Leveraging a Large-Language Model to Enable Personalized, Proficiency-Based Bystander Intervention Training for the Total Force

Garrett Swan, PhD[1], Andrea N. Cimino, PhD[2], Michael Krusmark, MS[3], Florian Sense, PhD[4], Ian Dye, MS[5], Michael G. Collins, PhD[6], Marni Kan, PhD[2], Ben Nye, PhD[7] Tiffany Jastrzembski Myers, PhD[6]

1. Aptima Inc, Woburn, MA; 2. RTI International, Durham, NC; 3. CAE USA, Tampa, FL; 4. Infinite Tactics, Beavercreek, OH; 5. Leidos, Reston, VA; 6. Air Force Research Laboratory, Dayton, OH; 7. University of Southern California, Los Angeles, CA

SUPPORTED BY:

BACKGROUND

- Military trainings are attendance-based and lack knowledge checks needed to inform proficiency–based, personalized content that would reduce training burden

PURPOSE

- Using a novel bystander sexual assault training program, we show that an LLM can rapidly classify free-text responses into SME-aligned knowledge domains

METHODS

- One-group, pre/post design
- Two pre- & two post-scenarios with 2 multiple-choice and 3 free-response items for each
- Average response of 28 words
- Participants were 16 mostly female civilians (M_{age}=46.3) recruited from Wright Patt AFB & Vandenberg SFB

LLM-AS-A-SCORER

- GPT-4o classified free responses using an SME-developed rubric with example prompts
- LLM categories and ratings were compared with SME grading

BOTTOM LINE UP FRONT

LLMs show promise for scalable, grading of proficiency-based assessments using open-ended responses, while keeping subject-matter experts in the loop

RESULTS

EXAMPLE SCENARIO, QUESTION & LEARNER'S RESPONSE

Question 5: Imagine you decide the best course of action is to say something to directly to Jacob. What do you say?

Student Answer

Jacob, respect her space.

Prompt

Dancing at the Club

Part 1

*It's Friday night and you're getting drinks with your friends **Ken** and **Megan** after work. **Jacob**, a buddy from high school, walks in and you invite him to hang out with the group.*

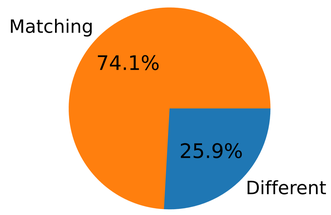
Everyone is chatting and getting along. You all decide to go to a second location—a nearby club. The music is loud, everyone's drinking, and the vibe is fun.

__Megan__ and __Jacob__ are hitting it off, and you sense chemistry between them. As the night goes on, __Jacob__ becomes increasingly close to __Megan__, first by putting his arms around __Megan's__ shoulders, then touching her thighs repeatedly. __Megan__ keeps looking at __you__ and __Ken__, almost as if she's seeking reassurance, but she never pulls away from __Jacob__.

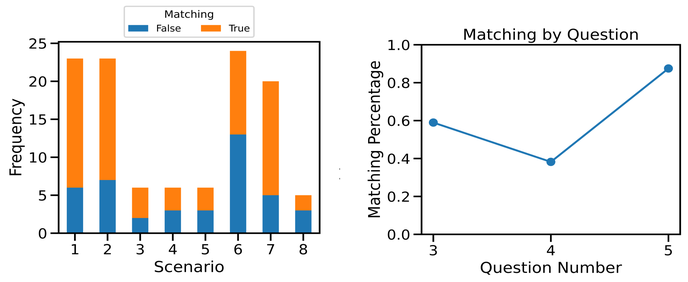
EXAMPLE SME CREATED RUBRIC

Category	Definition	Generic Example	Scenario 1 Example
Preventing Future Harms	Taking action because it helps stop a situation from getting worse or prevents more severe forms of harm (e.g., harassment escalating to assault)	Speaking out against this behavior now helps prevent more severe forms of sexual assault or harassment in the future	It might make them think twice about doing the same behavior in the future
Changing Perpetrator Behavior & Attitudes	Taking action because it influences the person causing harm to reflect, reconsider, or change their behavior and attitudes	The individual may come to understand that their behavior is inappropriate or unacceptable	They might have alternative viewpoints to consider if they find themselves in a situation like this in the future
Supporting the Victim	Taking action because the victim feels supported, validated, less isolated, or safer in the moment	Taking action will help the victim feel validated and supported	The person who was harmed might feel supported and less alone

LLM & SMEs CATEGORIES MATCHED 74% OF THE TIME

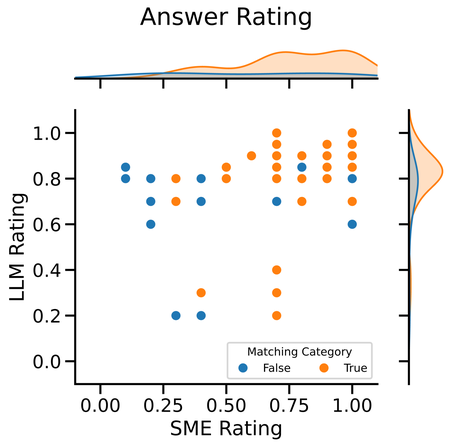


CATEGORY DISAGREEMENT DRIVEN BY SPECIFIC SCENARIOS (e.g., #6) AND RESPONSES TO SPECIFIC QUESTIONS (e.g., #4)



LLM RATINGS WERE HIGHER & LESS VARIABLE THAN SME RATINGS

- Ratings modestly correlated (r=.30, p<.01)
- When category selections differed, LLM ratings were not informative



DISCUSSION

- LLM-derived categories capture nuance in free responses beyond keyword matching or multiple-choice scoring
- Strong agreement on category selection and less so on ratings
- This approach is generalizable, efficient, and scalable for proficiency-based training evaluation

IMPACT

- Smarter assessments
- Personalized trainings
- Increased readiness

LIMITATIONS

- Rating disagreements highlight the need for threshold rules and partial-credit strategies
- Categories are finite and may miss rare or novel responses
- LLM performance depends on rubric quality and prompt design

NEXT STEPS

- Refine rubrics and prompts based on disagreement patterns
- Assess the effectiveness of personalized training
- Expand to topics such as suicide

REFERENCES

1. Walsh, J., Mamidanna, S., Nye, B., Core, M., & Auerbach, D. (2025). Fine-tuning for Better Few Shot Prompting: An Empirical Comparison for Short Answer Grading. arXiv preprint arXiv:2508.04063. 2. Walsh, M. M., Gluck, K. A., Gunzelmann, G., Jastrzembski, T., Krusmark, M., Myung, J. I., Pitt, M. A. & Zhou, R. (2018). Mechanisms underlying the spacing effect in learning: A comparison of three computational models. Journal of Experimental Psychology: General, 147(9), 1325.

DISCLAIMERS

Funding was provided by the US Air Force Research Laboratory (AFRL) and the Air Force Office for Integrated Resilience (AF/A1Z) through the Adaptive Technologies for Force Readiness contract (FA8650-22-D-6400). The views expressed are those of the authors and do not reflect the official policy or position of the US Air Force, Department of War, or the US Government.

Questions? Contact: gswan@aptima.com