AstroSpecLM: A Spectrum-Language Model for Evidence-Grounded Astronomical Spectral Analysis

Anonymous ACL submission

Abstract

Astronomical spectra encode rich physical information, but drawing scientific conclusions from spectral features typically requires expert interpretation. This paper presents AstroSpecLM, a spectrum-language model that connects one-dimensional DESI spectra with Qwen3-4B to answer questions and provide explanations grounded in spectral evidence. Instead of generating question-answer pairs directly from templates or raw catalog fields, we first distill each spectrum into a compact set of catalog- and spectrum-derived facts, then use these facts as references to generate instruction-following conversations. The resulting model is competitive with specialist supervised baselines on classification and redshift estimation, while additionally producing natural-language explanations that reference specific spectral features. Our results indicate that grounding a language model in one-dimensional scientific spectra is feasible, and that fact-mediated instruction data yields a model capable of both prediction and explanation.

1 Introduction

Astronomical spectra provide wavelength-resolved measurements that support the inference of object type, redshift, stellar temperature, and other physical properties. A typical astronomical spectrum consists of a one-dimensional (1D) flux array spanning thousands of wavelength pixels. Its interpretation hinges on linking localized spectral features, such as emission and absorption lines, and their wavelength shifts to the physical properties of the astronomical source. This process is inherently evidence-based: conclusions about an object's type or redshift should be supported by specific spectral features. Modern spectroscopic surveys such as DESI produce millions of such spectra (DESI Collaboration et al., 2024, 2026), and automated pipelines can efficiently extract structured measurements from them (Guy et al., 2023). However, the development of natural-language frameworks that explicitly map this spectral evidence to final conclusions lags behind. Ideally, such a model should process spectral inputs to generate interpretable predictions, accompanied by a clear rationale grounded in the underlying spectral features.

Large language models (LLMs) offer a path toward this form of grounded scientific interaction (Taylor et al., 2022; Zhang et al., 2024). When connected to modality encoders, they can condition generation on non-text inputs and produce flexible natural-language responses (Alayrac et al., 2022; Li et al., 2023). However, most multimodal language models are designed for natural images (Liu et al., 2023; Zhu et al., 2024), videos (Maaz et al., 2024; Lin et al., 2024), or documents (Hu et al., 2025), where the input is two-dimensional and semantically compositional. A 1D astronomical spectrum poses a different grounding problem: its meaning is tied to wavelength, and critical evidence often lies in narrow localized regions that only become meaningful through physical interpretation (Guy et al., 2023; DESI Collaboration et al., 2024; Zhong et al., 2024; Parker et al., 2024). Bridging this gap requires both a spectrum encoder that preserves wavelength-localized information and a training strategy that grounds language-model outputs in observable spectral evidence (He et al., 2022; Parker et al., 2024).

In this work, we present **AstroSpecLM**, a spectrum-language model that connects real 1D DESI spectra with Qwen3-4B (Yang et al., 2025) for spectrum-grounded question answering. AstroSpecLM uses a spectrum-specific masked autoencoder (He et al., 2022), **SpecMAE**, to encode the input spectrum, and a lightweight spectral projector to insert spectral representations into the language-model embedding space. To train the model faithfully, we construct **DESI-SpecInstruct**, an instruction-tuning dataset built through an intermediate fact extraction stage. The resulting model

answers spectrum-grounded questions in natural language with explanations that reference specific spectral features as evidence.

2 Related Work

Astronomical spectral analysis is an evidence-centered form of scientific interpretation: physical conclusions are supported by continuum shapes, emission and absorption features, spectral breaks, and wavelength shifts. Modern spectroscopic pipelines and expert systems have made this process increasingly scalable, producing calibrated spectra together with structured measurements and quality assessments (Guy et al., 2023; DESI Collaboration et al., 2026). Recent spectral models further move from task-specific prediction toward reusable representations, including image–spectrum contrastive models, stellar spectral foundation models, and native-resolution spectral encoders (Parker et al., 2024; Koblischke and Bovy, 2024; Parker et al., 2025; Islam and Fox, 2026). These systems are effective for processing spectra at scale, but their outputs — structured labels, numerical estimates, or visual interfaces — leave the evidence-to-conclusion chain implicit and dependent on expert interpretation.

LLMs have transformed the interface of AI systems from fixed-output prediction to open-ended instruction following and natural-language interaction (Ouyang et al., 2022; Chung et al., 2024; Grattafiori et al., 2024; Yang et al., 2025). Multimodal LLMs extend this capability to non-text inputs by connecting modality encoders with language models through adapters, projection layers, or cross-attention mechanisms (Liu et al., 2023; Li et al., 2024; Dai et al., 2024; Zhu et al., 2025; Bai et al., 2025). These advances suggest a general path for turning perceptual signals into language-conditioned systems. However, most general-purpose multimodal language models are optimized for natural images, videos, documents, or web-scale data, rather than real 1D scientific measurements whose interpretation depends on wavelength-localized physical evidence such as emission line profiles and continuum breaks.

Within astronomy, this language-centered paradigm has been adopted through domain-specific LLMs, astronomical vision-language models, and early spectroscopic LLM systems. Astronomy-oriented LLMs improve access to domain knowledge by continued pretraining or instruction tuning on astronomical literature and question-answer data (Nguyen et al., 2023; de Haan et al., 2025). Astronomical multimodal models further connect observations with language, including image–text retrieval, astronomical visual question answering, and radio-source analysis (Mishra-Sharma et al., 2024; Zaman et al., 2025; Riggi et al., 2025). More closely related to our setting, recent work has begun to bring language-model interfaces into spectroscopic workflows: Teaching LLMs to Speak Spectroscopy adapts LLaMA-3.1-8B with LoRA to predict galaxy redshifts from spectroscopic data while retaining part of its language ability (Ramachandra et al., 2025), and Spec-o3 develops a tool-augmented vision-language agent for rare-object vetting through spectral-image inspection and interleaved reasoning (Jia et al., 2026). These studies highlight the promise of bringing language-model interfaces to spectroscopic workflows, but they are still closer to limited quantitative prediction or tool-mediated visual inspection. AstroSpecLM bridges spectral expert systems and astronomy-oriented language models. It uses real observed 1D spectra as the direct evidence source, supporting unified classification, regression, spectral-feature analysis, and evidence-based explanation within a single generative framework.

3 Data Construction

A core challenge in spectrum-grounded question answering is encouraging generated answers to remain supported by observable spectral evidence. To this end, we design a pipeline that converts DESI spectra into language supervision through an intermediate *fact extraction* stage, producing the DESI-SpecInstruct dataset. Figure 1 illustrates the process: we first select high-quality spectra, distill each into a compact set of catalog- and spectrum-derived facts, and then use these facts as grounding references to generate instruction data. This design reduces the risk that generated answers introduce unsupported physical properties, while enabling rich, evidence-tied questions and explanations.

