MeasHalu: Mitigation of Scientific Measurement Hallucinations for Large Language Models with Enhanced Reasoning

Anonymous ACL submission

Abstract

The accurate extraction of scientific measurements from literature is a critical yet challenging task in AI4Science, enabling large-scale analysis and integration of quantitative research findings. However, Large Language Models (LLMs) frequently exhibit severe hallucinations, which significantly undermine the reliability of automated scientific document understanding systems. To address this problem, we propose **MEASHALU**, a novel framework for mitigating scientific measurement hallucinations through enhanced reasoning and targeted optimization. We first present a fine-grained taxonomy of measurement-specific hallucinations, categorizing errors across quantities, units, modifiers, and relations. Our approach incorporates a two-stage reasoning-aware fine-tuning strategy using augmented scientific data and process-based supervision. Furthermore, we introduce a progressive reward curriculum designed to penalize specific hallucination types, significantly improving extraction faithfulness. Experimental results demonstrate that MEASHALU substantially reduces hallucination rates and improves overall accuracy on the MeasEval benchmark. This work provides a targeted solution to a key bottleneck in automated scientific knowledge extraction, facilitating more trustworthy and scalable machine-assisted scientific literature analysis.

1 Introduction

The rapid expansion of scientific literature has created an unprecedented demand for reliable automatic extraction of quantitative knowledge, which lies at the core of modern AI4Science applications such as large-scale meta-analysis, knowledge base construction, and autonomous scientific discovery (Hanson et al., 2024; Chen et al., 2025). Central to this process is *scientific measurement extraction*—the task of identifying numerical quantities, their units, modifiers, and their relationships

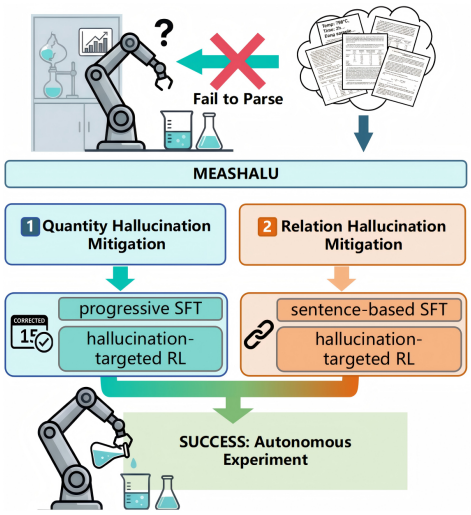

Figure 1: **Motivation of MeasHalu.** To rectify parsing failures, we propose a taxonomy-based approach to mitigate quantity and relation hallucinations.

to measured entities and properties. These quantitative statements form the evidential backbone of experimental sciences across disciplines ranging from materials science to biomedical research (Berrahou et al., 2013; Kononova et al., 2021). Although recent Large Language Models (LLMs) have demonstrated remarkable generalization abilities, they continue to perform unreliably on this task (Foppiano et al., 2024): even minor hallucinations in quantities or relations can invalidate entire experimental conclusions, severely limiting the trustworthiness of LLM-driven scientific understanding systems.

A key challenge underlying this failure is that *measurement hallucinations differ fundamentally from general textual hallucinations*. Unlike open-domain factual errors, measurement hallucinations exhibit fine-grained structural failures: models fabricate nonexistent values, misassociate quantities with wrong entities, overlook crucial qualifiers, or distort relations between scientific variables (Saier et al., 2024). Existing hallucination mitigation techniques, such as retrieval augmentation (Lewis et al.,

2020), generic instruction tuning, or conversational verification (Polak and Morgan, 2024), remain insufficient, as they are not designed to enforce the strict grounding and structural consistency required by scientific measurements. Yet, despite the importance of this problem, current research lacks both a systematic analysis of measurement-specific hallucination phenomena and targeted learning mechanisms for their mitigation. For instance, even state-of-the-art LLM-based extraction systems often compromise faithfulness by generating implicit information that is absent from the original text, such as inferring chemical formulas (Dagdelen et al., 2024).

