

LLM-Unjournal-BS research proposal

Proposal for comparing LLM-generated reviews to Unjournal evaluations

See: [Project home page in Coda](#)

[Github space](#), creates content [here](#) (public description of project/output)

“Also see '[other relevant links](#)'”

See '[Project Goals](#)' just below

Relates to a potential collaboration between [The Unjournal](#) and the [Black Spatula project](#), exploring the potential of AI tools.

Implementation and engineering notes (Unjournal):

☰ Automated research checking, evaluation, and discussion pipeline proposal ([link](#): David Reinstein's sketch, proposal ~with Black Spatula:)

The implementation synergizes with ☰ Unjournal LLM tools (Basic chat and feedback tools for Unjournal and potential partners)

Live chat space: Unjournal Slack – [tech-comms LLM](#) – (scratch space, not 'real') Also see [BlackSpatula chat](#)

Project goals

Project Goals

Both applied academic research and practical goals:

1. Provide insights and evidence on the **potential of using LLMs for checking and evaluating research** and providing useful feedback and suggestions, comparing these to human evaluations and to ‘hybrid collaborations’, informing the future of research practice in general. This can also aid our understanding of whether and how AI could be a research leader.
2. **Develop, implement, and demonstrate** these tools, integrating this into The Unjournal’s and Black Spatula’s approaches
(Here is Steve Newman’s sketch of some infrastructure for evaluating which tools, pipelines, etc. work best

 [Meta-Review: Infrastructure for Evaluating Review Pipelines](#))

“Compare LLM-generated reviews to The Unjournal’s human-generated evaluations”:

1. Assess LLMs’ potential to
 - Identify errors and weaknesses in research papers
 - Evaluate research using (or aligned with) The Unjournal evaluation criteria
 - Provide feedback for authors
 - Support human evaluators
2. Help test scalability and feasibility, generate insights for integrating LLMs into Unjournal workflows
3. Characterize (and cross-validate) the nature of human and machine evaluations, as well as their consistency
4. Additional benefits:
 - Using human (Unjournal) evaluations to train and improve LLM and ML research feedback and evaluation
 - Building pipelines & resources enabling a chatbot for engagement with research, evaluations, and policy implications

Other possible goals/approaches (are others interested?)

- DR 26 Apr 2025 – I suggest adding something more closely tied to the ‘ratings and predictions’ we ask evaluators to give; this seems concrete in a useful way
- “How do economists+ currently judge a paper’s quality, the methodological consensus, common themes”
- Consider a treatment where we share LLM reports with evaluators and encourage them to use these, we could compare these with both the automated evaluations and the ones where the evaluators perhaps did not use AI so much.

Do we have a clear ‘academic ~paper goal’ in mind (and does it matter?)

If we’re looking to get academic points for this, do we have a path?

Our absolute advantages:

- our evaluations are quantitative and concrete
- we have some ability to experiment and delay release of data

Research questions, nature of comparison, implementation

Pose questions here

1. What is the role of prompt engineering in the context of LLM capability assessments?
2. What are effective prompts to guide LLM-based evaluations of scientific papers?
 - DR: This (prompts, approaches, parameters) seems like a very useful research output.
 - I imagine training on our existing evaluations, and then testing which most resemble future evaluations
 - Alt: use out-of-the-box LLM models (excluding our data), and test these on our existing evaluations
 - Model that excludes 2022 and later data?
3. What aspects of evaluations are readily automatable by LLMs?
4. How well do LLMs or ML tools perform at predicting journal/bibliometric/measurable impact outcomes relative to human evaluators?
5. Are machine-generated evaluations rated as highly by readers as human ones?
6. How many errors can be detected by LLMs?
7. How much time does it save evaluators (using the existing LLM report to produce an equivalent output)
8. How would we design an evaluator LLM from scratch?
9. “Can LLM’s predict impact of research better than humans”?

Lorenzo – a few possible approaches

1. What tools to plug into The Unjournal’s pipeline and what do they get us?
2. A one-off investigation – using an agent to automate one part of the pipeline, compare to what humans do (~RCT). This involves deciding on a particular tool to evaluate and report results with that only

3. Create a 'benchmark' – standardized set of samples and metrics to measure performance. This allows us to evaluate tools existing today and also future ones under identical conditions.

- lorenzo.pacchiardi@gmail.com – can you follow up with a sketch of what this might look like?
- Something like a 'sample of best evaluations'? What is the 'ground truth'?