3.1 Spectrum Selection

We use 1D spectra from DESI DR1 (DESI Collaboration et al., 2026; Guy et al., 2023), defining two data sources for different training stages. The pretraining corpus for the spectral encoder, denoted as $\mathcal{D}_{\text{enc}}$, comprises 2.85 million spectra balanced across STAR, GALAXY, and QSO classes, and is

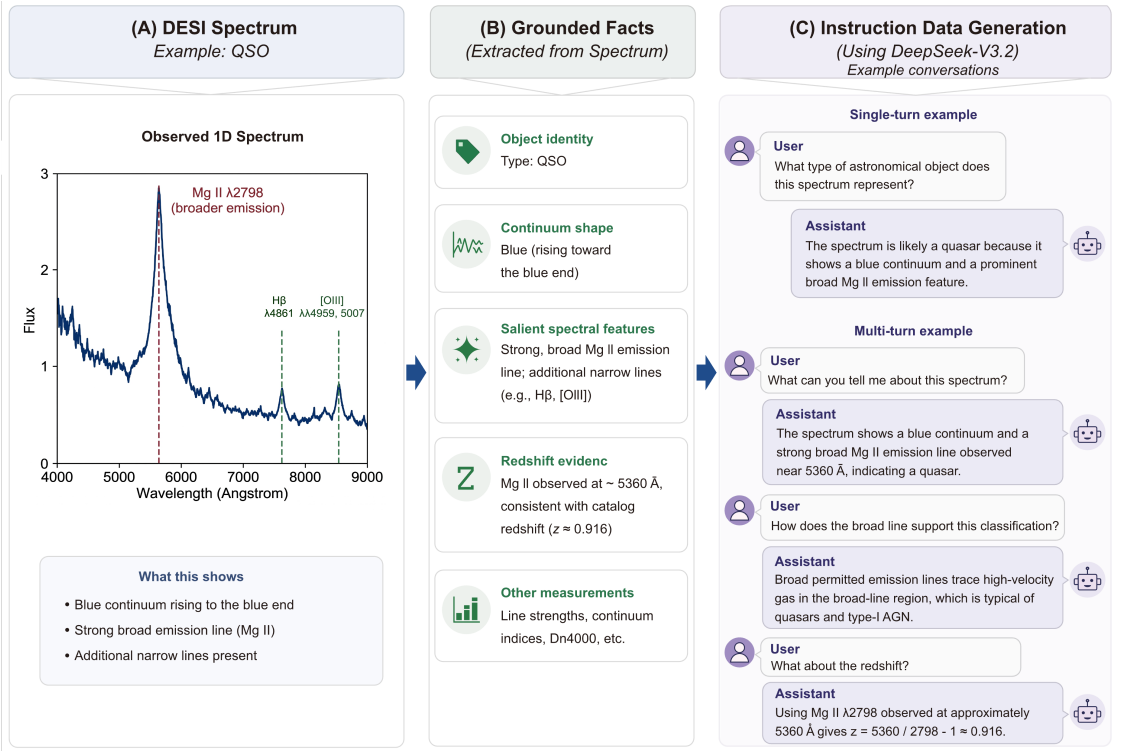

Figure 1: Instruction data construction pipeline. Raw DESI spectra are first distilled into grounded facts combining catalog annotations and spectral measurements; these facts then serve as references for generating instruction-following conversations.

used for masked reconstruction training. From the same underlying DESI collection, we further select a higher-quality subset $\mathcal{S}_{\text{inst}}$ of 58,558 spectra with reliable catalog annotations and clean spectral measurements for instruction-data construction. Detailed acquisition, preprocessing, and filtering criteria are provided in Appendix A.1.

3.2 Fact Extraction

For each spectrum in $\mathcal{S}_{\text{inst}}$, we construct a set of *grounded facts* that combine catalog labels (e.g., object class, redshift) with quantities directly measured from the spectrum (e.g., emission-line equivalent widths, continuum slopes). These facts are never shown to the model; they act solely as grounding references during instruction data generation, ensuring that every generated question and answer is anchored to spectral evidence. A full specification of fact types and measurement rules is given in Appendix A.2.

3.3 Instruction Data

Using the grounded facts as a conditioning signal, we employ DeepSeek-V3.2 (DeepSeek-AI, 2025) to generate user–assistant conversations, following the broader practice of constructing instruction data with LLMs (Wang et al., 2023; Ding et al., 2023). Both single-turn and multi-turn interactions are produced. To improve faithfulness to the input spectrum, the generation prompt explicitly forbids the model from introducing wavelengths, redshifts, or physical quantities not supported by the provided facts, thereby reducing the risk of hallucinated measurements. The full prompt is listed in Appendix A.3.

After automatic filtering for malformed conversations and leakage of reference wording, DESI-SpecInstruct contains 305,536 conversations (208,614 single-turn and 96,922 multi-turn).

3.4 Held-out Benchmark

We construct AstroSpecQA as a held-out benchmark from spectra disjoint from those used in DESI-SpecInstruct. It contains 11,698 questions across four spectrum-grounded tasks: object classification (5,000), redshift regression (3,247; galaxies and QSOs only), stellar subtype classification (1,753; stars only), and effective-temperature regression (1,698; stars only). This benchmark is used only for evaluation, and its label distributions are reported in Appendix A.4.

4 Method

AstroSpecLM is a spectrum-language model built on Qwen3-4B and a spectrum-specific masked autoencoder, SpecMAE. It connects 1D astronomical spectra to the language model by projecting spectral representations into the Qwen3-4B embedding space, and is trained through a two-stage procedure: spectrum masked autoencoding followed by spectral instruction tuning. Figure 2 illustrates the overall architecture.

4.1 AstroSpecLM

AstroSpecLM takes a normalized spectrum $\mathbf{X}_s \in \mathbb{R}^{7781}$ and a natural-language question $\mathbf{q}$ as input, and generates an answer $\mathbf{a}$ conditioned on both the spectral signal and the textual instruction. The spectrum is first encoded by a pretrained spectral encoder and then projected into the language-model embedding space.

Given a spectrum $\mathbf{X}_s$, the SpecMAE encoder $g_s(\cdot)$ produces a sequence of spectral patch representations:

$$\mathbf{Z}_s = g_s(\mathbf{X}_s) \in \mathbb{R}^{778\times 768}. \tag{1}$$

These features are passed to a lightweight spectral projector $P_\phi(\cdot)$. To reduce the spectral sequence length before insertion into the language model, we pad $\mathbf{Z}_s$ from 778 to 780 tokens and merge every six adjacent spectral tokens, producing 130 spectral tokens. The merged sequence is mapped to the hidden dimension of Qwen3-4B ($d_\ell = 2560$):

$$\mathbf{H}_s = P_\phi(\mathbf{Z}_s) \in \mathbb{R}^{130\times d_\ell}. \tag{2}$$

The projected spectral embeddings are inserted into the language-model input between two newly added special tokens, `<|spectrum_start|>` and `<|spectrum_end|>`. Let $E(\mathbf{q})$ denote the token embeddings of the question, and let $\mathbf{e}_{\text{start}}$ and $\mathbf{e}_{\text{end}}$ denote the embeddings of the two special tokens. The input embedding sequence is

$$\mathbf{E}_{\text{in}} = [\mathbf{e}_{\text{start}}, \mathbf{H}_s, \mathbf{e}_{\text{end}}, E(\mathbf{q})]. \tag{3}$$

The language model then generates the answer autoregressively:

$$p(\mathbf{a} \mid \mathbf{X}_s, \mathbf{q}) = \prod_{i=1}^{L} p_\theta(a_i \mid \mathbf{E}_{\text{in}}, a_{<i}), \tag{4}$$

where $\mathbf{a} = (a_1, \ldots, a_L)$ is the target answer and θ denotes the trainable parameters during spectral instruction tuning.