In this work, we present MEASHALU, a reasoning-enhanced framework that explicitly models and suppresses scientific measurement hallucinations in LLMs. Our central insight is that hallucinations in this domain arise from two intertwined sources: (1) unreliable quantitative reasoning that corrupts individual quantities and units, and (2) fragile long-range relational reasoning that breaks the alignment between quantities, entities, and scientific properties. MEASHALU addresses these failure modes through a unified learning pipeline that combines reasoning-aware supervised fine-tuning with targeted reinforcement learning via structured reward shaping, thereby internalizing scientific grounding constraints directly into model parameters.

Concretely, MEASHALU introduces a fine-grained taxonomy of measurement hallucinations, and leverages this analysis to design a progressive optimization strategy: an initial supervised stage that standardizes quantitative reasoning and extraction structure, followed by Group Relative Policy Optimization (GRPO) with carefully constructed rewards that penalize fabrication, out-of-scope predictions, misclassification, and relational incompleteness. Our framework is developed on top of the MeasEval annotation schema (Harper et al., 2021a) and integrates external quantity validators, including CQE (Almasian et al., 2023a) and Quantulum[1], during training. Extensive experiments on the MeasEval benchmark and our newly constructed MeasEval-Ext dataset demonstrate that MEASHALU substantially reduces hallucination rates and consistently outperforms strong supervised baselines and proprietary LLMs. Furthermore, we show that MEASHALU functions as a reliable external measurement extraction tool that significantly improves performance on downstream embodied scientific tasks, validating its practical utility for trustworthy AI4Science systems.

Our contributions are summarized as follows:

We provide the first fine-grained analysis of scientific *measurement hallucinations* in large language models, revealing their structural nature and identifying two fundamental sources of failure: unreliable quantitative reasoning and fragile relational grounding.

We propose MEASHALU, a unified reasoning-enhanced learning framework that systematically suppresses measurement hallucinations by integrating reasoning-aware supervised fine-tuning with targeted reinforcement learning via structured reward shaping.

We construct a new out-of-distribution evaluation benchmark, MEASEVAL-EXT, and demonstrate through extensive experiments that MEASHALU substantially reduces hallucination rates and consistently outperforms strong supervised baselines and proprietary LLMs on scientific measurement extraction.

We further show that MEASHALU serves as a reliable external measurement extraction tool that significantly improves performance on downstream embodied scientific tasks, validating its practical utility for trustworthy AI4Science systems.

2 Related Work

2.1 Hallucinations in Large Language Models

Hallucination, where language models generate ungrounded or factually incorrect content, has been extensively studied in general-purpose LLMs (Huang et al., 2025). Most prior work focuses on semantic and factual hallucinations in open-ended generation (Ji et al., 2023), with typical taxonomies including fabrication, inconsistency, and logical errors (Li et al., 2025). However, these taxonomies are largely developed for free-form text generation and do not capture the structural requirements of measurement extraction, where numerical faithfulness, unit consistency, and entity-quantity relational grounding are essential. We address this gap by proposing a fine-grained taxonomy of *measurement-specific hallucinations* and designing mitigation mechanisms tailored to these failure modes.

[1] https://github.com/nielstron/quantulum3

2.2 General Information Extraction vs. Measurement Extraction

Information extraction (IE) and named entity recognition (NER) are foundational NLP tasks (Nadeau and Sekine, 2007). While early systems relied on rule-based and feature-engineered pipelines, modern approaches increasingly leverage neural architectures and pre-trained language models. Nevertheless, *scientific measurement extraction* poses additional constraints beyond conventional IE: models must accurately capture numerical values, units, and modifiers, and preserve their structured relations to measured entities and properties under strict grounding. These constraints make the task particularly sensitive to hallucinations and motivate learning objectives that explicitly penalize fabrication, mis-scoping, and relational incompleteness.

