Journal of Scientific Exploration
Anomalistics and Frontier Science

CORRESPONDENCE

Testing Noetic Potential in Large Language Models

Chris Roe
University of Northampton, UK
chris.roe@northampton.ac.uk
orcid.org/0000-0001-8294-4758

Gavin Ritchie
Independent Researcher
g.p.ritchie@nightsborder.com
orcid.org/0000-0001-6774-8986

Michael Daw
University of Northampton, UK
Michael.Daw@northampton.ac.uk

https://doi.org/10.31275/20264011

We read with interest Benjamin Boyle's report of his experiment testing precognition with the Large Language Model (LLM) ChatGPT (Boyle, 2025). As the author noted, this topic has recently generated much interest with anecdotal reports of AI engines supposedly demonstrating various psychic abilities. This clearly raises questions around attributing consciousness to nonbiological systems. We would note particularly that Turing (1950) proposed that one way to differentiate between a machine and a human being would be a human's (but not a machine's) capacity to score above chance at a Rhinean forced choice ESP task. Whilst we welcome the initiative, we have some serious concerns about its instantiation in the paper.

Despite Turing's proposed test, it is not obvious why a forced choice ESP task might be most appropriate in testing psychic ability in an AI. This method has practical advantages in terms of ease of implementation but card-guessing has been superseded by more sophisticated approaches to testing psi, including remote viewing and presentiment protocols that are both referenced in the paper, which give effect sizes that are orders of magnitude larger than for card guessing (Cardeña, 2018).

Additionally, given recent concerns about the need to avoid Questionable Research Practices (QRPs) in parapsychology and other fields (e.g., Bierman et al., 2016), there is an obligation on researchers to ensure either that their study has sufficient statistical power to generate robust findings, or to provide a justification for why this is not the case. There is a substantial body of research that focuses on forced-choice precognition (most recently summarised by Storm & Tressoldi, 2023, in this journal) that gives an effect size estimate of .017. For 80% power this would require 27,200 trials.[1] It is surprising, then, that Boyle reports only one run of just 100 trials, especially when this is reported to have been completed in 30 minutes. Why were there no replication attempts? The study was not pre-registered so it is not clear whether any replications were planned and were unsuccessful. The paper does state that "no warm-up or practice runs were discarded" (p. 351), but this does not preclude attempted replication runs that might have been unsuccessful. It is disappointing that the study was published without a requirement to confirm findings.

Although the study is intended to provide a preliminary test of the capacity for Chat GPT accurately to guess future outcomes, there seems little lability within the system to allow for 'choice' given that ChatGPT employs pseudo-randomisation. In its own words:

> I use *algorithmic randomness*, not a true physical source of randomness. More specifically: The numbers I generate come from a *pseudorandom number generator (PRNG)* built into my model; A PRNG is deterministic—meaning that while the output *appears* random, it ultimately comes from mathematical processes, not physical randomness; I don't have access to external entropy sources (like atmospheric noise, radioactive decay, or hardware RNGs). So: *My "random" numbers are pseudo-random, not truly random*." (ChatGPT v5, 23 February 2026, emphasis in original)

It is surprising, then, that no randomness tests were conducted to assess the numbers generated by ChatGPT. For example, one well-known human bias (a form of the gambler's fallacy) is to avoid calling the same symbol two or even three times in a row (termed doublets and triplets respectively). In a string of 100 random numbers (1-5) we should expect to see on average 19.8 doublets and 3.9 triplets. The sequence of 'guesses' reported by Boyle (pp. 354-355) includes *no doublets or triplets at all*. For comparison, the actual GotPsi sequence of targets includes 23 doublets and 4 triplets, which reasonably reflects chance expectation. A second (related) human bias is to avoid the symbol that was the target for the previous trial. Repeats of this kind should occur by chance on average 19.6 times in 100 trials. The reported sequence contains *no such instances*. When one of us repeated the study (on 22 Oct 2025) using the same instructions reported by Boyle (2025), but using psychicscience.org/esp3 as the experimental platform, ChatGPT again yielded no doublets or triplets and no repeats of the previous target. This raises serious concerns about the adequacy of this source of 'random' guesses.

Although human beings are susceptible to these biases, it is not to this extreme degree, so that it would become a trivial matter to distinguish between a human's guesses and those of a machine attempting to fabricate "human-like" responses. This AI, at least, fails Turing's Imitation Game. It would have been useful to see some screenshot samples showing an interaction sequence between the systems involved, but we are left to imagine this based on a single table of hand-tabulated results.

Another issue we have with the article is that it feels like it was written by an AI tool, for example because it bases its rationale on at least one non-existent reference, Tressoldi and Paladino (2024), which the lead author has confirmed to us is spurious. AI tools are notorious for producing such 'hallucinations', which Chat GPT explains occur because the model tries to provide a helpful answer even when it lacks sufficient information and so predicts likely text based on patterns in training data, and not by "looking up" genuine facts. Academic formats (citations, study summaries) are easy for the model to mimic, and give the false impression that the author has engaged with the sources he cites.

In our view these concerns are sufficiently serious for us to recommend that the article and its findings are set aside until a more substantial and verifiable series of experiments has been conducted, and a rationale is constructed from genuine published work. Given the low resource and time needs, this should not be very challenging. We are not in a position to conclude whether the issues we have raised are indicative of poor scholarship or more concerningly suggest a 'sociological experiment' in the Sokal (1996) tradition.

END NOTE

[1] Calculated using Chat GPT, for study power 0.8 for an effect size (Z divided by the square root of n) of 0.017.

REFERENCES

Bierman, D. J., Spottiswoode, J. P., & Bijl, A. (2016). Testing for questionable research practices in a meta-analysis: An example from experimental parapsychology. *PLoS One*, *11*(5), e0153049. https://doi.org/10.1371/journal.pone.0153049

Boyle, B. J. A. (2025). Testing noetic potential in large language models: A 100-trial precognitive forced-choice study with ChatGPT-4.1-Mini. *Journal of Scientific Exploration*, *39*(3), 348–355. https://doi.org/10.31275/20253739

Cardeña, E. (2018, May 24). The experimental evidence for parapsychological phenomena: A review. *American Psychologist*, *73*(5), 663–677. https://doi.org/10.1037/amp0000236

Sokal, A. (1996). A physicist experiments with cultural studies. *Lingua Franca*, *6*(4), 62–64.

Storm, L., & Tressoldi, P. (2023). Assessing 36 years of the forced choice design in extra sensory perception research: A meta-analysis, 1987 to 2022. *Journal of Scientific Exploration*, *37*(3), 517–535. https://doi.org/10.31275/20232967

Tressoldi, P., & Paladino, P. (2024). Precognition research 1978–2023: A cumulative meta-analysis and assessment of evidential value. *Journal of Parapsychology*, *88*(1), 45–66.

Turing, A. M. (1950). Computing Machinery and Intelligence. *Mind*, *59*, 433–460. https://doi.org/10.1093/mind/LIX.236.433