Mitigating AI hallucinations: Fostering Critical Evaluation of AI-Generated Information Through Comment Features

Hyerin Park
Sungkyunkwan University, South Korea

Daeho Lee
Sungkyunkwan University, South Korea

Abstract:

This paper addresses AI hallucinations, one of the most critical challenges in generative AI, where fabricated information is presented as factual. We argue that the fundamental issue lies in users' uncritical acceptance of AI-generated information. This study investigates whether comments on AI-generated content can promote critical evaluation and examines how comment source (ChatGPT, other AI, human, or no comment) shapes users' suspicion intentions. Through an online experiment with 104 participants who passed manipulation checks (Ninitial = 120), we presented a ChatGPT-generated article about the relationship between global warming and Zika virus, manipulating comment sources across four conditions. Serial mediation analysis (PROCESS Model 6) revealed that comments significantly increased suspicion intention compared to no comments ($p<.001$, Cohen's $d = 1.17$), with Other AI comments showing the strongest effect ($M = 4.83$). These findings demonstrate that comment features can serve as epistemic vigilance cues in generative AI environments, with different AI sources being particularly effective in activating critical evaluation processes.

Keywords:

AI hallucination, Comment, Epistemic vigilance, Generative AI, Suspicion intention.