

AI Review Report

Paper: AI Agents as a Research Team: A Human-Supervised Case Study in Stellar Spectroscopy

Paper ID: 101

Review Service: Reviewer3.com

Generated: October 24, 2026 at 13:51 UTC

Round 1 - Completed

Paper version: v1

Submitted: October 01, 2026 at 23:47 UTC

Completed: October 01, 2026 at 23:53 UTC

Reviewer 1

On page 19, the text directly below Figure 8 (which is the 'Baseline Cr abundance sweep') describes the non-monotonic behavior of [Mg/H] and the Mg diagnostic: 'Absorption initially strengthens and the mean flux falls as [Mg/H] rises. At approximately +1.37 dex, however, the mean flux rises relative to the +0.67 dex case and the profiles become shallower. The bot-produced diagnostic therefore reveals a reversal of the earlier trend.' This text actually describes the Mg abundance sweep shown in Figure 10 on page 21, but its placement under Figure 8 on page 19 creates an internal contradiction and mismatch between the figure and the text immediately following it.

Reviewer 2

The comparison between the AI team's turnaround time of less than a week and the estimated human team's time of one academic semester is fundamentally flawed because it does not control for the starting state of the project. The AI team's work was built on the existing database and notebooks supplied by the authors. If a human student or researcher were provided with the exact same pre-existing database and fully functional notebooks, they could also easily run the code, generate the figures, and draft the manuscript in a few days. Comparing the AI team's execution of pre-existing workflows to a human team's typical semester-long research cycle (which usually involves database creation, code development, and troubleshooting from scratch) is highly misleading. To make a fair comparison, the authors should estimate the time a human would take given the same pre-existing assets, or temper the claim to avoid implying a massive speedup that is largely an artifact of the starting conditions.

Reviewer 3

The reported baseline MAE of 0.003681 is presented as a single global average across the entire test set of 21,657 spectra. However, stellar spectra vary dramatically in complexity across the parameter space; hot stars have very few absorption lines and are mostly flat continuum, whereas cool stars have extremely dense and complex line structures. A global average MAE can easily be skewed by the simpler spectra or by the large number of continuum pixels where the error is near zero, masking poor performance in line-rich regions or specific parameter regimes (such as cool stars, where the Mg limitation was noted). To provide a scientifically robust evaluation, the authors should analyze and report how the reconstruction error (MAE/RMSE) scales as a function of key physical parameters, particularly effective temperature and bulk abundance, rather than relying solely on a single global metric.

Reviewer 4

In Table 4, the authors compare five architectures on a small subset of 256 spectra and report single-point metrics to claim that the baseline MLP outperforms the alternatives. However, because the evaluation is restricted to a very small subset (~1.2% of the test partition) and no standard deviations, confidence intervals, or statistical significance tests are provided, it is impossible to determine whether the reported differences in performance are statistically significant or merely artifacts of the specific random subset chosen. To support this claim, the authors should either evaluate the models on the full test set or provide error bars and statistical significance testing for the metrics on the subset.

Reviewer 5

The manuscript presents a case study of a coordinated AI research team comprising nine specialized agents. However, it remains highly ambiguous whether there was any actual multi-agent system (e.g., an automated framework like AutoGen or CrewAI) executing these tasks, or if the coordination was entirely simulated by the human prompters manually copy-pasting messages between different chat threads of a single commercial LLM (Grok). The text mentions rooms and exchanges, but also admits that the available record does not include the complete instructions or configuration used to instantiate each role and does not establish whether the bots used different underlying models or performed their tasks simultaneously. If the team is simply a single human manually routing prompts to a single LLM under different persona instructions, then the claim of demonstrating a coordinated AI research team is fundamentally overstated. The authors should clarify the exact technical architecture of the agent system, specify whether message routing was automated or manual, and provide the system prompts or configurations used to instantiate the agents.

Reviewer 6

In Figure 10, the abundance sweep for Mg shows an uptick in the mean flux at high abundances, which is claimed to represent an unphysical non-monotonic behavior. However, Figure 10 only plots the generated CondGen spectra and mean flux, completely omitting the corresponding SYNSPEC reference curves for the same parameter sweep. To scientifically demonstrate that this behavior is indeed a model limitation (i.e., a departure from physical radiative transfer), the authors must plot the SYNSPEC reference trend alongside the CondGen trend in Figure 10. Relying solely on a qualitative description in the text without visual or quantitative comparison in the figure weakens the evidence for this limitation.