2.3 Scientific Measurement Extraction and Benchmarks

Scientific information extraction has been advanced by datasets such as SciERC (Luan et al., 2018) and MeasEval (Harper et al., 2021b). Among them, MeasEval provides the most fine-grained annotation schema for scientific measurements, including quantities, units, modifiers, and their relations, and has become a key benchmark for evaluating measurement extraction systems. Despite progress, numerically grounded and relation-consistent extraction remains challenging, especially for complex sentences containing multiple measurements and implicit constraints. Our work builds on the MeasEval schema and targets these persistent failure modes with a hallucination-aware optimization framework.

2.4 Mitigation Strategies for LLM Hallucinations

A wide range of techniques have been proposed to reduce hallucinations in LLMs, including retrieval-augmented generation (RAG) (Lewis et al., 2020), supervised fine-tuning (SFT) (Zhou et al., 2023), chain-of-thought prompting (Wei et al., 2022), process-based supervision (Lightman et al., 2023), reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022), and direct preference optimization (DPO) (Rafailov et al., 2023). While effective for open-ended generation, these methods are not explicitly designed to enforce the strict grounding and structural consistency required by scientific measurement extraction. In contrast, our approach integrates reasoning-aware SFT with targeted reinforcement learning and structured reward shaping, explicitly encoding measurement-specific constraints to suppress hallucinations at their structural root.

Despite significant progress in hallucination mitigation, prior work has neither systematically characterized hallucinations in scientific measurement extraction nor introduced specialized reward objectives tailored to its error patterns. We bridge this gap by unifying a fine-grained hallucination taxonomy with a progressive optimization framework designed specifically for measurement-specific error suppression.

3 Methodology

Informed by our analysis in Section 2, we design MeasHalu around a central hypothesis: *scientific measurement hallucinations arise from two fundamentally different failure modes—unreliable quantitative reasoning and fragile relational grounding*. Accordingly, our framework adopts a two-branch mitigation strategy, targeting **Quantity Hallucinations** and **Relation-based Hallucinations** respectively. As illustrated in Figure 2, MeasHalu integrates progressive supervised fine-tuning with hallucination-aware reinforcement learning, enabling the model to internalize strict scientific grounding constraints directly into its reasoning process. These hallucinations (see Table 8) significantly undermine the reliability of LLMs for this critical task.

3.1 Quantity Hallucination Mitigation

Unlike prior approaches that employ end-to-end joint training for quantities and relations, our method first trains quantity extraction independently. Furthermore, following the SFT stage, we incorporate a GRPO phase specifically driven by hallucination-targeted rewards to further mitigate hallucinations.

3.1.1 Progressive Supervised Fine-tuning

To endow the LLM with structured quantity reasoning capabilities, we adopt a progressive SFT strategy. Specifically, we first utilize $\mathcal{D}_{\text{aug}}$ to establish foundational quantity reasoning skills, followed by fine-tuning on $\mathcal{D}_{\text{trace}}$ to ensure rigorous alignment with MeasEval standards. The construction details of these two datasets are elaborated below.

