*, 12(12):e1001918, 2015. doi: 10.1371/journal.pmed.1001918.

[55] Narjust Duma, Jaime Vera Aguilera, Jermaine Paludo, Christine Haddox, Maria Gonzalez Velez, Yanjun Wang, Konstantinos Leventakos, Jane M. Hubbard, Andrea S. Mansfield, Alex A. Adjei, and et al. Representation of minorities and women in oncology clinical trials: Review of the past 14 years. *Journal of Oncology Practice*, 14(1):e1–e10, 2018. doi: 10.1200/JOP.2017.026674.

[56] Fangjun Chen and et al. Representation of older adults in clinical trials supporting fda drug approvals, 2011–2014. *JAMA Internal Medicine*, 179(8):in press, 2019. doi: 10.1001/jamainternmed.2019.1500.

[57] U.S. Food and Drug Administration. Diversity plans to improve enrollment of participants from underrepresented racial and ethnic populations in clinical trials (draft guidance). Technical report, FDA, 2022. URL https://www.federalregister.gov/documents/2022/04/14/2022-07800/diversity-plans-to-improve-enrollment-of-participants-from-underrepresented-racial-and-ethni

[58] Fred Paas, Alexander Renkl, and John Sweller. Cognitive load theory and instructional design: Recent developments. *Educational Psychologist*, 38(1):1–4, 2003. doi: 10.1207/S15326985EP3801_1.

[59] J. David Flory and Ezekiel J. Emanuel. Interventions to improve research participants' understanding in informed consent for research: A systematic review. *JAMA*, 292(13):1593–1601, 2004. doi: 10.1001/jama.292.13.1593.

[60] Christine Grady. Enduring and emerging challenges of informed consent. *New England Journal of Medicine*, 372(9):855–862, 2015. doi: 10.1056/NEJMra1411250.

[61] Dean Schillinger, John Piette, Kevin Grumbach, Frances Wang, Courtney Wilson, Carol Daher, Kim Leong-Grotz, Cesar Castro, and Andrew B. Bindman. Closing the loop: Physician communication with diabetic patients who have low health literacy. *Archives of Internal Medicine*, 163(1):83–90, 2003. doi: 10.1001/archinte.163.1.83.

[62] J. N. Haun, N. R. Patel, D. D. French, R. R. Campbell, D. D. Bradham, and W. A. Lapcevic. Evidence for the use of the teach-back method: A systematic review. *Patient Education and Counseling*, 98(10):1295–1305, 2015.

[63] Ruth R. Faden and Tom L. Beauchamp. *A History and Theory of Informed Consent*. Oxford University Press, New York, 1986. ISBN 9780195036862.

[64] Jon A. Krosnick. Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, 5(3):213–236, 1991. doi: 10.1002/acp.2350050305.

[65] Mirta Galesic and Michael Bosnjak. Effects of questionnaire length on participation and indicators of response quality in a web survey. *Public Opinion Quarterly*, 73(2):349–360, 2009. doi: 10.1093/poq/nfp031.

[66] Don A. Dillman, Jolene D. Smyth, and Leah Melani Christian. *Internet, Phone, Mail, and Mixed-Mode Surveys: The Tailored Design Method*. Wiley, Hoboken, NJ, 4 edition, 2014. ISBN 9781118456149.

[67] John Neter and Joseph Waksberg. A study of response errors in expenditures data from household interviews. *Journal of the American Statistical Association*, 59(305):18–55, 1964. doi: 10.1080/01621459.1964.10480699.

[68] Seymour Sudman and Norman M. Bradburn. Effects of time and memory factors on response in surveys. *Journal of the American Statistical Association*, 68(344):805–815, 1973. doi: 10.1080/01621459.1973.10481428.

[69] Norman M. Bradburn, Lance J. Rips, and Steven K. Shevell. Answering autobiographical questions: The impact of memory and inference on surveys. *Science*, 236(4798):157–161, 1987. doi: 10.1126/science.3563494.

[70] Clinical safety data management: Definitions and standards for expedited reporting (e2a), 1994. URL `https://database.ich.org/sites/default/files/E2A_Guideline.pdf`.