Scope: What aspects of evaluations can be compared, what do we care about most?

- Error Detection: Precision and recall for identifying calculation/computational, methodological, statistical, and logical errors.
- Critique Depth: Quality and specificity of feedback (methodology, data robustness, clarity of argumentation)
 - How to assess? Authors rating? Extent of authors' response [blind the authors to which is AI?]
- Tone and Constructiveness: Professional tone, usability for authors, and practical recommendations.
- Alignment with Human Evaluations: Agreement on quantitative ratings (e.g., novelty, rigor, impact) and qualitative judgments.
 - Note: human evaluations may themselves be limited; we may aim at alignment with meta-evaluations – see footnote.¹
- Predictive power for bibliometrics and impact
 - E.g., which predicts 'which journal tier a paper will be published in', or 'how many times it is cited'

¹Lorenzo: Human evaluations may be suboptimal, so evaluating alignment with them may not be ideal. An option to reduce this issue is having meta-reviewers scoring human and LLM reviews in a blind manner, similarly to how it is done here: <https://arxiv.org/abs/2409.04109>. I mean, we should compare with them, but we should not take human evaluations as gold standard.

DR: I agree somewhat but I still think there is also value in an approach where we say 'what if we took the human evaluations as the gold standard'? If the LLM's or more analytical ML tools can get close to that (when asked to do so ... asked to predict the issues raised and the ratings assigned) then that might be a sufficient justification for using them extensively. Getting galaxy-brained: I wonder if and how we can have confidence in a meta-review standard?

Lorenzo: consider LLMs A and B; A produces reviews that are identical to those humans produce (same novelty, impact and rigour scores). LLM B instead produces different scores, but those eventually (after a few years) turn out to be better predictive of the actual impact of the paper. If we rank A and B using the difference of their scores from human ones -> A wins, even though it may be suboptimal.

The question is: can we approximate "the review captures the impact the paper will eventually have" in another way? Maybe using meta-reviewers is a way to do so (although yes, they can be flawed too). Also, meta-reviews can also capture other things such as how much is a review useful for the authors to improve the paper or, more generally, extending the research they have done.

Probably there is not a single best metric to consider; we could study more than one and analyse them in parallel.

- In the future, expand and update the dataset with upcoming evaluations.
- Or later (Metaculus?) assessment of practical impact

How to evaluate/compare evaluations

- Determine metrics like precision, recall, and effort to edit/refine LLM outputs.
- Use blind reviews by independent evaluators?
- Avoid leakage/contamination and : in testing the LLM evaluations, the AI should not have access to Unjournal evaluators or proxies for other predictive outcomes (journal publication, etc.)
- Consider/avoid reverse causality: If evaluators access our tools or use LLM in their evaluation
- Multiple independent LLM evaluations (alex@herwix.com general suggestion, motivated by a Mühlhoff and Henningsen paper)

Select suitable research and Unjournal evaluations for testing:

- Different types of research, complexity
- High quality evaluations
- Evaluations raising specific issues
- ...

Alternative LLMs and how to “train”:

- Work with the Black Spatula project on this – they have engineering experience and are being somewhat systematic. I’m already engaging with them.
- Human feedback (RLHF) and improvement

Casey Wimsatt’s points (moved to

Automated research checking, evaluation, and discussion pipeline proposal)

Alex: the context we feed in and training data we use will determine the sort of outputs we are likely to get.

How may a review-generation benchmark look like

Considering the idea of building a benchmark leveraging The Unjournal’s evaluations

Novelty/strengths

- Economics Focus: Most available datasets are in NLP, Computer Science, or general science. The economics focus allows you to investigate domain-specific review characteristics, terminology,

and evaluation criteria. This could reveal insights into how review quality is perceived and assessed differently in economics compared to other fields

- Papers chosen based on their influence and potential impact.
- Integrating Reviewer Predictions: Few, if any, existing benchmarks incorporate reviewer predictions about a paper's future impact or journal acceptance.
 - Your benchmark could evaluate how well automatic review evaluation aligns with these human prediction
- The dataset can be dynamically updated
- The reviews are high-quality as they have been paid. Some previous datasets collect reviews of mixed quality as the aim is to train models to detect bad reviews from good ones. Instead here we are considering evaluating review generation!

