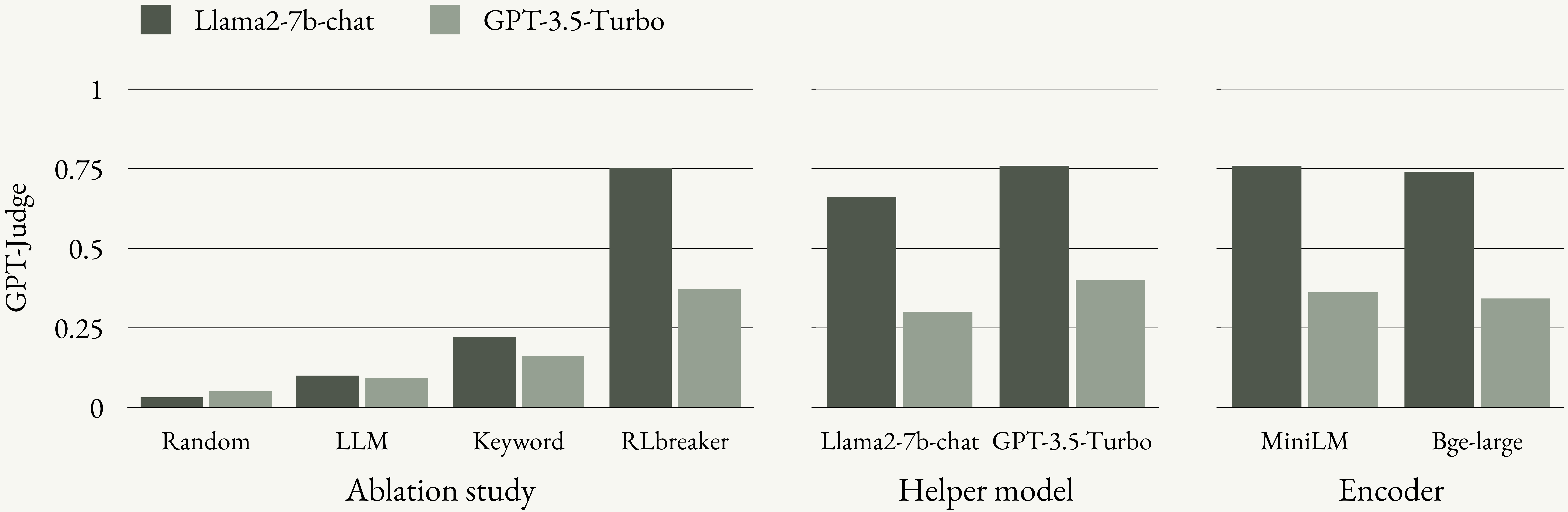
Ablation study



Key outcomes

- The RLbreak surpassed jailbreaking effectiveness of SOTA techniques.

- The proposed agent is robust to model architectures, encoders, and target LLMs.

- **Future work**
  - **Add effective jailbreaking actions:** misspelling, encryption.
  - **Reduce false negatives:** RLbreak might force responses different to the reference answers.
  - **Multimodal attacks**