# INFERENCING LEARNER MENTAL ATTITUDES WITH PROPOSITIONAL INFERENCES OF GENERATIVE AI MODELS

Muhammad Fusenig

Luke Butler

University of Maryland, College Park

**Abstract**

This proposal outlines a study evaluating the capability of current Generative AI models to accurately infer the mental attitudes and reasoning processes of learners in educational contexts. While current pedagogical theory emphasizes the important role of subjective mental attitudes like motivation and self-efficacy in learning, practical assessment currently relies on self-report or extensive observation. Generative AI presents an opportunity to interpret such mental qualities more efficiently. This study will first establish baseline Generative AI model performance on learner inferred Theory of Mind tasks. Models will then be evaluated against observed and self-reported inferences of mental attitudes. A new reasoning strategy enabled by propositional inferences will be tested and compared against baseline Generative AI model, human-rater, and self-report accuracy. This study will provide valuable insight into the capabilities of current Generative AI models to make accurate inferences of learner mental states and attitudes, clarifying the role such technologies can play in advancing research-based assessment and accommodation of subjective learner experiences.

## Theoretical Framework

Generative AI has proliferated all strata of society, with particular attention being paid to its impact on education (Office of Educational Technology 2023, UNESCO 2023). With 13% of students between the ages of 13-17 having used ChatGPT and other Generative AI tools for school related assignments, it seems the impact is underway (Pew Research Center 2023). Recent advances in the efficacy of Generative AI models on a variety of tasks and benchmarks, such as causal reasoning and Theory of Mind (ToM) capabilities, have only made the transition

to AI-enabled pedagogy more accessible (Jamali 2023, Kosinski 2023). While such research on Generative AI model capabilities has led to rapid advances in information retrieval (Khattab et al., 2023), text and code generation (OpenAI 2023), logical reasoning (Nye et al., 2021; Wei et al., 2022; Kojima et al., 2022), and other pedagogically relevant advances, little research has been devoted to the application of Generative AI model capabilities to downstream tasks relevant to educational psychological interest, namely the inferencing of learner mental attitudes.

Presupposing the recent progress in AI are established mental constructs within the field of educational psychology. Deeper examination of motivation, self-perception, self-efficacy, and other learner mental attitudes place particular interest in the subjective, learner experience on knowledge acquisition and development (Alexander, 1997). Efforts by Alexander and Murphy demonstrate the influence of such mental attitudes on both the process and outcomes of learner achievement (Murphy and Alexander, 2000). This joint focus on motivation and self-perception is corroborated by findings from the literature of Positive Psychology, by which increased self-efficacy is tied to positive self-appraisal and performance outcomes within educational settings (Smith et al., 2023). Such findings demonstrate the link between a wide variety of identified mental attitudes and performance outcomes (Acosta-Gonzaga 2023, Alexander 1997). However, as with recent successes in the field of Artificial Intelligence, the implementation of findings from Educational and Positive Psychology, respectively, leaves much to be desired.

One problem with the application of such constructs in classrooms is the formulation of the construct itself. Among a number of definitions, Motivation is defined as *"the physiological process involved in the direction, vigor, and persistence of behavior"* (Alexander 2000), with self-perception and self-efficacy being similarly subjective (Bandura 1977). The preferencing of

the individual learner's experience has advantaged the subjective learner, who is privy to private mental attitudes and thoughts. But this preference may simultaneously disadvantage educators, for whom inferencing mental attitudes and individual student thought requires high levels of investment and psychological know-how. Coincidentally, this gap between theory and application may be responsible for the rapid adoption of Generative AI in the classroom.

With ChatGPT's release in November of 2022, Generative AI has rapidly adopted for its potential to provide contextually relevant, personalized instruction (Javaid 2023, Dan et al. 2023, UNESCO 2023). The open access educational service Khan Academy echoed these sentiments in early 2023, stating that their Generative AI service Khanmigo could *"give every student a guidance counselor, academic coach, career coach, life coach."* (Khan 2023). Yet despite the rapid adoption of AI tutors, chatbots, and other Generative AI technologies amongst learners and educators, alarms are still being raised.

The unreliability of Generative AI models, as well as their implicit biases, have raised fears concerning the use of Generative AI in classroom settings (Kamalov). It is clear that individuals construct mental representations of AI models (Holliday). The agentic nature of chatbots raises a sobering question about how learners and educators perceive, interact, and infer the interpretive faculties Generative AI models are (or are not) assumed to have. Resultantly, there is an open question over whether Generative AI possesses the capacity to provide accurate learner appraisal.