Weaknesses

- Fairly small sample

How to score

Ie, what metrics to use, how to rely on the human evaluations. See [this](#) for more considerations

- Use a rubric with some criteria (critical depth, tone)... This would require human raters or LLM-as-a-judge
- Alignment with Human Evaluations: Agreement on quantitative ratings (e.g., novelty, rigor, impact) and qualitative judgments.
 - Note: human evaluations may themselves be limited; we may aim at alignment with meta-evaluations – see footnote.²

²Lorenzo: Human evaluations may be suboptimal, so evaluating alignment with them may not be ideal. An option to reduce this issue is having meta-reviewers scoring human and LLM reviews in a blind manner, similarly to how it is done here: <https://arxiv.org/abs/2409.04109>. I mean, we should compare with them, but we should not take human evaluations as gold standard.

DR: I agree somewhat but I still think there is also value in an approach where we say 'what if we took the human evaluations as the gold standard'? If the LLM's or more analytical ML tools can get close to that (when asked to do so ... asked to predict the issues raised and the ratings assigned) then that might be a sufficient justification for using them extensively. Getting galaxy-brained: I wonder if and how we can have confidence in a meta-review standard?

Lorenzo: consider LLMs A and B; A produces reviews that are identical to those humans produce (same novelty, impact and rigour scores). LLM B instead produces different scores, but those eventually (after a few years) turn out to be better predictive of the actual impact of the paper. If we rank A and B using the difference of their scores from human ones -> A wins, even though it may be suboptimal.

The question is: can we approximate "the review captures the impact the paper will eventually have" in another way? Maybe using meta-reviewers is a way to do so (although yes, they can be flawed too). Also, meta-reviews can also capture other things such as how much is a review useful for the authors to improve the paper or, more generally, extending the research they have done.

Probably there is not a single best metric to consider; we could study more than one and analyse them in parallel.

- Predictive power for bibliometrics and impact
 - E.g., which predicts 'which journal tier a paper will be published in', or 'how many times it is cited'

[Related work: moved to next tab]

[Meeting notes – following tab]

Related work

Related work

[\[2505.23824\] Reviewing Scientific Papers for Critical Problems With Reasoning LLMs: Baseline Approaches and Automatic Evaluation](#)
– <https://arxiv.org/html/2505.23824v1>

Commenting on this version in Alphaxiv: <https://www.alphaxiv.org/abs/2505.23824>

Work coming from International Conference on Learning Representations (ICLR) machine learning conference

General works

[\[2501.04306\] LLM4SR: A Survey on Large Language Models for Scientific Research](#) survey on LLMs for scientific research. Section 5 is about LLMs for peer review. It splits the use of LLMs in peer review in two approaches: automated review generation and LLM-assisted review workflows. Section 5.2 lists papers developing ways to generate reviews automatically, splitting them into single-model and multi-model system. Sec 5.3 is about LLM-assisted review writing, enhance “three primary functions in the scientific review process: (1) information extraction and summarization, which helps reviewers quickly grasp paper content (eg <https://github.com/WING-NUS/SciAssist/tree/CocoSciSum>); (2) manuscript validation and quality assurance, which supports systematic verification of paper claims; and (3) review writing support, which assists in generating well-structured feedback.”

The most interesting part for our purposes is instead, Sec 5.4 is on benchmarks for peer reviews. Table 5 lists datasets and the metrics they used, but it is not extremely informative. It looks like only a few datasets employed human evaluation. Need to look more into the datasets they mention there.

<https://arxiv.org/abs/2502.05151> Another survey on LLMs for scientific research. They have section 4.5 on peer reviews.

[What Can Natural Language Processing Do for Peer Review?](#) Mostly a position paper for challenges and opportunities for using NLP in peer review. Also discusses how and where NLP should be used. Interesting is the repo with collection of datasets: [GitHub - OAfzal/nlp-for-peer-review](#). Also, their section 7 has some points on data confidentiality which may apply to us when building a benchmark. Section 8 is highly relevant to how to measure the performance of automated review systems. They stress how the assessment of peer reviews are hard to formalise, and therefore makes it hard to evaluate NLP systems. Beyond the formalisation, the experimental setup can also be hard to define. Sec 8.1 discusses how the variables of interest for peer review are not directly observed, rather they are *constructs*.

Therefore, they need to be clearly defined and measured, ensuring that measurements are reliable and valid.

They have an example of operationalisation of review quality, which we could consider adopting with some adaptation. It involves breaking down review quality into multiple elements and relying on NLP tools to assist with their evaluation; also, empirically estimate the reliability and validity. Sec 8.2 talks about how to design experiments, different kinds of experiments (gold-standard evaluation, observational studies, RCT...) and how choosing between them depends on what needs to be measured.