4.2 Spectrum Masked Autoencoding

We pretrain SpecMAE on $\mathcal{D}_{\text{enc}}$ (Section 3.1) using a masked reconstruction objective. Many diagnostic spectral features are spectrally localized; we therefore mask continuous spectral spans rather than isolated tokens, encouraging the encoder to infer missing regions from surrounding wavelength context. Specifically, for each spectrum, we sample 10 masked chunks, each containing 30 consecutive patch tokens.

The encoder first applies a 1D convolutional patch embedding with kernel size 20 and stride 10, producing overlapping spectral patches with boundary padding. This results in 778 patch tokens, which are processed by an 8-layer Transformer encoder (Vaswani et al., 2017) with hidden size 768, 8 attention heads, MLP ratio 4, and learned positional embeddings.

During pretraining, the encoder observes only the visible patch tokens, while a lightweight decoder reconstructs the masked flux values. The decoder is discarded after pretraining. The reconstruction loss is computed only on wavelength pixels covered by masked patches. Let f_i be the original flux at pixel i, $\hat{f}_i$ the reconstructed flux, $m_i \in \{0, 1\}$ the strict mask indicator, and $v_i \in \{0, 1\}$ a validity mask that excludes pixels with invalid inverse variance or known artifacts. We use the inverse variance ivar_i from the DESI pipeline (Guy et al., 2023) as a reconstruction weight, clipped at 1000 to reduce the influence of extreme weights:

$$w_i = \min(\text{ivar}_i, 1000). \tag{5}$$

The pretraining loss is

$$\mathcal{L}_{\text{MAE}} = \frac{\sum_i m_i v_i w_i \left(\hat{f}_i - f_i\right)^2}{\sum_i m_i v_i w_i}. \tag{6}$$

After pretraining, we retain only the encoder $g_s(\cdot)$ and freeze it in all subsequent training stages.

4.3 Spectral Instruction Tuning

We perform spectral instruction tuning on DESI-SpecInstruct (Section 3.3). Each training example consists of a spectrum $\mathbf{X}_s$, a user question $\mathbf{q}$, and a target assistant answer $\mathbf{a}$. During this stage, the SpecMAE encoder and the base parameters of Qwen3-4B are kept frozen. We train only the spectral projector P_ϕ, the embeddings of the newly added special tokens, and LoRA adapters (Hu et al.,

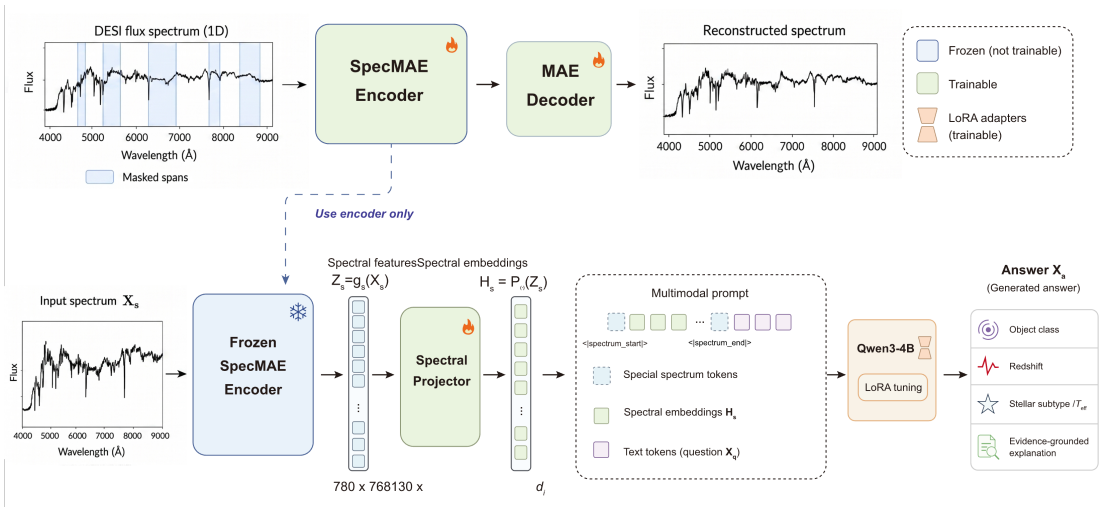

Figure 2: Overview of AstroSpecLM. A normalized DESI spectrum is encoded by the SpecMAE encoder, compressed by a spectral projector, and inserted into Qwen3-4B as continuous spectral embeddings. During spectral instruction tuning, the SpecMAE encoder is frozen, while the projector, special-token embeddings, and LoRA adapters are trained.

2022) inserted into the linear layers of Qwen3-4B. The LoRA rank is $r = 16$, with scaling factor $\alpha = 32$.

For each example, the SpecMAE encoder produces $\mathbf{Z}_s = g_s(\mathbf{X}_s)$. The projector maps $\mathbf{Z}_s$ to $\mathbf{H}_s$, which is inserted into the language-model input together with the question tokens as defined in Section 4.1. The model is optimized with the standard autoregressive language modeling objective, with loss applied only to answer tokens:

$$\mathcal{L}_{\text{sft}} = -\sum_{i=1}^{L} \log p_\theta(a_i \mid \mathbf{E}_{\text{in}}, a_{<i}). \qquad (7)$$

5 Experiments

5.1 Experimental Setup

Models and Training. We first pretrain SpecMAE on $\mathcal{D}_{\text{enc}}$ and then train AstroSpecLM on DESI-SpecInstruct through spectral instruction tuning. Both stages are conducted on 8 NVIDIA A100-40GB GPUs. SpecMAE pretraining takes approximately 24 hours, and AstroSpecLM instruction tuning takes approximately 3 hours. Detailed optimization hyperparameters for both stages are provided in Appendix B.1.

Baselines. We compare AstroSpecLM against three baselines. SpecMAE is evaluated by attaching task-specific MLP heads to the frozen spectral encoder used by AstroSpecLM. For AstroCLIP (Parker et al., 2024), we use its pretrained spectral encoder as a frozen feature extractor and train task-specific MLP heads on our training spectra for each classification or regression task. These supervised baselines are not designed for language tasks or free-text generation. Qwen3-4B (Yang et al., 2025) is included as a text-only reference model without additional training, and is evaluated only on the general language benchmarks to indicate the original language capability before spectral instruction tuning.