[71] Lorna Hazell and Saad A. W. Shakir. Under-reporting of adverse drug reactions: A systematic review. *Drug Safety*, 29(5):385–396, 2006. doi: 10.2165/00002018-200629050-00003.

[72] Elena Lopez-Gonzalez, Maria Teresa Herdeiro, and Adolfo Figueiras. Determinants of under-reporting of adverse drug reactions: A systematic review. *Drug Safety*, 32(1):19–31, 2009. doi: 10.2165/00002018-200932010-00002.

[73] Su Golder, Gill Norman, and Yoon K. Loke. Systematic review on the prevalence, frequency and comparative value of adverse events data in social media. *British Journal of Clinical Pharmacology*, 80(4):878–888, 2015. doi: 10.1111/bcp.12659.

[74] CIOMS Working Group VIII. *Signal Detection in Pharmacovigilance*. Council for International Organizations of Medical Sciences (CIOMS), Geneva, 2016. URL `https://cioms.ch/publications/working-group-reports/signal-detection-in-pharmacovigilance/`.

[75] Ethan Basch, Allison M. Deal, Amylou C. Dueck, Howard I. Scher, Mark G. Kris, Clifford Hudis, and Deborah Schrag. Overall survival results of a trial assessing patient-reported outcomes for symptom monitoring during routine cancer treatment. *JAMA*, 318(2):197–198, 2017. doi: 10.1001/jama.2017.7156.

[76] I. Ralph Edwards and Jeffrey K. Aronson. Adverse drug reactions: definitions, diagnosis, and management. *The Lancet*, 356(9237):1255–1259, 2000. doi: 10.1016/S0140-6736(00)02799-9.

[77] Mick P. Couper. *Designing Effective Web Surveys*. Cambridge University Press, Cambridge, 2008. ISBN 9780521869156. doi: 10.1017/CBO9780511499371.

[78] Todd C. Knepper and Howard L. McLeod. When will clinical trials finally reflect diversity? *Nature*, 557(7704):157–159, 2018. doi: 10.1038/d41586-018-05049-5.

[79] Jonathan M. Loree, Swati Anand, Arvind Dasari, Joseph M. Unger, Amit Gothwal, Lauren M. Ellis, Gauri R. Varadhachary, Scott Kopetz, and Kanwal P. S. Raghav. Disparity of race reporting and representation in clinical trials leading to cancer drug approvals from 2008 to 2018. *JAMA Oncology*, 5(10):e191870, 2019. doi: 10.1001/jamaoncol.2019.1870.

[80] Brian Knutson, Scott Rick, G. Elliott Wimmer, Drazen Prelec, and George Loewenstein. Neural predictors of purchases. *Neuron*, 53(1):147–156, 2007.

[81] Colin Camerer, George Loewenstein, and Drazen Prelec. Neuroeconomics: How neuroscience can inform economics. *Journal of Economic Literature*, 43(1):9–64, 2005. doi: 10.1257/0022051053737847.

[82] Richard H. Thaler and Cass R. Sunstein. *Nudge: Improving Decisions About Health, Wealth, and Happiness*. Penguin, New York, 2008.

[83] Hilke Plassmann, John O'Doherty, Baba Shiv, and Antonio Rangel. Marketing actions can modulate neural representations of experienced pleasantness. *Proceedings of the National Academy of Sciences*, 105(3):1050–1054, 2008.

[84] Dan Ariely. *Predictably Irrational: The Hidden Forces That Shape Our Decisions*. HarperCollins, New York, 2008.

[85] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021. doi: 10.1145/3458723. URL `https://dl.acm.org/doi/10.1145/3458723`.

[86] The National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. The belmont report: Ethical principles and guidelines for the protection of human subjects of research, 1979. URL `https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/read-the-belmont-report/index.html`.