Automated peer review generation benchmarks

- For instance see this paper for a collection of papers and reviews from NLP conferences: [NLPeer: A Unified Resource for the Computational Study of Peer Review - ACL Anthology](#)

Tools for automated review generation

- David R: [Automated Peer Reviewing in Paper SEA: Standardization, Evaluation, and Analysis](#) looks interesting, for example. They are building tools we may be able to leverage (partner with them?). But I wonder how many of them were able to use human evaluations for training the models.
 - Main idea is a process to train an automated peer reviewer system. LP: I don't think the paper or the approach are that great.
 - They rely on this corpus: <https://huggingface.co/snsf-data> (grant review corpus)
- [DocETL](#) is a tool to build pipelines for analysing documents with LLMs. It could be used to build automated peer review tools. Findings: "Our evaluation on four different unstructured document analysis tasks demonstrates that DocETL finds plans with outputs that are 25 to 80% more accurate than well-engineered baselines"

Tools to analyze reviews

- [A Supervised Machine Learning Approach for Assessing Grant Peer Review Reports](#) builds a NLP system that analyses the peer review reports of grant applications, rather than the grant applications themselves. In particular, they create 12 categories that describe how each sentence of the report tackles different aspects. Then they annotate sentences of a collection of reports with humans and fit a NLP system to predict the annotations in a supervised learning fashion. So not really about evaluating or building automated review generation, but helping to understand existing reviewers. Though one could use their method to score the reviews, for instance understanding how much each review touches upon different aspects. However this would rely on a brittle NLP system which may not work well out[side] of the training distribution.
- [Automatic Analysis of Substantiation in Scientific Peer Review](#) collects a dataset of "550 reviews from NLP conferences annotated by domain experts." and then train an NLP system that determines how much the claims of the reviews are substantiated.

Works on using LLMs for other aspects of scientific research

[\[2409.04109\] Can LLMs Generate Novel Research Ideas? A Large-Scale Human Study with 100+ NLP Researchers](#) they recruit human experts to write novel research ideas and other experts to blindly evaluate those ideas and the ones written by a LLM agent they develop. Very extensive study. We can build on their experimental protocol (even though their focus is different).

?Which tools would we use, if any?

[PaperQA2: Superhuman scientific literature search](#) introduces an agent to do literature search tasks.

Interesting part for this project:

- They evaluate that agent by asking it to write Wikipedia-style summaries of properties of some proteins and conduct a study where human evaluators blindly score those and previously-(human-)written articles.
- Also: they use their system to detect **contradictions** in the Literature (related to the Black Spatula project, slightly different focus though).
- Context: Mainly in biology

meeting notes

23 Jun 2025

<https://arxiv.org/html/2505.23824v1> – read and take notes here with hypotheses

6 May 2025 - Reinstein and Valentin

Granola notes from chat with @Valentin Klotzbücher on our progress on the ‘automated evaluation’ project: <https://notes.granola.ai/d/8f2da435-2544-494c-9e9b-b3efa78523a2> – queryable transcript

Project Status & Technical Progress

- Valentin has created initial code files with R and Python implementations
- Explored multiple API approaches:
 - Pure OpenAI implementation
 - LMAR package
 - LangChain (showing particular promise)
- LangChain capabilities identified:
 - PDF splitting and processing
 - Systematic evaluation pipelines
 - Multiple LLM integration
 - JSON/Excel output formatting
- Current focus: Creating reproducible evaluation examples with clear parameters
- Temperature settings and seed variables discussed for controlling LLM output variation

Research Questions & Methodology [Revisited]

- Key research questions identified: [Revisited]
 - How do automated LLM reviews compare to human evaluations?
 - Can LLMs identify errors/bad practices that human reviewers miss?
 - How consistent are LLM reviews across multiple runs?
- Potential validation approaches:
 - Journal publication outcomes as benchmark
 - Expert consensus validation
 - Pre-LLM evaluation comparison
 - Working paper vs final version analysis
- Data sources discussed:
 - Unjournal evaluations
 - “Economics eJournal” reviews
 - Nature Human Behavior referee reports

Next Steps & Action Items

- Valentin to:
 - Clean up and document current code implementation
 - Prepare Quarto document with examples
 - Share updated implementation within next day
- Contact to:
 - Review existing human evaluations to identify themes
 - Book next week's follow-up meeting
 - Consider integration with teaching/student evaluations
- Joint tasks:
 - Set up GitHub repository for project
 - Define minimum working example with 1-2 papers
 - Establish evaluation criteria and benchmarks