Evaluation. We evaluate AstroSpecLM on the AstroSpecQA benchmark described in Section 3.4. Specifically, we report classification accuracy for both object types and stellar subtypes, alongside Mean Absolute Error (MAE) and R^2 coefficients for redshift and effective temperature (T_{eff}) regressions. AstroSpecLM is prompted to first analyze the spectrum and then provide a concise answer after a "Final Answer" marker. To assess whether spectral instruction tuning preserves general language capability, we additionally evaluate on three standard benchmarks: MMLU-Redux (Gema et al., 2025) (general knowledge), IFEval (Zhou et al., 2023) (instruction following), and MATH-500 (Hendrycks et al., 2021; Lightman et al., 2024) (mathematical reasoning).

For classification tasks, we extract the predicted label from the final-answer field using deterministic rules. For numerical tasks, we extract scalar values, round them to at most four decimal places, and compute MAE and R^2 against the catalog tar-

Model	Spectrum-grounded Tasks				Language Benchmarks		
	Object Class Acc.	Stellar Subtype Acc.	Redshift MAE / R^2	T_{eff} MAE / R^2	MMLU-Redux Acc.	IFEval Acc.	MATH-500 Acc.
SpecMAE	98.12	85.62	0.029/**0.993**	**162.3/0.955**	N/A	N/A	N/A
AstroCLIP (Parker et al., 2024)	97.82	86.37	0.099/0.969	330.3/0.942	N/A	N/A	N/A
Qwen3-4B (Yang et al., 2025)	N/A	N/A	N/A	N/A	**75.82**	**69.30**	**83.60**
AstroSpecLM	**98.56**	**87.56**	**0.015**/0.992	335.2/0.668	74.00	65.53	80.20

Table 1: Main results on spectrum-grounded tasks and language benchmarks. Best results are shown in bold. N/A indicates tasks that are not applicable due to model input or output constraints.

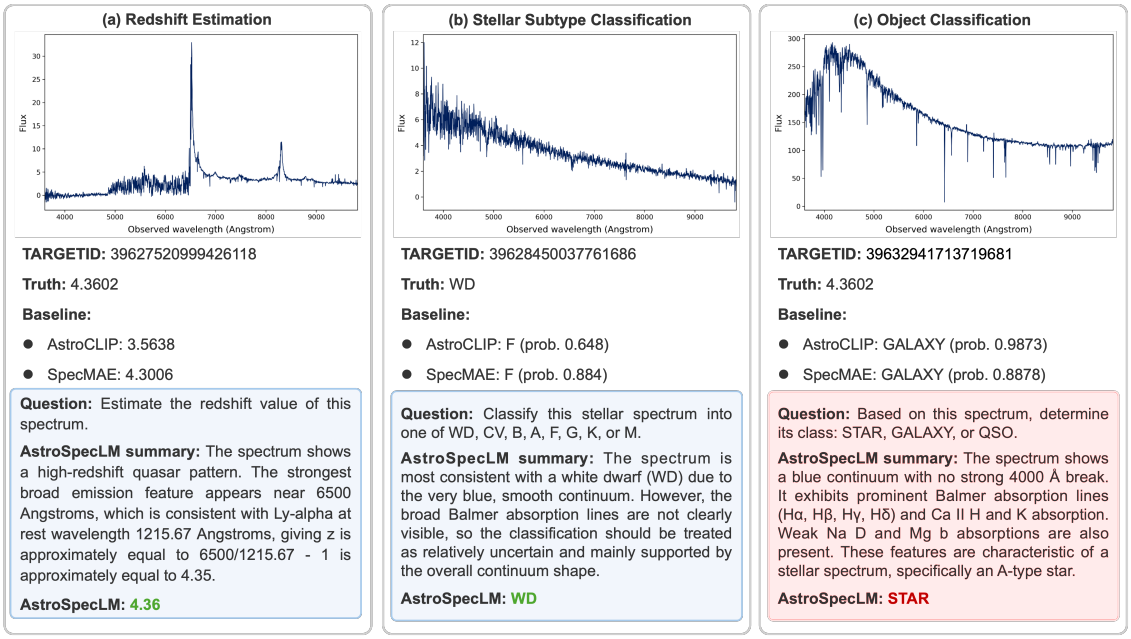

Figure 3: Qualitative examples comparing AstroSpecLM with SpecMAE and AstroCLIP. **Left:** redshift estimation for a QSO at $z = 4.3602$. **Center:** stellar subtype classification for a white dwarf. **Right:** a galaxy misclassified as a star by AstroSpecLM.

get. Outputs that cannot be parsed by rules are passed to DeepSeek-V3.2 for format normalization only, without access to ground-truth labels. We further evaluate explanation faithfulness on all AstroSpecQA outputs using DeepSeek-V3.2 as an LLM judge, which scores whether each explanation is supported by the corresponding held-out spectral facts. The full judging prompt and scoring rubric are provided in Appendix B.2. We also conduct a small human meta-evaluation on 100 randomly sampled outputs to assess the reliability of the automatic judge.

5.2 Main Results

Table 1 reports the results on spectrum-grounded tasks and general language benchmarks.

Spectrum-grounded tasks. AstroSpecLM achieves the highest accuracy on both classification tasks, surpassing the specialist supervised baselines on object type (98.56% vs. 98.12%) and stellar subtype (87.56% vs. 86.37%). This suggests that the unified generative model can remain competitive with dedicated classifiers while supporting free-form QA.

On redshift estimation, AstroSpecLM attains an MAE of 0.015, lower than SpecMAE's 0.029, while both models achieve comparable R^2 scores (0.992 vs. 0.993). This suggests that the generative model can effectively exploit line-position evidence when estimating redshift. Since redshift is closely tied to the observed-to-rest wavelength ratio of identifiable spectral lines, it is well aligned with the model's training objective of producing feature-grounded explanations.

Effective-temperature estimation is more challenging for the generative model. AstroSpecLM's MAE of 335.2 K and R^2 of 0.668 lag substantially behind SpecMAE's 162.3 K and 0.955. Un-

like redshift, effective temperature depends on distributed spectral cues such as continuum shape, line strengths, and molecular-band depths, and is less directly recoverable from a single localized measurement. Autoregressive generation may also introduce additional instability when producing continuous numerical values. Section 5.4 shows that alignment pretraining can substantially improve this task, suggesting that stronger spectral-language alignment is important for numerical parameter estimation.

Language benchmarks. AstroSpecLM retains much of the language capability of the text-only Qwen3-4B baseline after spectral instruction tuning, with MMLU-Redux decreasing from 75.82% to 74.00%, IFEval from 69.30% to 65.53%, and MATH-500 from 83.60% to 80.20%. This 1.8–3.8 point drop indicates moderate degradation rather than severe loss of general language ability, consistent with forgetting effects commonly observed in multimodal fine-tuning (Zhai et al., 2023; Zheng et al., 2024). Preserving this capability is important for AstroSpecLM, since the model relies on natural-language generation to express evidence-grounded explanations.