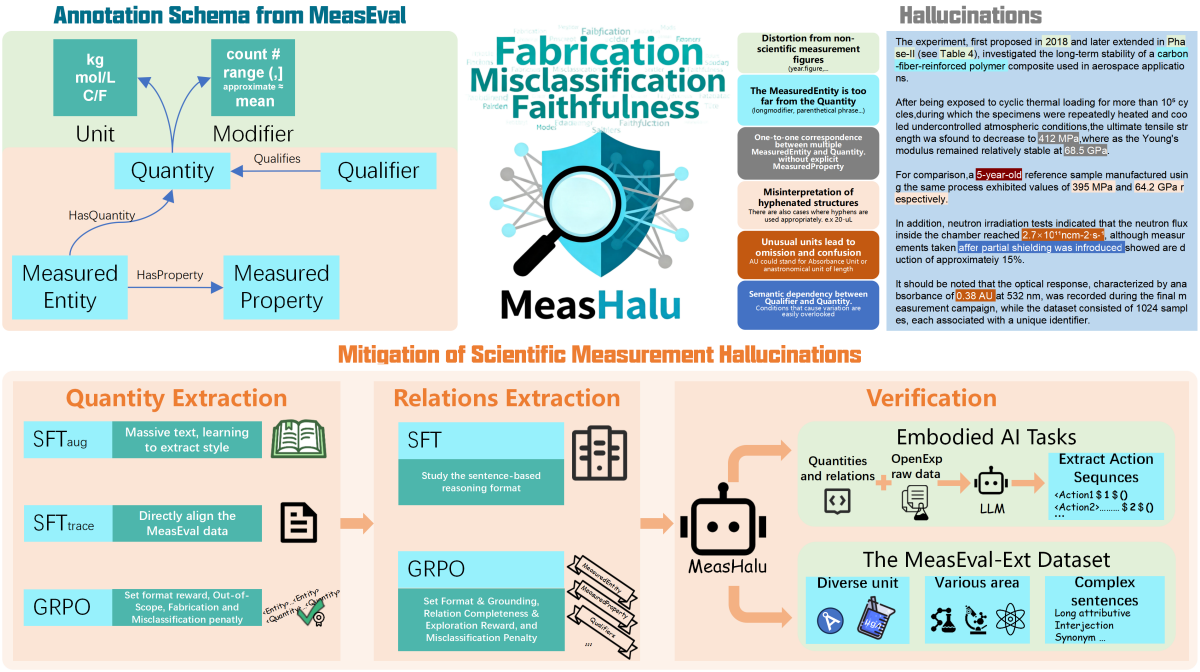

Figure 2: Overview of our method consisting of two stages, Supervised Fine-Tuning & GRPO based Reinforcement Learning.

$\mathcal{D}_{\text{aug}}$ We curate an unlabeled corpus $\mathcal{X}_{\text{un}}$ from arXiv paper abstracts (Cohan et al., 2018). Lacking gold quantity annotations, we use Quantulum3 (f_{qtm}, [2]) to extract quantity candidates, then leverage an augmentation template $\mathcal{P}_{\text{aug}}$ to prompt $\mathcal{M}$ to verify these anchors and generate a reasoning trajectory h_{aug}. Formally, for $x \in \mathcal{X}_{\text{un}}$:

$$\tilde{y} \leftarrow f_{\text{qtm}}(x), \quad h_{\text{aug}} \leftarrow \mathcal{M}(x, \tilde{y}; \mathcal{P}_{\text{aug}}) \quad (1)$$

where $\tilde{y}$ is noisy pseudo-labels from f_{qtm}, and $\mathcal{P}_{\text{aug}}$ guides $\mathcal{M}$ to filter false positives via semantics. Valid trajectories form $\mathcal{D}_{\text{aug}} = \{(x, h_{\text{aug}})\}_{i=1}^{20K}$.

$\mathcal{D}_{\text{trace}}$ We leverage the MeasEval dataset with human-annotated gold quantity labels y_{gt}, and adopt a *traceback template* $\mathcal{P}_{\text{trace}}$ to guide reasoning reconstruction: given $y$gt, the model generates a stepwise reasoning trajectory h_{trace} leading to the gold conclusion, formulated as:

$$h_{\text{trace}} \leftarrow \mathcal{M}(x, y_{\text{gt}}; \mathcal{P}_{\text{trace}}) \quad (2)$$

To ensure correctness of the reasoning trajectory, we enforce strict consistency validation via $\mathbb{I}(\cdot)$, where concl$(\cdot)$ extracts the final quantity from h_{trace}. The filtered dataset is constructed as:

$$\mathcal{D}_{\text{trace}} = \{(x, h_{\text{trace}}) \mid \mathbb{I}\big(\text{concl}(h_{\text{trace}}), y_{\text{gt}}\big) = 1\} \quad (3)$$

[2]https://github.com/nielstron/quantulum3

The prompts $\mathcal{P}_{\text{trace}}$ and $\mathcal{P}_{\text{aug}}$ are provided in Appendix A.