Technical Infrastructure Decisions

- [NOT QUITE] Moving away from n8n platform due to:
 - Interface complexity
 - Performance issues
 - Resource limitations
- Adopting LangChain for:
 - Better code transparency
 - Easier scaling
 - Systematic process chains
- Need to implement:
 - API key management
 - Proper GitHub setup
 - Reproducible evaluation pipelines

Feb 7, 2025 |

📅 LLM-generated reviews & Unjournal evaluations (Org...

Attendees: Emmanuel Orkoh Casey Wimsatt david reinstein Lorenzo Pacchiardi
Valentin Klotzbücher

The process, piloting, tools, collaborations

Casey – at Black Spatula, some ~dissapointing results; LLM’s not finding the high value errors

Todo – David Reinstein to suggest one or more papers to feed into “the Demo”

1. Try again with the one you fed in before “reducing pollution in China” ... see if it does better
 2. DR finds 1-2 papers where we had our ‘strongest evaluations’ and evaluations that spotted fairly concrete limitations/errors
 - a. See 📖 Unjournal-relevant papers to try to ‘Black Spatula’
 3. Choose one or more that are ‘currently being evaluated’ as a way of avoiding contamination
 - a. See 📖 Unjournal-relevant papers to try to ‘Black Spatula’
- DR – Let’s be clear about the approach to replicate the recipe on multiple papers
 - DR – feedback loops to improve this?
 - Casey: We don’t have a systematic method to iterate and improve on the prompts etc. (only a spreadsheet that is updated and considered qualitatively)
 - “DSP does this” [DR: what is DSP?]

Casey W:

- James Heathers is doing this systematically, looking for the lower-hanging fruit. His grad students as the ‘human in the loop’.
- Elisabeth Mick [sp?]

→ DR: Maybe we should work with them, apply for a joint grant

Our project and steps forward

Lorenzo, considering literature –

We work in a ‘more interesting’ field perhaps (not just AI and CS research)

We have strong evaluations to compare these to, and to train them on

The LLM evaluations can be a sort of ‘benchmark’

Casey – ‘one of the things the tools do best is to suggest things for people to review/consider’

“*Cruxes*”

- Are others already doing this/ are we novel? (Seems like yes, we have some novelty and differentiation)
- VOI? – Casey says ‘projects in other fields have not had success’ so a ‘failure to perform’ might not be high value
- Are we well-positioned, do we have the expertise?

- Can the LLM's do this at least plausibly, at scale?
- Do people on our team have the bandwidth/funding to pursue this?

DR: Most of this seems high-value for The Unjournal anyways– e.g. people are already asking me whether our evaluations ‘could be done by AI anyways’.

Takeaways

- Try out Casey's steps on our own papers or papers we are very familiar with – so we get a sense of it
- Evaluators may be willing to participate – Casey, please do let me know what you propose

Jan 17, 2025 |

📅 LLM-generated reviews & Unjournal evaluations (Org...

Attendees: alex@herwix.com Emmanuel Orkoh Casey Wimsatt david reinstein
Lorenzo Pacchiardi Valentin Klotzbücher

Agenda:

Brief introductions, goals, how we see the collaboration going

[Potential collaborators](#) – fill in your details/interest and what role you might play

[Project Goals](#)

Consider: Academic/research output, integration with Unjournal and Black Spatula, roles, timing, funding.

[Discuss: research questions, nature of comparison, implementation](#)

Any ‘demos’ or screenshares of content

Relevant 'demos' to consider

On a side note, I assume everyone is busy, but if anyone has time to quickly review/sanity check this triangulated llm review of an old Economic report, that would be great? Pasting part of this morning's thread from Black Spatula whatsapp. If you can post a reply on whatsapp, great, but happy to just update for you.

Here's the result of using the Bookito3 prompt across three different models for a 2007 report from the pps.org site: o1 deep research, gemini fast thinking with temp =0, and DeepSeek, with the responses from the latter two fed back to o1 for synthesis.