Round 2 - Failed

Paper version: v2

Submitted: October 05, 2026 at 15:00 UTC

Completed: October 05, 2026 at 15:05 UTC

Review status: failed

Round 2 - Approved

Paper version: v3

Submitted: October 05, 2026 at 15:06 UTC

Completed: October 05, 2026 at 15:09 UTC

Comment 1

Score: 4/4 - Fully addressed

Original comment:

On page 19, the text directly below Figure 8 (which is the 'Baseline Cr abundance sweep') describes the non-monotonic behavior of [Mg/H] and the Mg diagnostic: 'Absorption initially strengthens and the mean flux falls as [Mg/H] rises. At approximately +1.37 dex, however, the mean flux rises relative to the +0.67 dex case and the profiles become shallower. The bot-produced diagnostic therefore reveals a reversal of the earlier trend.' This text actually describes the Mg abundance sweep shown in Figure 10 on page 21, but its placement under Figure 8 on page 19 creates an internal contradiction and mismatch between the figure and the text immediately following it.

Author response:

Element-specific prose now immediately precedes each abundance-sweep figure. Cr text remains with the Cr figure only; all Mg non-monotonic discussion sits with the Mg figure (and the appendix abundance-diagnostics note). Image paths and science claims were unchanged (reorder/attach only). Abundance sweeps Sc-Mg are interleaved with matching prose in Results (float order typically PDF Figures 7?11), rather than stacked at section end.

Reviewer assessment:

The authors have successfully resolved the figure-text mismatch. In the revised manuscript, the text describing the non-monotonic Mg response has been moved to Section 6.3 and placed adjacent to Figure 11 (the Mg sweep figure), while the Cr text remains properly aligned with Figure 9 (the Cr sweep figure).

Comment 2

Score: 4/4 - Fully addressed

Original comment:

The comparison between the AI team's turnaround time of less than a week and the estimated human team's time of one academic semester is fundamentally flawed because it does not control for the starting state of the project. The AI team's work was built on the existing database and notebooks supplied by the authors. If a human student or researcher were provided with the exact same pre-existing database and fully functional notebooks, they could also easily run the code, generate the figures, and draft the manuscript in a few days. Comparing the AI team's execution of pre-existing workflows to a human team's typical semester-long research cycle (which usually involves database creation, code development, and troubleshooting from scratch) is highly misleading. To make a fair comparison, the authors should estimate the time a human would take given the same pre-existing assets, or temper the claim to avoid implying a massive speedup that is largely an artifact of the starting conditions.

Author response:

Timing language was tempered. "Less than a week" is scoped to AI-assisted analysis, diagnostics, figures, and drafting on a pre-existing SYNSPEC database, notebooks, and research program- not a from-scratch database or methods build. The Abstract and Discussion state that no comparison with a human team starting from the same resources was performed, so the elapsed-time account does not establish a productivity or labor saving. An explicit week-vs-semester comparison is no longer present.

Reviewer assessment:

The authors have fully addressed this concern by removing the misleading week-vs-semester comparison. They have added clear, prominent caveats in the Abstract, Section 7.1.1, and the Conclusion, explicitly stating that no direct comparison with a human team starting from the same resources was performed and that the elapsed-time account does not establish a productivity or labor saving.

Comment 3

Score: 4/4 - Fully addressed

Original comment:

The reported baseline MAE of 0.003681 is presented as a single global average across the entire test set of 21,657 spectra. However, stellar spectra vary dramatically in complexity across the parameter space; hot stars have very few absorption lines and are mostly flat continuum, whereas cool stars have extremely dense and complex line structures. A global average MAE can easily be skewed by the simpler spectra or by the large number of continuum pixels where the error is near zero, masking poor performance in line-rich regions or specific parameter regimes (such as cool stars, where the Mg limitation was noted). To provide a scientifically robust evaluation, the authors should analyze and report how the reconstruction error (MAE/RMSE) scales as a function of key physical parameters, particularly effective temperature and bulk abundance, rather than relying solely on a single global metric.

Author response:

Results now report verified full-TEST metrics as they appear in Main.tex:

? Global TEST MAE = 0.003681 for n = 21 657 spectra

? RMSE = 0.009794

? correlation = 0.9979

? R2 = 0.9958

Explicit prose states that this average does not describe every region of the label space equally. We cite the binned MAE figure (versus T_{eff} / $[\text{Fe}/\text{H}]$ / mean abundance). Example T_{eff} -bin MAEs from the text: ?0.00647 (4000-4750 K) and ?0.00254 (8500-9250 K).

Reviewer assessment:

The authors have thoroughly addressed this comment by adding a new figure (Figure 6) that presents binned MAE

summaries across key physical parameters: effective temperature (T_{eff}), iron abundance ($[\text{Fe}/\text{H}]$), and mean abundance. They also added corresponding discussion in Section 6.1 explaining how the reconstruction error scales across these different regimes.

Comment 4

Score: 4/4 - Fully addressed

Original comment:

In Table 4, the authors compare five architectures on a small subset of 256 spectra and report single-point metrics to claim that the baseline MLP outperforms the alternatives. However, because the evaluation is restricted to a very small subset (~1.2% of the test partition) and no standard deviations, confidence intervals, or statistical significance tests are provided, it is impossible to determine whether the reported differences in performance are statistically significant or merely artifacts of the specific random subset chosen. To support this claim, the authors should either evaluate the models on the full test set or provide error bars and statistical significance testing for the metrics on the subset.