3.1.2 Hallucination-targeted Reward Function

The total reward $R(y)$ is a weighted sum of four components targeting distinct quantity-related hallucinations:

$$R = w_1 r_{\text{fmt}} + w_2 r_{\text{scope}} + w_3 r_{\text{fab}} + w_4 r_{\text{mis}} \quad (4)$$

Format compliance reward (r_{fmt}): A binary reward is assigned for strict adherence to the predefined structure <ARABIC>, ..., <CONCLUSION>, enforcing schema compliance and parsability of generated reasoning chains.

Out-of-scope hallucination penalty (r_{scope}): A penalty is imposed when the model extracts out-of-scope entities—such as figure labels (e.g., "Fig. 1")—that do not constitute valid numerical data. This mechanism utilizes pattern recognition to identify and penalize specific noisy strings, while simultaneously penalizing any generated answers that fail to match the ground truth, ensuring that the model avoids generating arbitrary numbers that deviate from the objective definitions.

Fabrication hallucination penalty (r_{fab}): This penalty targets invalid quantity fabrication by verifying each extracted entity against a hybrid phys-

ical parser $\mathcal{T}_{\text{parse}}$. A penalty is triggered if the extracted string fails to be parsed as a valid physical or numerical quantity, preventing the model from inventing nonsensical values.

Misclassification hallucination reward (r_{mis}): A reward is assigned based on token-level precision to mitigate misclassification hallucinations. This mechanism imposes a penalty if the model generates excessively long spans that erroneously incorporate surrounding components, such as the MeasuredEntity, into the quantity extraction.

Detailed mathematical derivations and implementation specifics for these reward components are provided in Appendix D.

3.2 Relation-based Hallucination Mitigation

The extraction of relation-based scientific measurements is particularly challenging due to long-range contextual dependencies that frequently induce hallucinations. Compared to traditional rule-based approaches that generate answers after exhaustively processing complex constraints, our approach first pinpoints the quantity-containing sentence, with subsequent reasoning anchored to this local context to extract the quantity and its relations—eliminating cross-sentence hallucination triggers. We implement this strategy via SFT for schema establishment and GRPO for hallucination-targeted alignment.

3.2.1 Quantity-Guided Relation Extraction

Given the document text x and a list of candidate quantities $\mathcal{Q}$in, the extraction pipeline Ahalu follows a two-stage chain-of-thought reasoning process.

First, the model identifies the evidence sentences $S = s_1, \ldots, s_n$ that contain quantities in $\mathcal{Q}$in:

$$S \leftarrow A_{\text{halu}}(x, \mathcal{Q}_{\text{in}}) \tag{5}$$

where S denotes the target sentences.

The model then performs fine-grained reasoning over S to resolve quantity attributes (e.g., units and modifiers) and associate them with their corresponding measured entities or properties, yielding the final structured relations $\mathcal{R}$.

This two-stage reasoning is learned via supervised fine-tuning on rule-derived traces from MeasEval annotations.

3.2.2 Hallucination-targeted Reward Function

While sentence-based extraction excels at local entity identification, it often struggles to capture long-range dependency chains (e.g., *MeasuredEntity*, *Qualifier*), which frequently induces inference bias and leads to the under-extraction of sparse components. To mitigate these reasoning biases and suppress the resulting hallucinations, we design a composite reward function R optimized via GRPO. The reward function are formulated as:

$$R = w_1 r_{\text{fmt}} + w_2 r_{\text{comp}} + w_3 r_{\text{mis}} \tag{6}$$

The design rationale for each reward component is detailed below:

Format compliance reward (r_{fmt}) A composite reward is assigned to enforce strict adherence to the quantitative schema and ensure textual grounding. It imposes two constraints: first, validating the structural segmentation of reasoning sections to prevent schema collapse; second, verifying that each extracted sentence can be mapped to a valid text span in the source document.

Relational completeness reward (r_{comp}) To mitigate inference-induced and role-definition hallucinations stemming from broken dependency links, this reward is designed to enforce the structural integrity of the reasoning chain. The mechanism drives comprehensive exploration through a two-tier incentive structure: (1) a stepwise reward for incremental component extraction and a weighted exploration term that prioritizes harder-to-predict components to drive model exploration (2) a completeness bonus awarded only upon full recovery of the gold-standard relation group to enforce the structural integrity of the reasoning chain.