Explanation faithfulness. We further evaluate the faithfulness of AstroSpecLM's generated explanations on the full AstroSpecQA benchmark using DeepSeek-V3.2 as an LLM judge. AstroSpecLM obtains an average faithfulness score of 4.2/5.0, indicating that most explanations are well supported by the corresponding spectral facts. To calibrate the automatic judge, we manually score 100 randomly sampled outputs using the same rubric and find that the human assessments are generally consistent with the LLM-judge scores, supporting its use as a scalable faithfulness proxy.

5.3 Qualitative Analysis

To complement the quantitative results, we examine three representative cases that illustrate the strengths and limitations of AstroSpecLM relative to the specialist baselines. Figure 3 presents the spectra and model outputs for each case.

Case 1: Redshift estimation through line identification. The first example (Figure 3, left) shows a high-redshift quasar with true $z = 4.3602$. AstroSpecLM produces the closest estimate ($z = 4.36$) and explicitly identifies Lyα at ≈ 6500 Å and C IV at ≈ 8305 Å, computing the implied redshifts separately and verifying their mutual consistency. This illustrates how the generative model can outperform direct regression by explicitly modeling the relationship between observed line positions and the redshift formula.

Case 2: Rare subtype classification. The second example (Figure 3, center) shows a white dwarf (WD) spectrum that both SpecMAE and AstroCLIP misclassify as an F-type star. AstroSpecLM correctly identifies it as WD based on the blue continuum and absence of molecular bands, while explicitly noting that the usual broad Balmer absorption lines are not visible and that the classification is "relatively uncertain." This behavior—providing the correct answer while acknowledging evidence limitations—is a qualitative benefit of the language interface.

Case 3: Misclassification due to feature-level ambiguity. The third example (Figure 3, right) shows a GALAXY spectrum that both baselines classify correctly. AstroSpecLM misclassifies it as an A-type star, citing prominent Balmer absorption lines and a blue continuum with no strong 4000 Å break. These features are indeed standard indicators of an A-type stellar atmosphere, and their presence in this galaxy creates genuine ambiguity between an early-type stellar continuum and a galaxy dominated by young stellar populations. The error suggests that the model can over-rely on local absorption patterns when they closely match those of a different class.

5.4 Ablation Study

We conduct ablation experiments to validate the key design choices in AstroSpecLM, examining whether each component contributes to the model's ability to understand and reason about spectra. Table 2 summarizes the results.

Training strategy. We examine two variations of the training pipeline. Adding an intermediate alignment stage — where the projector and LoRA adapters are first trained on spectrum–caption pairs derived from the same grounded facts before instruction tuning — substantially improves performance on tasks requiring precise numerical reasoning, particularly effective temperature (R^2: 0.668 $\rightarrow$ 0.942). However, it also reduces MMLU-Redux from 74.00 to 70.40 and introduces an additional training stage. We therefore report the SFT-only model as our main configuration because

Model Variant	Object Class Acc.	Stellar Subtype Acc.	Redshift MAE / R^2	T_{eff} MAE / R^2	MMLU-Redux Acc.
Training strategy					
AstroSpecLM (4B, SFT)	<u>98.56</u>	**87.56**	<u>0.015</u> / **0.992**	335.2 / 0.668	**74.00**
+ Alignment pretraining	98.46	86.82	**0.011 / 0.992**	**206.2 / 0.942**	70.40
+ 2 training epochs	**98.78**	85.62	0.031 / <u>0.975</u>	350.6 / 0.712	72.10
Model scale & token length					
AstroSpecLM (1.7B, SFT)	88.18	71.03	0.090 / 0.847	374.6 / 0.437	64.40
AstroSpecLM (4B, SFT, 260 tok)	97.72	84.88	0.084 / 0.958	<u>285.4</u> / <u>0.912</u>	<u>73.30</u>
LoRA capacity					
AstroSpecLM (4B, SFT, rank=32)	98.12	<u>87.27</u>	0.018 / 0.963	327.7 / 0.712	73.10

Table 2: Ablation results. The first row (4B, SFT, 130 tok, rank=16) is our final configuration. Best results are in bold, second-best results are underlined.

it provides a simpler and more balanced trade-off between spectrum-grounded performance and retained language ability. The alignment variant represents a stronger option when numerical accuracy is prioritized. Details of the alignment data are provided in Appendix A.5. Extending instruction tuning to two epochs shows signs of overfitting: object classification improves slightly, but both regression tasks degrade, suggesting that longer instruction tuning does not consistently improve spectral grounding.

Model scale and token length. Reducing the language model from 4B to 1.7B parameters causes consistent degradation across all tasks, with the sharpest drop in T_{eff} (R^2: 0.668 $\rightarrow$ 0.437). This suggests that larger language-model capacity is beneficial for integrating spectral representations while preserving instruction-following ability. The spectral token count reflects a fundamental trade-off in how the model attends to the spectrum. With 260 tokens (twice the default), the model improves T_{eff} estimation (R^2: 0.668 $\rightarrow$ 0.912), likely because this task benefits from broader continuum shape and distributed line-ratio information. However, redshift accuracy decreases substantially (MAE: 0.015 $\rightarrow$ 0.084), suggesting a trade-off between broad spectral coverage and compact representations for localized line-based prediction. Since this mechanism is not directly verified by attribution analysis, we treat it as a hypothesis and retain 130 tokens as a practical default.

LoRA capacity. Increasing the LoRA rank from 16 to 32 produces only small changes across all tasks, confirming that rank-16 provides sufficient adaptation capacity for the model to learn meaningful spectral features without over-parameterization.

6 Conclusion

We presented AstroSpecLM, a spectrum-language model that connects 1D DESI spectra with Qwen3-4B for spectrum-grounded question answering. By constructing instruction data through an intermediate fact extraction stage, the model learns to produce answers grounded in observable spectral evidence. Experiments show that the generative approach is viable: AstroSpecLM matches or surpasses specialist supervised baselines on classification and redshift estimation, while effective-temperature regression remains a challenge for token-by-token decoding. In addition, the explanation-faithfulness evaluation indicates that its generated explanations are largely supported by the corresponding spectral facts. These results suggest that 1D scientific measurements can be connected to language models through fact-mediated instruction tuning. Future work may improve numerical grounding through stronger spectral-language alignment, finer wavelength-aware representations, and broader instruction data covering more physical parameters and scientific analysis scenarios.

Limitations

A central limitation of AstroSpecLM is that its main advantage lies in evidence-grounded explanation rather than uniformly superior numerical prediction. While it matches or surpasses specialist baselines on classification and redshift estimation, it remains weaker on effective-temperature regression, where dedicated supervised models are more suitable for applications requiring only precise numerical estimates. The generative formulation also introduces additional computational cost and possi-

ble decoding instability for continuous values.

Although we evaluate explanation faithfulness on the full AstroSpecQA benchmark using an LLM-as-a-judge protocol and a small human meta-evaluation, this does not replace large-scale expert astronomical assessment. Generated explanations may still contain unsupported or overconfident statements, especially for noisy spectra, rare subtypes, or physically ambiguous cases. In addition, AstroSpecLM is trained and evaluated only on DESI spectra, so its performance may degrade on spectra from instruments with different wavelength coverage, resolution, calibration, or noise characteristics.