They seem to agree there are serious errors in the paper. Here is the full chat on o1:

<https://chatgpt.com/share/67a4e81d-0344-8006-bedc-0b412bd1cb39>

and here is the paper

https://cdn.prod.website-files.com/581110f944272e4a11871c01/5f0df6dfa50297f98899f9ce_pps_public_markets_eis.pdf

David Reinstein very quick response

- This seems very strong and high-value at a quick skim. Nice!
- Is the pipeline here automated/easy to automate for say 10-12 paper?
- I'd love to try this on our previously evaluated papers as well as other work in our pipeline

n8n looks promising; maybe the pipeline could be done in this

Or a developers tool like 'replit'

Crewai less familiar (it's the multi-agent thing) – interesting, but these can go off the rails quickly and be a net loss of time and resources

Propose/agree on concrete next steps

Who does what?

Who should take the lead? Is anyone here ready to take the "PI" role and set a research agenda etc?

- Emmanuel – can help with academic writing
- Valentin – has time and very interested
- Casey: Steve Newman is the leader and the rest is an organic thing

Next steps

David Reinstein: I will be in touch with Black Spatula and reach out to Steve Newman, glean their thoughts on the best approach to the “compare machine evaluation to UJ evaluations” preliminary testing. Also will consider grant application [Note they will be releasing the ‘pipeline protocol’ today– I will try to test it out while being aware of avoiding contamination.]

Lorenzo: Will sketch out a benchmark, look at related literature to consider if we are novel

Valentin: Will keep reading and sketching

Emmanuel: Will look at the ML reviews and literature; how we would fit best

Next meeting in about 2 weeks

Simple way to test with own papers or papers which are in a domain in which you have expertise:


Steps:

<https://aistudio.google.com/app>

1. select Gemini 2.0 Fast Thinking as the model and set temperature = 0
2. then set system instruction to something like "you are a professor of meta science who is concerned about the proliferation of low-quality AI science publications",
3. upload the pdf
4. paste in the prompt, eg: Here are a partial list of Black Spatula prompts (looking for other sheet).

https://docs.google.com/spreadsheets/d/1Eo5BH_shOZXKf63_kA7cuDXTBkva0vteB9CBAddq2TI/edit?gid=1704622449#gid=1704622449

Other relevant discussions

 Unjournal and Black Spatula – conversation with Steve Newman

Updated – see additional notes. Steve really wants a ‘corpus of research with identified and vetted errors’ to use for training and testing models designed to pick this up. I had an idea (with Jeroen van de Ven) to use *changes in paper versions* as a good initial source, assisted by AI/LLM comparison, discussion and classification of these. I’d love your thoughts on [this](#).

VK:

Just found [this potentially relevant paper](#) where they "propose a model for recommending [Wikipedia] edit summaries generated by a language model trained to produce good edit summaries given the representation of an edit diff".

Also note again my comment that for Nature (Human behaviour etc) there should be many cases where we could fully track draft(s)+reports(!)+final paper (where authors did not opt out, see [here](#) for example the now-published "In Review" preprints). I quickly tried with a paper I know ([here](#)), the summary is great there are just no really interesting "errors" caught between versions or maybe some of the points count? I'm trying to find a better case

CW. March 9, 2025:

Is there an added/bonus benefit from this approach - e.g. the AI diff adds intrinsic value so would be eagerly adopted, or these papers are less likely to be in the training data? - or this is just a convenient corpus? Seems worth a try in either case. There were other corpus' for other AI peer review projects, yes, so I assume they were not convenient to access? Another possibility might be to benchmark against deterministic forensic software tools.

How are you all viewing the potential for AI to help reduce noise in Peer Review/publication? I am more convinced that current LLMs will not achieve top notch logical thought because genuine logic (vs learned patterns of logic that exists in texts) is absent from the architecture/design (see Yann Lecun's lectures/projects to add logical thinking via a different architecture/approach). When I add to this the inherent divergent/non-repeatable nature of LLMs, it suggests to me that it is best thought of as an "additional set of eyes" of a grad student/research assistant level. Expect that it will make mistakes and hope that it will have some insights/finds that others have missed.

other relevant links

☰ [Proposal: Comparing LLM-generated Reviews to Unjournal evaluations](#)
(Current doc)

☰ [Automated research checking, evaluation, and discussion pipeline proposal](#) (edited)
(edited)

Here is also the doc with examples:

☰ [Unjournal-relevant papers to try to 'Black Spatula'](#)

And I also found these again now:

☰ [Unjournal and Black Spatula – conversation with Steve Newman](#)

WIP in chatGPT (careful about leakage issues):

☰ [LLM/GPT evaluation and rating of papers in our pipeline](#)