[87] An-Wen Chan, Jennifer M. Tetzlaff, Peter C. Gøtzsche, Douglas G. Altman, Helen Mann, Jesse A. Berlin, Kay Dickersin, Asbjorn Hrobjartsson, Kenneth F. Schulz, Wendy R. Parulekar, and et al. Spirit 2013 statement: Defining standard protocol items for clinical trials. *Annals of Internal Medicine*, 158(3):200–207, 2013. doi: 10.7326/0003-4819-158-3-201302050-00583.

A Roadmap for Integrating Cognitive Science and ML into Clinical Trials

A.1 Phase 1: Foundations—Cognitive instrumentation of clinical data practices

The first phase equips existing workflows with cognitive overlays that respect trial operations. For consent, we pair the standard PDF with brief teach-back checks delivered by a conversational interface, using readability-aware rewrites and adaptive clarification to reduce cognitive load. For ePRO survey fatigue, we layer adaptive phrasing and timing atop the current app, modeling habituation while preserving endpoints and schedules. For AE reporting, we introduce clinical cover stories that normalize sensitive categories and add temporal anchors tied to daily routines to mitigate telescoping; outputs are written back as structured fields with confidence scores and an audit trail. Each workflow must pass BCA gates before claims of improvement: for consent, a predefined comprehension-gain threshold and stable performance across reading levels; for ePROs, reduced missingness without loss of calibration; for AEs, bounded timestamp error relative to clinician adjudication and increased yield without unacceptable false positives. Mechanism checks remain surgical (e.g., a representational test showing the model tracks time-anchor features), and development remains minimal (dataset/model cards, prompt/version control, weekly drift snapshots). Preconditions include IRB-approved guardrails, privacy governance aligned with site requirements, and non-disruptive integration to the trial's EDC.

Goal: Address overlooked challenges by building cognitive models of participant behavior.

1. Refining Consent
   (a) Apply cognitive load theory to redesign consent forms and test comprehension experimentally.
   (b) Pilot LLM-mediated consent conversations that adaptively clarify participant doubts.
   (c) Rather than replacing existing PDFs, introduce a "two-factor" sampling mechanism: participants review a short excerpt or summary and are then prompted with comprehension checks or clarifying dialogues before proceeding, ensuring comprehension without overhauling existing infrastructure.
2. Combating Survey Fatigue

    (a) Develop models of habituation by simulating how repeated prompts reduce attention and engagement.
    (b) Explore adaptive ePROs that vary question framing/timing using reinforcement learning to sustain engagement.
    (c) Integrate these adaptive mechanisms as overlays to existing ePRO applications rather than entirely new systems, minimizing disruption.

3. Improving Event Reporting
    (a) Test hybrid systems where participants' AE narratives (elicited through LLM dialogue) are temporally anchored by cognitive recall models.
    (b) Benchmark against standard self-report logs for accuracy.
    (c) Deploy these conversational probes at pre-specified intervals (e.g., at routine check-ins) so that they augment, rather than replace, existing AE reporting workflows.

Outcome: Cognitive-aware clinical data collection protocols that can be layered onto existing trial infrastructures (e.g., PDF forms, ePRO apps, routine AE logs) with minimal overhead, enabling rigorous evaluation alongside traditional methods.

A.2 Phase 2: Patient-in-the-loop models for trial operations

Having instrumented workflows, we integrate our patient-first cognitive model more tightly with site practice while retaining human control. Behaviorally, the model is initialized with behavioral priors (Centaur-style training on human choices and latencies) and adapted via the thin clinical layer to trial tasks: consent comprehension, AE timing/reporting, preference trade-offs, and bias-aware disclosure support. Clinicians receive calibrated summaries with explicit uncertainty; low-confidence or out-of-distribution cases are automatically deferred. We evaluate dropout-risk prediction, tacit-knowledge yield from naturalistic dialogue, and clinician workload, always reporting subgroup-parity metrics. Processing commitments add one causal test per task (e.g., an activation-patching or ablation study showing that disrupting a "time-anchor" representation degrades AE dating), linking internal processes to reliability. Development expands to shadow training and safe online adaptation under SOPs: updates occur only when BCA gates and parity thresholds are met, with rollback plans and versioned audits. Preconditions include site-level SOPs for escalation/deferral, staff training, and live interoperability checks with CDISC mappings.