Ethical Considerations

Our work uses publicly available DESI DR1 spectra and does not involve human subjects, personally identifiable information, or private data. The instruction dataset DESI-SpecInstruct is generated by DeepSeek-V3.2 conditioned on extracted spectral facts, and we performed automatic filtering and manual spot-checking to reduce factual errors. However, AstroSpecLM is a research prototype and its explanations are generated from learned representations and should not be treated as guaranteed scientific interpretations. We therefore recommend that model outputs be treated as assistive references rather than definitive scientific conclusions. We provide anonymized code and derived datasets in the supplementary materials to support reproducibility during review. Upon publication, we will make these resources publicly available, subject to the licenses and terms of the underlying resources.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. 2022. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 23716–23736.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. 2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Wenliang Dai, Nayeon Lee, Boxin Wang, Zhuolin Yang, Zihan Liu, Jon Barker, Tuomas Rintamaki, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2024. Nvlm: Open frontier-class multimodal llms. *arXiv preprint arXiv:2409.11402*.

Tijmen de Haan, Yuan-Sen Ting, Tirthankar Ghosal, Tuan Dung Nguyen, Alberto Accomazzi, Azton Wells, Nesar Ramachandra, Rui Pan, and Zechang Sun. 2025. Achieving gpt-4o level performance in astronomy with a specialized 8b-parameter large language model. *Scientific Reports*.

DeepSeek-AI. 2025. DeepSeek-V3.2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*.

DESI Collaboration, M. Abdul Karim, A. G. Adame, D. Aguado, J. Aguilar, et al. 2026. Data release 1 of the dark energy spectroscopic instrument. *The Astronomical Journal*, 171(5):285.

DESI Collaboration, A. G. Adame, J. Aguilar, S. Ahlen, S. Alam, G. Aldering, D. M. Alexander, R. Alfarsy, C. Allende Prieto, et al. 2024. The early data release of the dark energy spectroscopic instrument. *The Astronomical Journal*, 168(2):58.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Enhancing chat language models by scaling high-quality instructional conversations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3029–3051, Singapore. Association for Computational Linguistics.

Aryo Pradipta Gema, Joshua Ong Jun Leang, Giwon Hong, Alessio Devoto, Alberto Carlo Maria Mancino, Rohit Saxena, Xuanli He, Yu Zhao, Xiaotang Du, Mohammad Reza Ghasemi Madani, Claire Barale, Robert McHardy, Joshua Harris, Jean Kaddour, Emile van Krieken, and Pasquale Minervini. 2025. Are we done with mmlu? In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 5069–5096. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

J. Guy, S. Bailey, A. Kremin, S. Alam, D. M. Alexander, et al. 2023. The spectroscopic data processing pipeline for the dark energy spectroscopic instrument. *The Astronomical Journal*, 165(4):144.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16000–16009.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the MATH dataset. In *Advances in Neural Information Processing Systems*, volume 34, pages 7949–7962.

Anwen Hu, Haiyang Xu, Liang Zhang, Jiabo Ye, Ming Yan, Ji Zhang, Qin Jin, Fei Huang, and Jingren Zhou. 2025. mPLUG-DocOwl2: High-resolution compressing for OCR-free multi-page document understanding. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5817–5834, Vienna, Austria. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Md Khairul Islam and Judy Fox. 2026. Omnispectra: A unified foundation model for native resolution astronomical spectra. *arXiv preprint arXiv:2601.15351*.

Minghui Jia, Qichao Zhang, Ali Luo, Linjing Li, Shuo Ye, Hailing Lu, Wen Hou, and Dongbin Zhao. 2026. Spec-o3: A tool-augmented vision-language agent for rare celestial object candidate vetting via automated spectral inspection.

Nolan Koblischke and Jo Bovy. 2024. Spectrafm: Tuning into stellar foundation models. *arXiv preprint arXiv:2411.04750*.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let's verify step by step. In *International Conference on Learning Representations*.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. 2024. Video-LLaVA: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5971–5984, Miami, Florida, USA. Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. 2024. Video-ChatGPT: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12585–12602, Bangkok, Thailand. Association for Computational Linguistics.

Siddharth Mishra-Sharma, Yiding Song, and Jesse Thaler. 2024. Paperclip: Associating astronomical observations and natural language with multi-modal models.

Tuan Dung Nguyen, Yuan-Sen Ting, Ioana Ciuca, Charles O'Neill, Ze-Chang Sun, Maja Jabłońska, Sandor Kruk, Ernest Perkowski, Jack Miller, Jason Jason Jingsh Li, Josh Peek, Kartheik Iyer, Tomasz Rozanski, Pranav Khetarpal, Sharaf Zaman, David Brodrick, Sergio J. Rodriguez Mendez, Thang Bui, Alyssa Goodman, Alberto Accomazzi, Jill Naiman, Jesse Cranney, Kevin Schawinski, and Roberta Raileanu. 2023. AstroLLaMA: Towards specialized foundation models in astronomy. In *Proceedings of the Second Workshop on Information Extraction from Scientific Publications*, pages 49–55, Bali, Indonesia. Association for Computational Linguistics.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744.

Liam Parker, Francois Lanusse, Siavash Golkar, Alberto Bietti, Miles Cranmer, Michael Eickenberg, Geraud Krawezik, Michael McCabe, Ruben Ohana, Mariel Pettee, Bruno Regaldo-Saint Blancard, Kyunghyun Cho, and Shirley Ho. 2024. Astroclip: A cross-modal foundation model for galaxies. *Monthly Notices of the Royal Astronomical Society*, 531(4):4990–5011.

Liam Parker, Francois Lanusse, Jeff Shen, Ollie Liu, Tom Hehir, Leopoldo Sarra, Lucas Meyer, Micah Bowles, Sebastian Wagner-Carena, Helen Qu, Siavash Golkar, Alberto Bietti, Hatim Bourfoune, Pierre Cornette, Keiya Hirashima, Geraud Krawezik, Ruben Ohana, Nicholas Lourie, Michael McCabe, Rudy Morel, Payel Mukhopadhyay, Mariel Pettee, Kyunghyun Cho, Miles Cranmer, and Shirley Ho. 2025. Aion-1: Omnimodal foundation model for astronomical sciences. In *Advances in Neural Information Processing Systems*, volume 38, pages 95386–95428. Curran Associates, Inc.

Nesar Ramachandra, Yuan-Sen Ting, Zechang Sun, Azton Wells, and Salman Habib. 2025. Teaching llms to speak spectroscopy. *arXiv preprint arXiv:2508.10075*.

S. Riggi, T. Cecconello, A. Pilzer, S. Palazzo, N. Gupta, A. M. Hopkins, C. Trigilio, and G. Umana. 2025. Evaluating small vision-language models as ai assistants for radio astronomical source analysis tasks. *arXiv preprint arXiv:2503.23859*.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report.

