

We recently submitted a paper to a prestigious journal and received a rather fascinating review, that was long, in perfect English, yet made almost no sense. Due to the suspicion that (all or part of) the review was written with AI, we asked ChatGPT itself if it could write a critical review that would lead to a rejection (with no other prompt engineering or tuning). In the following table, we compare points from ChatGPT with the “human” review. In each row, one of those was part of a ChatGPT review, and the other is part of the “human” review. Later we will clarify which are the “human” ones and which are the ones generated by AI.

A) Nevertheless, similar relationships have been previously noted by Ramdas et al. (2023, Statistical Science). It is not entirely clear whether the current manuscript extends these established results in a substantial manner. A deeper discussion on how these connections provide new insights or improvements in performance would be beneficial.	A) The central theoretical claim—that admissibility in online closed testing requires e-values—is a natural extension of the martingale-based testing framework already established in prior literature (e.g., Ramdas et al., 2020; Vovk and Wang, 2021). The formalization provided here does not introduce fundamentally new theoretical tools or paradigms, but rather reinterprets existing results under a slightly different lens.
B) The proposed SeqE-Guard algorithm appears to be a technical simplification of the closure principle, but its practical utility is not convincingly demonstrated. There is no compelling argument or empirical evidence to suggest that this method will be preferable to existing online FDR control methods like LORD (Javanmard & Montanari, 2015) or SAFFRON (Ramdas et al., 2018) in realistic settings.	B) Given the recent developments of adaptive online FDR procedures (for example, through modified alpha-investing strategies), existing works such as LORD++ and ADDIS can overcome these limitations. The manuscript should provide a more compelling justification for why overcoming this limitation is a significant contribution.
C) The statement on page 3 that the proposed method “uniformly improves all existing online true discovery procedures” appears to be an overgeneralization. More careful phrasing accompanied by empirical or theoretical substantiation would be necessary to support such a strong claim.	C) The authors assert that their procedure offers uniform improvements over state-of-the-art methods. However, the evidence provided does not convincingly support such a broad claim. The manuscript would benefit from a more nuanced discussion of the conditions under which their method outperforms existing techniques and acknowledging scenarios where it may not offer advantages.
D) The notion of admissibility is emphasized throughout, but the practical consequences of violating admissibility in real-world online testing are not made clear. The insistence on e-values as the only admissible tool may be overly restrictive, especially considering the trade-offs between power and control in finite samples. This comes off more as a theoretical artifact than a meaningful constraint for practitioners.	D) The manuscript appears to rely on the assumption of data independence in constructing e-values, which may compromise their adaptability and efficiency in sequential designs, where it would be of substantial interest to utilize past data and side information. To reinforce the contribution of the algorithm, the authors should elaborate on the construction process, discuss strategies for enhancing adaptability (possibly through techniques that leverage the anytime validity of p-value/e-values), and provide a more rigorous justification for their approach.

The “human” review is in rows A and C on the left side (and in rows B and D on the right side). The reviews have the following points in common, which is why we think the “human” review could have been generated by AI (in large part) as well:

1. Most of the claims are not only blatantly wrong, but also wrong in an implausible way. For a non-expert reader, the sentences may seem plausible, but for an expert reader (like us), many sentences are nonsensical, like point (B). (For example, “The authors do not convincingly show that quantum gravitational lensing cannot cause black holes to lose mass.” may seem plausible for a non-physicist but gibberish to a physicist.)
2. The cited works are superficially related, but do not support the actual statement the reviewer is trying to make in any way. E.g., the Ramdas et al. 2023 paper in point (A) is a review paper on game-theoretic statistics that is related to e-values, but does not mention closed testing or online multiple testing at all. This is not a novice versus expert mistake: even novices would not cite irrelevant papers in this specific but incorrect way.
3. The reviewer mentions that certain points should be addressed, but these are already clearly in the paper. E.g. in (D), they say it would be of interest to use past data for the construction of sequential e-values. We devote a paragraph to this in our paper, where we explicitly say that sequential e-values are well suited for such adaptive designs and that it is straightforward to include past data into the calculation of the e-values. Admittedly, this mistake could potentially be a human mistake.
4. The claims made are vague and do not provide any details. E.g., in (C), the claim is made that “uniformly improves all existing online true discovery procedures” appears to be an overgeneralization. In our paper we clearly say which existing procedures we are aware of and explicitly show how each of them can be improved by our method.
5. The reviewer does not use any self-reference. For example, many reviewers use phrases like “I did not understand”, “I am confused by”, etc. This may of course be the style of the reviewer, but such self-references would at least be an indication of a human reviewer. In fact, we looked at many reviews of rejected and accepted papers, and they are often written in first person, or have other human-like features of being specific, somewhat accurate, etc.
6. As far as we can tell, the review is written in perfect English and contains no typos. Once again, this is of course not a clear indication of AI generation. But it is at least conspicuous in comparison to the very sloppy content of the report.

We have also pasted the review into AI detection tools. GPTZero is “moderately confident this text was AI generated” and Scribbr says that 52% of the “human” review is likely AI generated. In comparison, Scribbr assigned the review we completely generated via ChatGPT as 71%. Out of three other human reviews, two were rated 0% and one (quite short review) was rated 42%.