These systematic issues draw attention to the unique predicament educators and learners now find themselves in. A position in which they must navigate the reliability of Generative AI for pedagogical use, with no established reference to the current paradigm of educational

psychology constructs. Such asymmetries leave educators with few pedagogical anchors, and in a media environment rife with misinformation and speculation about the capabilities of Generative AI there is the risk of further engraining or producing new existing pedagogical myths that undermine research-based educational practices, a phenomenon still pervasive in education writ large.

While progress has been made in reducing outdated or erroneous beliefs concerning pedagogy, the achievement of that goal is yet forthcoming. As we have seen, educators themselves may be purveyors of such mythologies, marrying pop psychological ideas to reputable pedagogical theory (Kirschner 2013). One pertinent example being the advancing literature on subjective learning experiences and the conjoined myth of learning styles (Nancekivell 2020). As with all paradigm shifts, there is the risk of engraining biases and myths concerning Generative AI's interpretive abilities. It is thus imperative to establish a clear and shared understanding of the scope of Generative AI capabilities for pedagogical use, especially concerning the importance that learner mental attitudes have on acquisition and demonstration.

As demonstrated in the educational psychology literature, pedagogical use is not limited to the veracity of a statement or fact, but the direct relation of such statement or fact to the mental attitude of the learner themselves (Alexander et al. 2009). Further, it is generally accepted that complex theories of mind develop at a relatively young age, with competing theories claiming such developments occur within 15 months to four years of life (Onishi & Baillargeon 2005). While integration of Generative AI spans all levels of education, there is no data on Generative AI's ability to assess or accommodate learner mental attitudes or complex theories of

mind of any age—components necessary to gage the effectiveness of teaching, as well as learning outcomes and objectives.

ToM benchmarks have recently emerged within the Machine Learning and Artificial Intelligence literature, namely the Simple Stories task and FANToM. These tasks are modeled from traditional ToM tasks and draw inferences of the mental states of individuals (Kim et al. 2023). Yet no formal studies have been conducted within the field of educational psychology to determine the validity of such benchmarks in a real or academic setting, much less across demographics or spans of cognitive development.

Current advances in causal reasoning, information retrieval, and humanlike conversational ability in Generative AI have led to its rapid adoption in educational settings. Yet the authoritative nature of its output and its now commonplace use amongst learners and educators belie a more serious question of Generative AI's capabilities to interpret or supply information accommodating of learner mental attitudes. Assessing current and near capabilities of Generative AI to infer subjective, learner mental attitudes will provide greater clarity on the role Generative AI is to play in research-based pedagogy.

## Research Objectives

The primary objective of this study is to evaluate the capabilities of current Generative AI models to accurately inference mental attitudes and theories of mind within learner populations. As outlined in the Theoretical Framework, the ability to assess subjective learner experiences is critical for effective pedagogy and learning outcomes and the application of Generative AI thereof. However, current methods of assessing mental attitudes like motivation, self-efficacy,

and other attitudinal constructs rely heavily on learner self-reports or time-intensive observations by educators. Generative AI presents an opportunity to infer these attitudes more efficiently and without educator investment.

This study has three central aims:

(1) Establish current performance baselines of Generative AI on existing ToM benchmarks.
(2) Develop a novel method using propositional analysis of texts to inference mental attitudes of learners.
(3) Evaluate Generative AI models on their ability to make accurate ToM inferences.

More specifically, the first objective is to validate existing ToM benchmarks within an educational context. Models such as GPT-4, Llama-2, and Mistral-7B will be evaluated on established ToM tests, such as traditional False Belief tasks, the Strange Stories metric, and the FANToM task, (Kim et al. 2023). Several methods for making these propositional inferences will be developed and tested, including zero-shot prompting (Kojima et al. 2022), multi-shot prompting (Brown 2020), tree-of-thought prompting (Yao et al. 2023), symbolic reasoning (Pan et al. 2023), and propositional inferencing. Performance will be assessed by analyzing model accuracy in identifying learner mental attitudes by a team of raters.

The second objective is to introduce and evaluate a novel method for learner mental attitude inferences. We will evaluate the propositional analysis of idea units within educator supplied materials and learner generated responses (i.e. text). This new approach will move beyond existing benchmarks by appending Generative AI models with an interpretable framework for analyzing and conducting inferences on components of learner generated

responses. The model will identify key idea units in assessment materials and student responses, making propositional inferences about the mental attitudes, beliefs, and causal factors that learners may have experienced while engaged in their response to the material.