Sharaf Zaman, Michael J. Smith, Pranav Khetarpal, Rishabh Chakrabarty, Michele Ginolfi, Marc Huertas-Company, Maja Jabłońska, Sandor Kruk, Matthieu Le Lain, Sergio José Rodríguez Méndez, and Dimitrios Tanoglidis. 2025. Astrollava: Towards the unification of astronomical data and natural language.

Yuexiang Zhai, Shengbang Tong, Xiao Li, Mu Cai, Qing Qu, Yong Jae Lee, and Yi Ma. 2023. Investigating the catastrophic forgetting in multimodal large language models. *arXiv preprint arXiv:2309.10313*.

Yu Zhang, Xiusi Chen, Bowen Jin, Sheng Wang, Shuiwang Ji, Wei Wang, and Jiawei Han. 2024. A comprehensive survey of scientific large language models and their applications in scientific discovery. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 8783–8817. Association for Computational Linguistics.

Junhao Zheng, Qianli Ma, Zhen Liu, Binquan Wu, and Huawen Feng. 2024. Beyond anti-forgetting: Multimodal continual instruction tuning with positive forward transfer. *arXiv preprint arXiv:2401.09181*.

Fucheng Zhong, Nicola R. Napolitano, Caroline Heneka, Rui Li, Franz Erik Bauer, et al. 2024. Galaxy spectra neural network (GaSNet). II. using deep learning for spectral classification and redshift predictions. *Monthly Notices of the Royal Astronomical Society*, 532(1):643–665.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2024. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. In *International Conference on Learning Representations*.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, et al. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

A Data Details

A.1 Spectrum Corpus and Preprocessing

Background on DESI. The Dark Energy Spectroscopic Instrument (DESI) is a large-scale multi-object spectroscopic survey designed to obtain optical spectra for millions of astronomical sources, including galaxies, quasars, and stars. DESI operates on the Mayall 4-meter telescope and uses thousands of fibers to collect spectra over a wide field of view, with its spectrographs covering the optical wavelength range of approximately 3600–9800 Å. Its public data releases provide calibrated 1D spectra together with pipeline-derived annotations such as object class, redshift, and quality indicators (DESI Collaboration et al., 2026). These properties make DESI particularly suitable for our setting: the spectra are large-scale, homogeneous, and physically interpretable, while the accompanying catalog information enables the construction of grounded supervision for spectrum-language modeling.

Fact Type	Fields	Source
Object labels	spectral type, subtype, redshift	catalog
Continuum	slope label, red-to-blue ratio	measured
D_n4000	value, strength label	measured
Spectral lines	rest/observed wavelength, EW, significance, status	measured
Molecular bands	CaH, TiO, VO indices (K/M stars)	measured
Stellar parameters	effective temperature (when available)	catalog

Table 3: Structured fact types used as grounding references.

Object Class	Representative Spectral Features
GALAXY	[O II], [O III], Hα, Hβ, [N II], [S II], Ca H/K, Mg b, Na D
QSO	Lyα, N V, Si IV+O IV], C IV, C III], Mg II, Hα, Hβ, [O III]
STAR	Ca H/K, Balmer lines, Mg b, Na D, Hα, Ca II triplet
K/M STAR	CaH, TiO molecular bands

Table 4: Class-specific spectral lines and bands used in fact extraction.

We use DESI DR1 1D spectra. Each spectrum provides a flux sequence and an inverse-variance sequence defined on a common wavelength grid of 7781 pixels (3600–9824 Å, step 0.8 Å). For encoder pretraining, we use 2,853,354 spectra balanced across STAR (951,118), GALAXY (951,118), and QSO (951,118), split 95/5 into training and validation. All pretraining spectra satisfy $\texttt{ZWARN} = 0$, finite redshift uncertainty, and SNR > 2.

For instruction-data construction, we sample a higher-quality subset of 58,558 spectra from the same collection, requiring $\texttt{ZWARN} = 0$, $\texttt{COADD_FIBERSTATUS} = 0$, finite $\texttt{ZERR}$ and $\texttt{DELTACHI2}$, $0 \leq \texttt{ZERR} < 0.005$, $\texttt{DELTACHI2} > 30$, and predominantly SNR > 5 (with a small fraction of lower-SNR spectra retained for robustness).

Each flux sequence is normalized to zero mean and unit variance before being fed to the spectrum encoder. The inverse variance is used only for weighting the masked reconstruction loss and for quality filtering, not as encoder input.

Data license. DESI DR1 is publicly available under the Creative Commons Attribution 4.0 International License. We use the spectra and catalog annotations in accordance with the DESI data-access terms and citation requirements.

A.2 Grounded Fact Extraction

For each spectrum selected for language-data construction, we extract a compact set of *grounded facts* that combine catalog annotations with quantities measured directly from the spectrum. Table 3 summarizes the fact types.

Spectral lines are measured in the rest frame using local continuum subtraction. For each predefined line, we integrate positive (emission) or negative (absorption) residuals within a line-centered window, using continuum sidebands for local baseline estimation. Lines are marked invalid if they fall in regions of insufficient wavelength coverage, have too few valid pixels, or yield physically implausible equivalent widths. The set of targeted lines depends on the object type; Table 4 lists the representative features used for each class. K and M stars additionally include molecular-band indices computed from flux ratios between feature and continuum bands.

Object Class	Count	Ratio
STAR	1,753	35.06%
GALAXY	1,648	32.96%
QSO	1,599	31.98%

Table 5: Object-class distribution in the AstroSpecQA object-classification split.

Stellar Subtype	Count	Ratio
K	620	35.37%
G	386	22.02%
M	322	18.37%
F	274	15.63%
WD	94	5.36%
A	53	3.02%
B	3	0.17%
CV	1	0.06%

Table 6: Stellar-subtype distribution in the AstroSpecQA stellar-subtype split.

A.3 Instruction Generation

We use DeepSeek-V3.2 to generate instruction-following conversations conditioned on the grounded facts. The full generation prompt is shown in Figure 4.

Act like an expert with extensive experience writing in the field of astrophysics.

Objective:
You are an AI data generator for instruction tuning of a spectral encoder + language model.
The final model will NOT see the text description below. During training, spectrum placeholders are inserted by the data collator, so the JSON output must contain natural user questions only. Do not include the literal token `<spectrum>`.
You are given a hidden reference description of one astronomical spectrum. Use it only as grounding information to create instruction conversations.
Your task is to generate instruction-tuning conversations for a model that answers questions about the spectrum.