Misclassification hallucination reward (r_{mis}) A reward is assigned based on token-level precision to mitigate misclassification hallucinations. This mechanism imposes a penalty if the model generates excessively long spans that erroneously incorporate surrounding components. The detailed mathematical formulations for these reward components are provided in Appendix E.

4 Experiments

In this section, We evaluate our method on quantity extraction and relation identification using the MeasEval benchmark. Additionally, we introduce

MeasEval-Ext, a specialized dataset annotated from recent literature to target novel units and complex expressions absent from the training distribution. Further analyses are conducted including entropy dynamic analysis and its utility as a functional tool within downstream embodied AI tasks.

4.1 Effectiveness of Quantity Hallucination Mitigation Strategies

Setup We utilize the Quantity subset of the MeasEval dataset as our primary evaluation benchmark. To verify the scalability and robustness of our approach, we employ the *Qwen2.5-Instruct* series across three different scales (0.5B, 3B, and 7B) as the base models.

Model Setting	0.5B	3B	7B
MeasHalu-Quant	$0.749_{\pm 0.006}$	$0.812_{\pm 0.006}$	$0.849_{\pm 0.006}$
w/o ($\mathcal{D}_{\text{trace}}$ + GRPO)	$0.475_{\pm 0.011}$	$0.481_{\pm 0.001}$	$0.465_{\pm 0.011}$
w/o ($\mathcal{D}_{\text{aug}}$ + GRPO)	$0.408_{\pm 0.028}$	$0.346_{\pm 0.008}$	$0.397_{\pm 0.027}$
w/o GRPO	$0.596_{\pm 0.008}$	$0.539_{\pm 0.002}$	$0.585_{\pm 0.018}$

Table 1: Quantity Information Extraction Performance of 0.5B, 3B and 7B Models (Mean $\pm$ Std)

Results Our full model MEASHALU-QUANT achieves consistent performance advantages across all model scales in Table 1. Compared to the baseline without GRPO, the integration of GRPO drives SOTA results of 0.749, 0.812, and 0.849 for 0.5B, 3B, and 7B models, validating that our rule-based reward system enables stable anti-hallucination alignment even for low-capacity models.

Using only gold-standard data (w/o ($\mathcal{D}_{\text{aug}}$ + GRPO)) gives the lowest scores (e.g., 0.346 for 3B), showing models cannot capture complex multi-domain quantitative annotation rules without prior measurement extraction schema scaffolding.

The single first stage (w/o ($\mathcal{D}_{\text{trace}}$ + GRPO)) brings marginal gains from data scaling but is sub-optimal, while (w/o GRPO) delivers substantial improvements (e.g., 3B score raised to 0.539). It confirms that the 1st stage initializes quantitative schema adherence, while the 2nd stage enhances generalization across diverse scientific contexts by leveraging multi-domain scientific knowledge.

4.2 Effectiveness of Relation-based Hallucination Mitigation Strategies

In this section, we evaluate our relation-based hallucination mitigation method on the MeasEval dataset and *MeasEval-Ext*, a newly annotated dataset containing novel expressions absent from MeasEval, designed to assess the model's generalization and robustness.

Setup We use the *Qwen2.5-Instruct* models (0.5B, 3B, 7B) to assess performance across parameter scales. Although MeasEval is a high-quality benchmark, its limited size and dated sources underrepresent emerging units. To evaluate robustness under distribution shift, we introduce *MeasEval-Ext*, annotated strictly following the MeasEval schema.

We employ an adversarial strategy by selecting recent literature containing novel units and complex expressions absent from the training distribution, rigorously testing model generalization beyond memorized vocabulary. Annotations followed MeasEval guidelines (see Appendix C for agreement analysis).