The third objective is to evaluate the inferences made by Generative AI models. Inferred mental attitudes will be compared against human-rater judgments and learner self-reports to determine the validity of the approach. Models that demonstrate the highest rates of agreement with human judgment and learner self-report will then be recommended as having the strongest current capability for assessing learner attitudes.

## Research Plan

### Participants

The participants will be 90 children aged 8-10 recruited from local elementary schools. Parental consent and child assent will be obtained prior to participation. Participants will be randomly assigned to one of three groups (n=30 per group): a control group receiving no feedback, a positive affect group receiving encouraging feedback, and a negative affect group receiving apathetic feedback.

### Materials

Participants will complete 5 linguistic problems, with each problem increasing in conceptually difficulty. Problems will assess comprehension, reasoning, and application of linguistic concepts. After each problem, participants will answer questions about their reasoning process, attitudes, motivation, and emotions during the task.

## Generative AI Inference Strategies

Three Generative AI models, GPT-4, Llama-2, and Mistral-7B will be utilized with five different processing strategies:

*Zero-shot prompting*

The model will be given a standardized prompt and participant transcript, then prompted to make inferences on participant mental attitudes without any additional context or follow-up questions.

*Multi-shot prompting*

The model will be given a standardized prompt and participant transcript, in addition examples of successful ToM inference tasks. The model will then be prompted to make inferences on participant mental attitudes without any additional context or follow-up questions.

*Tree-of-thought analysis*

The model will be given a specialized tree-of-thought prompt and participant transcript. The model will then be prompted to diagram the participant's inferred chain of reasoning in order to infer participant mental attitudes.

*Symbolic reasoning*

The model will be given a specialized symbolic reasoning prompt and participant transcript. The model will identify key idea units in the responses and use symbolic logic to make inferences about participant attitudes and beliefs.

*Propositional inferencing*

The model will be given a specialized propositional inferencing prompt and participant transcript. The model will break down survey questions and participant responses into discrete idea units. The model will then construct propositions to make targeted inferences about attitudes, reasoning chains, and other mental states.

## Procedure

Participants will first complete three habituation problems to become accustomed to the format. They will then be presented with five test problems:

Problem 1: Low difficulty

Problem 2: Low-moderate difficulty

Problem 3: Moderate difficulty

Problem 4: High-moderate difficulty

Problem 5: High difficulty

After each question, participants will explain their reasoning process and describe their mental attitudes during the task by responding to standardized questions posed by the researcher. After explaining their reasoning process, the control group will receive no feedback. The positive affect group will receive encouraging praise about their reasoning process. The negative affect group will be asked probing questions about their reasoning process with no positive feedback.

All responses will be transcribed and inputted into the Generative AI model, which will be prompted to infer the mental attitudes and reasoning process of the participant based on their responses. Transcripts will also be evaluated by teachers familiar with the participants who will judge the attitudes and reasoning of participants.

## Hypotheses

Based on prior research on the ToM capabilities of Generative AI models, the authors hypothesize that models with higher performance on established ToM benchmarks will demonstrate greater accuracy in inferring mental attitudes of learners. More specifically, it is predicted that models such as GPT-4 with strong benchmark scores will show higher agreement with human-rater and self-report judgments of learner attitudes compared to models with poorer benchmark performance, such as Llama-2 and Mistral-7B. Furthermore, it is hypothesized that the degree of accuracy on existing ToM tests will correlate positively with the accuracy of mental state inferences made by models on the current educational tasks.

Further, the authors hypothesize that the implementation of reasoning strategies, such as tree-of-thought prompting and symbolic reasoning, will improve the accuracy of Generative AI models in inferring learner mental attitudes relative to baseline models, but that human-raters and

self-report will still exceed reasoning enabled models. More specifically, it is predicted that symbolic reasoning and tree-of-thought approaches will result in more accurate inference rates compared to zero-shot and multiple-shot prompting but will fall short of human made inferences.

Lastly, the authors hypothesize that the propositional inference approach conducting fine-grained propositional analysis of idea units will result in the highest rates of accuracy in inferring learner mental attitudes compared to all other models and human judgments. More specifically, by breaking down both measurement questions and participant responses into discrete propositions linked to mental states, attitudes, and beliefs, inferences made by this model will be predicted to reach parity with, if not outperform, educator and participant appraisal.