Goal: Close the blind spots across ML, cognitive science, and clinical research.

1. Modeling Human Comprehension
    (a) Develop integrated ML and cognitive systems that explicitly simulate participant comprehension and decision-making (e.g., probabilistic models of consent understanding).
    (b) Incorporate dropout risk predictors that factor in cognitive states and engagement histories.
2. Tacit Knowledge Extraction
    (a) Deploy cognitive-LLMs in trial settings to conduct structured yet conversational interviews, surfacing information participants might not disclose in formal surveys.
    (b) Compare yields with standard structured reporting to quantify hidden knowledge capture.
3. AE Detection
    (a) Train models to cross-reference conversational data, response latencies, and linguistic markers to infer unreported AEs.
    (b) Validate against clinician follow-ups and medical records.

Outcome: Integrated ML and cognitive models that reduce dropout, increase AE reporting fidelity, and provide richer behavioral data streams.

A.3 Phase 3: Transformation: Cognitive-integrated trials and regulatory alignment

The final phase reframes trials as cognitive ecosystems in which patient narratives are first-class signals, i.e., auditable, calibrated, and ethically governed. Behaviorally, cognitive-interaction data (e.g., comprehension checks, temporal-anchor success, disclosure comfort) are specified as exploratory endpoints alongside traditional PROs and biomarkers, with preplanned analyses for clinical relevance. Processing commitments mature into stable mechanism specifications—the small set of internal signals that are continuously monitored for drift because they are causally tied to reliability (e.g., the anchor-tracking channel for AE dating). Development formalizes a lifecycle: retraining cadence, bias and safety audits, subgroup parity dashboards, and governance for change management across sponsors and sites. We outline regulatory mapping so that cognitive-interaction outputs are validated within existing guidance, and we document patient-centric benefits (comprehension, trust, retention) alongside operational impact (AE fidelity, clinician workload). Preconditions include cross-stakeholder governance (sponsor, CRO, site, IRB), privacy/compliance sign-off, and agreed thresholds for promotion from exploratory to supportive evidence.

Goal: Create epistemic and ethical shifts in how clinical trials are designed, interpreted, and regulated.

1. Ethical Diversity Modeling
   (a) Use cognitive ML-driven simulations to explore how diverse patient groups interact with trial protocols.
   (b) Develop adaptive recruitment and retention strategies sensitive to cognitive and cultural differences.
2. Incentive-Aware Trial Design
   (a) Incorporate behavioral economics insights to align participant incentives (e.g., framing adherence not as compliance but as agency).
   (b) Test cognitive LLM-mediated motivational feedback loops to sustain engagement.
3. Regulatory and Epistemic Shifts
   (a) Propose frameworks for regulatory acceptance of conversational and cognitive-response data as valid clinical endpoints.
   (b) Argue for expanding trial evidence bases beyond biomarkers and structured ePROs to include cognitive-interaction data streams.

Outcome: A new paradigm of "cognitive-integrated trials" where human responses are not noise to be controlled but signals to be modeled, analyzed, and ethically incorporated into clinical decision-making.

By moving from cognitive overlays (Phase 1) to patient-in-the-loop integration (Phase 2) and finally to cognitive-integrated trials (Phase 3), the roadmap delivers measurable improvements, i.e., higher AE yield with better timing fidelity, stronger consent comprehension, and reduced fatigue without disrupting infrastructure. The result is a patient-first, auditable pathway that demonstrates behavioral competence, links reliability to tractable internal processes, and establishes a sustainable development and governance model for clinical deployment.

B Case Study Design: Enhancing AE Reporting

Objective. Evaluate whether a cognitively informed, patient-first conversational layer (our cognitive ML model for clinical research) improves the completeness and temporal accuracy of AE reporting—relative to standard structured questionnaires—while maintaining calibration, subgroup parity, and clinician control.

Setting and patients. The study is embedded in a Phase III cardiovascular trial with 100 adult patients observed during a four-week intensive AE-monitoring window.