Author response:

Table tab:arch comparison reports MAE with bootstrap 95% confidence intervals for the 256-spectrum architecture pilot (2000 resamples; MAX TEST = 256). Values as in current Main.tex:

? Baseline in the pilot: 0.003655 [0.003353, 0.003971]

? Wavelength attention: 0.009588 [0.008962, 0.010233]

? Wider MLP: 0.011624 [0.010877, 0.012434]

? Multi-scale Conv1D: 0.012988 [0.011955, 0.014072]

? FiLM residual: 0.018974 [0.016314, 0.022330]

Caption and nearby prose state that intervals are not used to assert a paired significance test or a general architecture ranking; no paired difference intervals and no full-TEST $\times$ 5 comparison of all five architectures are presented. The pilot decision to retain the baseline is described as exploratory model selection, not a statistically established ranking.

Reviewer assessment:

The authors have fully addressed this comment by updating Table 4 to include bootstrap 95% confidence intervals for all five compared architectures. They also added text in Section 6.4 and Appendix A.5 clarifying that these intervals represent exploratory pilot runs and are not used to assert a general architecture ranking.

Comment 5

Score: 3/4 - Well acknowledged

Original comment:

The manuscript presents a case study of a coordinated AI research team comprising nine specialized agents. However, it remains highly ambiguous whether there was any actual multi-agent system (e.g., an automated framework like AutoGen or CrewAI) executing these tasks, or if the coordination was entirely simulated by the human prompters manually copy-pasting messages between different chat threads of a single commercial LLM (Grok). The text mentions rooms and exchanges, but also admits that the available record does not include the complete instructions or configuration used to instantiate each role and does not establish whether the bots used different underlying models or performed their tasks simultaneously. If the team is simply a single human manually routing prompts to a single LLM under different persona instructions, then the claim of demonstrating a coordinated AI research team is fundamentally overstated. The authors should clarify the exact technical architecture of the agent system, specify whether message routing was automated or manual, and provide the system prompts or configurations used to instantiate the agents.

Author response:

We adopted an honest description: work in Grok Bot multi-agent chats and rooms; human-authorized Orchestrator; named specialist personas; artifact handoffs. We do not claim AutoGen, CrewAI, or similar concurrent autonomous frameworks (those names do not appear in the manuscript). Incomplete prompt/configuration records are acknowledged explicitly (team setup and topology / communication sections).

Reviewer assessment:

The authors have added clear, honest disclaimers in Section 2, Section 2.1, and Section 2.3 clarifying that the workflow relied on role-based assignments and handoffs in Grok Bot chats/rooms under human authorization, rather than an

automated concurrent multi-agent framework. They also explicitly acknowledge that the complete original prompts and configuration records are missing, which prevents full replication but accurately tempers the claims.

Comment 6

Score: 3/4 - Well acknowledged

Original comment:

In Figure 10, the abundance sweep for Mg shows an uptick in the mean flux at high abundances, which is claimed to represent an unphysical non-monotonic behavior. However, Figure 10 only plots the generated CondGen spectra and mean flux, completely omitting the corresponding SYNSPEC reference curves for the same parameter sweep. To scientifically demonstrate that this behavior is indeed a model limitation (i.e., a departure from physical radiative transfer), the authors must plot the SYNSPEC reference trend alongside the CondGen trend in Figure 10. Relying solely on a qualitative description in the text without visual or quantitative comparison in the figure weakens the evidence for this limitation.

Author response:

(Decision: no new Synspec runs.) At the held atmosphere, the database is reported to contain an exact SYNSPEC spectrum only for Mg = +0.67 dex. Additional synthesis for a matched sweep was not performed; no reference curves are shown. The figure remains CondGen-only; caption and Results/Appendix state that a matched SYNSPEC sweep is not available, so the plot identifies a potential problem for investigation rather than demonstrating a physical-model discrepancy. There is no Synspec overlay and no nearest-neighbor substitution as Synspec truth. The limitation is twofold: CondGen non-monotonic Mg response, plus missing Synspec Mg grid at the held ATM. Note: in the compiled PDF float order the Mg sweep is typically Figure 11; label references remain correct. Abundance Figures 7-11 are interleaved with matching element prose.

Reviewer assessment:

The authors did not perform new SYNSPEC simulations to plot the reference curves. However, they meaningfully revised the text in Section 6.3, Appendix A.4, and the caption of Figure 11 to explicitly state that a matched SYNSPEC sweep is unavailable, thereby clarifying that the plot identifies a potential network anomaly for future investigation rather than a proven physical discrepancy.