## Evaluating Inference Sources

To evaluate and compare the accuracy of inferences made by human-raters, learner self-appraisals, and respective Generative AI models, inference agreement scores will be calculated between each pair of sources.

*Human-Model Agreement:*

Proportion of inferences with matching propositional content between human-raters and Generative AI model processing strategies.

*Self-Model Agreement:*

Proportion of inferences with matching propositional content between learner self-appraisals and Generative AI model processing strategies.

*Human-Self Agreement:*

Proportion of inferences with matching propositional content between human-raters and learner self-appraisals.

One-way repeated measures ANOVAs will be conducted for each processing strategy to determine if there are significant differences in agreement rates depending on the source of inferences (human, self, model). Post-hoc t-tests with a Bonferroni correction will be used to make pairwise comparisons between sources for each processing strategy. It is predicted that for most processing strategies, human-model and human-self agreement will significantly exceed self-model agreement. The role of propositional inferencing in closing this gap will be examined.

## Significance

The capability of Generative AI models to accurately infer subjective mental attitudes and learner reasoning processes carries major significance for the field of educational psychology and pedagogical practice. As outlined in the theoretical framework, constructs like motivation, self-efficacy, and other attitudinal factors play a decisive role in learning outcomes and the realization of student potential. However, practical assessment of such subjective experiences currently relies on learner self-report or extensive observation by educators—approaches that are not only resource and time heavy but are retroactive and bypass critical points of intervention. As

such, there exists a practical barrier between the emphasis placed on mental attitudes and reasoning chains within pedagogical theory and the implementation of such findings in practice.

Generative AI offers the potential to infer these subjective learner states more efficiently without demanding additional effort from students or teachers. Preliminary applications of Generative AI in education have focused predominantly on content generation, information retrieval, and conversational ability. Yet as this technology becomes further embedded within pedagogical contexts, it becomes necessary to evaluate benchmark performance on tasks requiring inference of learner mental states and theories of mind. Without established benchmarks tailored to educational settings, the suitability of Generative AI for research-backed assessment and accommodation of subjective learner experiences remains unclear.

This study tackles this open question by validating performance of leading Generative AI models and reasoning processes on an array of ToM tasks situated within an academic context. In doing so, the study provides an empirically grounded point of reference for practitioners seeking to integrate AI within assessment practices emphasized by prevailing learner-centered paradigms. Perhaps more significantly, the introduction of a novel propositional inference approach enables finer-grained analysis of the reasoning processes and attitudes undertaken by students during learning and reasoning activities.

The propositional method put forth stands to significantly advance the capability of AI systems to make valid appraisals of the chains of thought and motivational states underlying observable learner behaviors. Given the decisive role that factors like self-efficacy, perceived competence, curiosity, and reasoning ability play in achievement, such a methodology for

efficiently surfacing these currently unobservable elements promises to augment the sensitivity of assessments to the learner experience specified by research-backed models.

By equipping Generative AI technologies with the means to infer associated mental attitudes from learner responses, the proposed approach provides practitioners with an interpretable window into subjective processes that would otherwise require extensive individual analysis or self-report. If validated against educator judgments and self-appraisals, the propositional inference framework stands to grant educators and learners unprecedented insight into the rich array of learner mental attitudes, further reducing the gap between theory and practice.

Overall, the current study tackles a critical barrier in realizing efficient and valid assessment of the subjective learner experiences central to contemporary pedagogical thought. In establishing benchmark AI capabilities on situated ToM tests and piloting a methodology for the inferencing of mental attitudes, this proposal carries substantial import for practitioners seeking to integrate research-backed insights on motivation and self-perception with the application of Generative AI learning. Should our hypotheses be true, the novel propositional inference approach promises a more accurate appraisal of learner reasoning and the mental attitudes underlying overt learner behaviors. By equipping Generative AI with the means to interpret and inference on observed learner behavior, this study sets the stage for pedagogically sound integration of Generative AI into educational practice.

References

Acosta-Gonzaga, E. (2023). The Effects of Self-Esteem and Academic Engagement on University Students' Performance. Behavioral Sciences, 13(4), 348. https://doi.org/10.3390/bs13040348

Alexander, P. A. (1997). Mapping the multidimensional nature of domain learning: The interplay of cognitive, motivational, and strategic forces. Advances in Motivation and Achievement, 10, 213-250.