Results over MeasEval Table 2 compares complex quantitative relation extraction on the MeasEval test set. MEASHALU-7B achieves an overall F1 of **0.512**, closely matching the competition winner *LIORI* (Davletov et al., 2021) (0.519)[3], and substantially outperforming other supervised baselines such as *CONNER* (Cao et al., 2021) (0.473) and *Counts* (Gangwar et al., 2021) (0.432).

Our model also surpasses state-of-the-art proprietary LLMs (e.g., GPT-5, Gemini-2.5-Pro). Even with optimized sentence-based prompting, GPT-5 reaches only 0.406 F1, leaving MEASHALU-7B ahead by over 10 points. This result highlights the necessity of our quantitative domain alignment pipeline (SFT + composite reward optimization) for mitigating relational Quantity hallucinations.

Across all baseline LLMs, sentence-based prompting consistently outperforms rule-based prompting (e.g., Gemini-2.5-Pro improves from 0.359 to 0.440), supporting our hypothesis that sentence-level localized reasoning is more effective than rigid global rule-based deduction for complex quantitative relation extraction.

As shown in Table 3, the results on *MeasEval-Ext* expose a significant performance gap: while general LLMs exhibit non-uniform shifts—often struggling with novel expressions—MEASHALU demonstrates robust generalization to unseen distributions, substantially widening its lead over all baselines. Detailed statistics with standard deviations can be found in Table 9 and Table 10.

[3]LIORI uses a six-model ensemble and does not release weights.

Model	Overall	Quantities			Entities		Properties		Qualifiers	
		Quantity	Unit	Modifier	ME	HasQuantity	MP	HasProperty	Qualifier	Qualifies
Top Ranked Systems from the MeasEval Competition										
Baseline	0.239	0.827	0.561	0.000	0.053	0.075	0.064	0.007	0.005	0.000
Counts	0.432	0.861	0.804	0.614	0.406	0.311	0.245	0.183	0.077	0.064
CONNER	0.473	0.855	0.719	0.523	0.398	0.424	0.437	0.257	0.000	0.000
LIORI	0.519	0.861	0.722	0.642	0.437	0.482	0.467	0.318	0.163	0.092
Rule-based Prompting										
Qwen2.5-7b-inst	0.171	0.491	0.478	0.106	0.088	0.045	0.057	0.000	0.040	0.017
Qwen2.5-72b-inst	0.286	0.644	0.826	0.236	0.196	0.147	0.164	0.001	0.076	0.021
DeepSeek-R1	0.253	0.569	0.586	0.240	0.216	0.163	0.163	0.024	0.085	0.029
DeepSeek-V3	0.271	0.657	0.768	0.214	0.239	0.135	0.113	0.001	0.085	0.014
Gemini-2.5-Pro	0.359	0.712	0.784	0.464	0.306	0.266	0.287	0.090	0.146	0.076
GPT-5	0.371	0.804	0.742	0.395	0.361	0.270	0.355	0.020	0.152	0.052
Sentence-based Prompting										
Qwen2.5-7b-inst	0.073	0.151	0.160	0.027	0.066	0.059	0.044	0.031	0.003	0.005
Qwen2.5-72b-inst	0.212	0.403	0.516	0.232	0.204	0.137	0.113	0.087	0.038	0.012
DeepSeek-R1	0.304	0.589	0.711	0.356	0.260	0.198	0.182	0.118	0.113	0.057
DeepSeek-V3	0.320	0.567	0.726	0.355	0.303	0.225	0.226	0.149	0.019	0.000
Gemini-2.5-Pro	0.440	0.782	0.882	0.486	0.436	0.376	0.386	0.280	0.143	0.056
GPT-5	0.406	0.724	0.817	0.500	0.397	0.351	0.355	0.226	0.138	0.042
MeasHalu-0.5B	0.333	0.649	0.734	0.277	0.292	0.254	0.249	0.175	0.240	0.048
w/o GRPO	0.312	0.649	0.717	0.239	0.263	0.241	0.238	0.140	0.069	0.022
MeasHalu-3B	0.448	0.806	0.861	0.446	0.380	0.399	0.384	0.262	0.093	0.038
w/o GRPO	0.433	0.782	0.850	0.449	0.370	0.377	0.365	0.245	0.084	0.043
MeasHalu-7B	0.512	0.848	0.860	0.607	0.455	0.472	0.442	0.310	0.170	0.100
w/o GRPO	0.479	0.846	0.863	0.610	0.429	0.433	0.397	0.272	0.155	0.063

Table 2: Experimental results over the MeasEval Benchmark.Comparing MeasHalu with competition leaders and rule/sentence-based LLM baselines. Top ranks are shaded orange (1st), yellow (2nd), and teal (3rd).