Important rules:
1. Do NOT mention the hidden reference description, summary, JSON, fields, annotations, facts, labels, or provided text.
2. The user questions must sound like natural questions asked while viewing a spectrum.
3. The assistant answers must sound as if the model is analyzing the spectrum directly.
4. Use only information supported by the hidden reference description.
5. Do not invent spectral lines, parameters, object types, wavelengths, or physical interpretations not present in the hidden reference.
6. It is acceptable to use approximate physical language such as "around 3800 K" or "consistent with an M-type dwarf".
7. Avoid questions such as "What does the summary say?", "According to the provided data...", and questions about catalog identifiers, dataset metadata, or annotation fields.
8. Prefer questions such as "What type of object does this spectrum suggest?", "What features dominate this spectrum?", "What can be inferred from the continuum shape?", and "Why does this look like an M-type star?"
9. Generate samples that may include either or both of the following types:
   - single-turn simple QA samples or reasoning QA samples
   - multi-turn conversational samples with follow-up questions that dive deeper into the spectrum

Output format:
The output format should follow this style:

```
[
  {
    "type": "single_turn",
    "messages": [
      {"role": "user", "content": ""},
      {"role": "assistant", "content": ""}
    ]
  },
  {
    "type": "multi_turn",
    "messages": [
      {"role": "user", "content": ""},
      {"role": "assistant", "content": ""},
      {"role": "user", "content": ""},
      {"role": "assistant", "content": ""}
    ]
  }
]
```

Multi-turn samples can contain user/assistant conversations of any number of turns.
Generate diverse questions covering:
- object type or spectral subtype
- redshift
- continuum shape
- dominant spectral features
- molecular absorption bands
- stellar temperature
- physical interpretation of the spectrum
- and other spectrum-related questions that are supported by the hidden reference

Return valid JSON only. Do not include markdown. Return a JSON array, not an object.

Figure 4: Full prompt used for instruction data generation.

Hyperparameter	SpecMAE	AstroSpecLM
Training data	$\mathcal{D}_{\text{enc}}$	DESI-SpecInstruct
Epochs	100	1
Batch size	360	32
Optimizer	AdamW	AdamW
Learning rate	1×10^{-4}	2×10^{-5}
Weight decay	0.001	0.01
Scheduler	Cosine decay	Cosine decay
Warmup	5 epochs	100 steps
Precision	bf16	bf16
Training time	∼24 hours	∼3 hours

Table 7: Training hyperparameters for SpecMAE pretraining and AstroSpecLM spectral instruction tuning.

A.4 AstroSpecQA Benchmark Details

Table 5 reports the object-class distribution for the object-classification split. The three major classes are approximately balanced, reducing the risk that object-level accuracy is dominated by a single class. Table 6 reports the stellar-subtype distribution. This split is naturally imbalanced: common stellar types such as K, G, M, and F dominate the evaluation set, while rare categories such as B stars and cataclysmic variables contain very few examples.

A.5 Alignment Data

For the alignment pretraining experiment (Section 5.4), we construct 474,051 spectrum–caption pairs from the same grounded facts used for instruction data generation. Each caption is a concise natural-language description of the spectrum generated by DeepSeek-V3.2, constrained to use only information present in the fact structure. The full generation prompt is shown in Figure 5.

B Additional Experimental Details

B.1 Training Hyperparameters

Table 7 summarizes the main optimization hyperparameters used for SpecMAE pretraining and AstroSpecLM spectral instruction tuning.

B.2 Explanation Faithfulness Evaluation

We use DeepSeek-V3.2 as an automatic judge to evaluate explanation faithfulness. The judge receives the generated explanation and the corresponding held-out spectral facts, but not the ground-truth final answer label. It assigns a score from 1 to 5, where 1 indicates an unsupported or hallucinated explanation and 5 indicates that the explanation is fully supported by the spectral facts. The full judging prompt is shown in Figure 6.

You are an experienced astronomical spectroscopist. Your task is to write a natural and objective textual description based on the provided structured 1D spectral feature report. This description will be used to align a spectral encoder with a language model. Therefore, your goal is not to generate a measurement table or a lengthy physical interpretation, but to describe the observed spectral features in natural language while preserving key information useful for spectrum-language alignment, such as the object type, continuum morphology, main spectral lines, approximate observed positions of important lines, absorption or emission nature, relative strength, and reliable physical parameter context.

General style requirements:

- Do not mention JSON, field names, structured reports, confidence labels, measurement tables, or phrases such as "the report shows", "the data indicate", "the input information shows", or "the feature analysis shows".
- The description should sound like a natural observational description written by an astronomical spectroscopist, rather than a sentence-by-sentence conversion of a measurement table. Never use "identifies", "identified as", "classifies", or similar classifier-like wording.
- For stellar spectra, prefer describing radial velocity or the blueshift/redshift of stellar lines, rather than treating redshift as the central quantity.
- The continuum should only be described as an observed spectral morphology. Describe its color, slope, or overall shape, such as blue, red, flat, or smoothly varying, only when the continuum information is reliable.
- For spectral features that are explicitly indicated as significant, strong, or visible, you may use modifiers such as clear, strong, prominent, or visible. You may naturally include their observed wavelengths, absorption or emission nature, and approximate equivalent widths. However, do not mechanically list detailed information for every spectral line.
- For weak or uncertain spectral features, use cautious wording such as weak, faint, shallow, subtle, or tentative.
- Mention the 4000 Å break only when it is genuinely useful for describing the object type and spectral appearance. For QSOs, do not emphasize the 4000 Å break unless it is clearly important.
- When physical parameters are reliable, you may naturally incorporate T_{eff}, $\log g$, [Fe/H], redshift, or radial velocity, but do not mechanically list all available parameters.
- Avoid excessive physical interpretation or over-explaining the physical mechanisms.
- Ensure that the description is scientifically accurate and suitable for spectrum-language model alignment.

Figure 5: Full prompt used to generate spectrum–caption pairs for alignment pretraining.

You are an expert astrophysicist and a rigorous data validation specialist. Your task is to evaluate whether a spectrum-language model's generated natural-language explanation is faithful to the reference spectral facts and whether it contains any hallucinations.

You will be provided with:
- **Reference Facts:** catalog-derived physical parameters and measured spectral features used as the evaluation reference.
- **Question:** the user's query or instruction.
- **Model Answer:** the generated explanation to be evaluated.

Please objectively evaluate the **Model Answer** strictly based on the **Reference Facts**. Do not use external knowledge that conflicts with the provided facts.

Evaluation Dimensions:
- **Hallucination:** Determine whether the model fabricates features not present in the Reference Facts (e.g., inventing non-existent emission or absorption lines) or generates statements that contradict the facts (e.g., stating an incorrect redshift or object class).
- **Faithfulness Score (1–5 Scale):** Assess how well the model's answer aligns with the given facts:
  - **1 (Severe):** Completely contradictory or filled with severe misleading hallucinations.
  - **2 (Poor):** Partially correct, but contains major factual errors or critical fabrications.
  - **3 (Fair):** Mostly correct, but includes minor numerical deviations or trivial fabricated details that do not alter the main physical interpretation.
  - **4 (Good):** Highly accurate and faithful. No obvious errors, though it may omit some minor facts.
  - **5 (Excellent):** Perfectly accurate. Flawlessly covers the key facts without any hallucinations or contradictions.

Output Format:
You must output your evaluation strictly in JSON format with the following keys:

```
{
  "reasoning": "...",
  "has_hallucination": true or false,
  "hallucination_type": "None" or "Contradiction" or "Fabrication",
  "faithfulness_score": 1--5
}
```

Figure 6: Full prompt used for the LLM-as-a-judge faithfulness evaluation.