Alexander, P. A., Schallert, D. L., & Reynolds, R. E. (2009). What is learning anyway? A topographical perspective considered. Educational Psychologist, 44(3), 176-192. https://doi.org/10.1080/00461520903029006

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. Psychological Review, 84(2), 191–215. https://doi.org/10.1037/0033-295X.84.2.191

Dan, Y., Lei, Z., Gu, Y., Li, Y., Yin, J., Lin, J., Ye, L., Tie, Z., Zhou, Y., Wang, Y., Zhou, A., Zhou, Z., Chen, Q., Zhou, J., He, L., & Qiu, X. (2023). EduChat: A Large-Scale Language Model-based Chatbot System for Intelligent Education. arXiv. https://doi.org/10.48550/arXiv.2308.02773

Holliday, N. R. (2021). Perception in Black and White: Effects of Intonational Variables and Filtering Conditions on Sociolinguistic Judgments With Implications for ASR. Frontiers in Artificial Intelligence, 4. https://doi.org/10.3389/frai.2021.642783

Jamali, M., Williams, Z. M., & Cai, J. (2023). Unveiling theory of mind in large language models: A parallel to single neurons in the human brain. arXiv. https://doi.org/10.48550/arXiv.2309.01660

Javaid, M., Haleem, A., Singh, R. P., Khan, S., & Khan, I. H. (2023). Unlocking the opportunities through ChatGPT Tool towards ameliorating the education system. BenchCouncil Transactions on Benchmarks, Standards and Evaluations, 3(2), 100115. https://doi.org/10.1016/j.tbench.2023.100115

Kamalov, F., Calonge, D. S., & Gurrib, I. (2023). New era of artificial intelligence in education: Towards a sustainable multifaceted revolution. Sustainability, 15(16), 12451. https://doi.org/10.3390/su151612451

Khan, S. (2023, May). How AI could save (not destroy) education [Video]. TED. https://www.youtube.com/watch?v=hJP5GqnTrNo

Khattab, O., Santhanam, K., Li, X. L., Hall, D., Liang, P., Potts, C., & Zaharia, M. (2022). Demonstrate-Search-Predict: Composing retrieval and language models for knowledge-intensive NLP. arXiv. https://doi.org/10.48550/arXiv.2212.14024

Kim, H., Sclar, M., Zhou, X., Le Bras, R., Kim, G., Choi, Y., & Sap, M. (2023). FANToM: A benchmark for stress-testing machine theory of mind in interactions. https://doi.org/10.48550/arXiv.2310.15421

Kirschner, P. A., & van Merriënboer, J. J. G. (2013). Do learners really know best? Urban legends in education. Educational Psychologist, 48(3), 169–183. https://doi.org/10.1080/00461520.2013.804395

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large Language Models are Zero-Shot Reasoners. arXiv. https://doi.org/10.48550/arXiv.2205.11916

Kosinki, M. (2023). Theory of Mind Might Have Spontaneously Emerged in Large Language Models. arXiv. https://doi.org/10.48550/arXiv.2302.02083

Murphy, P. K., & Alexander, P. A. (2000). A motivated exploration of motivation terminology. Contemporary Educational Psychology, 25(1), 3-53. Academic Press.

Nancekivell, S. E., Shah, P., & Gelman, S. A. (2020). Maybe they're born with it, or maybe it's experience: Toward a deeper understanding of the learning style myth. Journal of Educational Psychology, 112(2), 221–235. https://doi.org/10.1037/edu0000366

Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? Science, 308(5719), 255–258. https://doi.org/10.1126/science.1107621

Pan, L., Albalak, A., Wang, X., & Wang, W. Y. (2023). Logic-LM: Empowering Large Language Models with Symbolic Solvers for Faithful Logical Reasoning. arXiv. https://doi.org/10.48550/arXiv.2305.12295

Pew Research Center. (2023, November 16). About 1 in 5 U.S. teens who've heard of ChatGPT have used it for schoolwork. https://www.pewresearch.org/short-reads/2023/11/16/about-1-in-5-us-teens-whove-heard-of-chatgpt-have-used-it-for-schoolwork/

Smith, A. C., Ralph, B. C., Smilek, D., & Wammes, J. D. (2023). The relation between trait flow and engagement, understanding, and grades in undergraduate lectures. British Journal of Educational Psychology, 93(3), 742–757. https://doi.org/10.1111/bjep.12589

UNESCO, Miao, F., & Holmes, W. (2023). Guidance for generative AI in education and research. United Nations Educational, Scientific and Cultural Organisation.

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., & Narasimhan, K. (2023). Tree of Thoughts: Deliberate Problem Solving with Large Language Models. arXiv. https://doi.org/10.48550/arXiv.2305.10601