Design. We use a parallel-arm randomized design comparing standard AE questionnaires at fixed intervals to an ePRO application instrumented with a lightweight conversational overlay. The overlay sits atop existing AE logging and maps outputs to standard EDC/CDISC fields without altering endpoints, visit schedules, or data standards. The model is initialized with behavioral priors derived from a behavioral foundation model (Centaur-style training on human choices and latencies) and adapted via a thin clinical layer; the artifact of record is our clinical model, not the base foundation model.

Intervention. The conversational layer deploys clinical cover stories that normalize sensitive categories in order to reduce social desirability bias and encourage disclosure, and it uses temporal anchors—such as prompts tied to meals or wake/sleep cycles—to mitigate telescoping and improve event dating. Prompt phrasing and timing adapt when hesitation cues are detected in language or latency. All outputs carry calibrated confidence; when confidence falls below a preregistered threshold, the system defers to clinician review by design. Integration is intentionally lightweight: the overlay augments the current ePRO workflow, pairs consent-period comprehension checks with brief teach-back items rather than lengthy retraining, and produces a structured AE summary (event type, onset/offset, severity, impact, confidence) with an auditable trace of prompts and responses.

Outcomes and endpoints. The co-primary endpoints are incremental AE yield—defined as additional clinician-adjudicated AEs per patient versus control—and temporal accuracy—defined as absolute error in AE onset time against clinician or record-based adjudication. Secondary endpoints include patient comfort and willingness-to-disclose measured on a brief Likert index, probabilistic calibration assessed via expected calibration error and Brier score with selective-prediction curves, and clinician workload measured as review minutes per patient. Exploratory outcomes quantify linguistic and latency markers of hesitation, subgroup parity in incremental yield and timestamp error across age, sex, and race, and weekly drift in calibration and yields.

BCA gates. Prior to any deployment claims, the model must meet preregistered gates: a BCA-Temporal threshold specifying that median timestamp error does not exceed a predefined window (for example, 24 hours) relative to adjudication; a BCA-Disclosure threshold demonstrating a statistically significant increase in clinically relevant AE categories without an unacceptable rise in false positives; a calibration threshold requiring expected calibration error below a preset bound with a deferral policy that routes a prespecified share of low-confidence cases to clinicians; and a parity threshold bounding subgroup gaps in incremental yield and timestamp error by a small δ selected with clinical input.

Analysis plan. Primary analyses follow intention-to-treat with difference-in-means or rank-based estimators and 95 percent confidence intervals. Hierarchical models adjust for baseline symptom burden and visit adherence. Sensitivity analyses include per-protocol estimates, clinician-review time effects, and robustness to prompt variants using preregistered A/B phrasing. Errors are categorized into non-AE false positives, misdated events, and severity misclassification, and are summarized with confusion matrices and calibrated risk plots.

Safety, ethics, and governance. The clinician remains in the loop for all flagged events; the model cannot modify treatment or finalize outcomes. Data are processed locally or within an approved secure enclave, and all interactions are logged for audit. Weekly bias and drift audits evaluate parity and calibration; failure to meet gates triggers rollback to baseline prompts. Documentation comprises a model card, dataset card for the clinical layer, and a standard operating procedure detailing deferral and escalation pathways.

Expected Contributions. The study demonstrates that a thin, auditable clinical layer can translate cognitive-behavioral priors into higher AE yield and better timing fidelity without disrupting infrastructure, while preserving safety through calibration and deferral, building trust via auditability, and promoting equity through explicit parity gates.

While our roadmap outlines a broad vision for integrating cognitive science and ML into clinical trial design, concreteness is essential for evaluation and critique. To this end, we present exemplar case studies that operationalize the approach in specific trial contexts. Each targets a distinct blind spot from our problem framing—patient disengagement (survey fatigue) and adverse-event underreporting—and shows how cognitively informed, LLM-mediated interventions can be embedded as lightweight overlays on existing workflows. These are not full-scale deployments but structured, auditable prototypes—using clinical cover stories, temporal anchoring, calibration and clinician deferral—that demonstrate feasibility, stimulate empirical inquiry, and clarify the methodological shifts we contend are necessary.