Model	Overall	Quantities			Entities		Properties		Qualifiers	
		Quantity	Unit	Modifier	ME	HasQuantity	MP	HasProperty	Qualifier	Qualifies
Rule-based Prompting										
GPT-5	0.383	0.833	0.706	0.357	0.400	0.282	0.310	0.046	0.135	0.019
DeepSeek-R1	0.252	0.553	0.514	0.217	0.213	0.169	0.141	0.045	0.099	0.052
DeepSeek-V3	0.312	0.724	0.726	0.234	0.270	0.189	0.121	0.012	0.113	0.037
Gemini-2.5-Pro	0.386	0.766	0.707	0.444	0.351	0.312	0.291	0.151	0.140	0.055
Qwen2.5-72b	0.296	0.675	0.747	0.203	0.246	0.187	0.142	0.000	0.091	0.066
Qwen2.5-7b	0.181	0.501	0.353	0.148	0.108	0.078	0.069	0.001	0.038	0.006
Sentence-based Prompting										
GPT-5	0.402	0.750	0.759	0.458	0.435	0.303	0.303	0.239	0.100	0.070
DeepSeek-R1	0.324	0.622	0.648	0.255	0.299	0.235	0.215	0.146	0.075	0.069
DeepSeek-V3	0.299	0.521	0.565	0.229	0.296	0.229	0.209	0.155	0.053	0.048
Gemini-2.5-Pro	0.462	0.827	0.832	0.444	0.472	0.399	0.402	0.332	0.100	0.072
Qwen2.5-72b	0.202	0.343	0.383	0.183	0.207	0.145	0.127	0.106	0.048	0.050
Qwen2.5-7b	0.033	0.065	0.056	0.022	0.030	0.028	0.017	0.008	0.016	0.020
MeasHalu-7B	0.578	0.861	0.832	0.539	0.555	0.551	0.522	0.459	0.159	0.097

Table 3: Experimental results over the MeasEval-Ext.

4.3 Further Analysis

Mechanism of Hallucination Suppression via Entropy Dynamics Inspired by Cui and Ding (2025), we quantify Cognitive Hesitation via entropy dynamics, adapting the analysis to our task by distinguishing the quantity group (Quantity, Unit, Modifier) from the relation group (MeasuredEntity, MeasuredProperty, qualifier).

We focus on tokens strictly bounded by square brackets (e.g., parsing `70 m` from the tagged sequence `...surface form [70 m]...`). To capture micro-level certainty, we report four key statistics:

Bracket Entropy Mean (H_B) and Std (σ_B) measure the average confidence level; Spike Rate (R_B) for the proportion of brackets containing high-entropy tokens; and Sample Spike Ratio (R_{sample}) quantifies the proportion of samples containing at least one high-risk fluctuation.

Group	Metric	w/o GRPO	MeasHalu	Change
Quantity	H_B	0.0071 bits	**0.0034 bits**	↓ 52.1%
	σ_B	0.0729	**0.0477**	↓ 34.6%
	R_B	0.16%	0.32%	-
	R_{sample}	0.77%	1.54%	-
Relation	H_B	0.1147 bits	**0.0662 bits**	↓ 42.3%
	σ_B	0.3253	**0.2359**	↓ 27.5%
	R_B	13.01%	**5.23%**	↓ 59.8%
	R_{sample}	33.85%	**14.62%**	↓ 56.8%

Table 4: Fine-grained Entropy Statistics by Semantic Role.

Table 4 reveals a clear dichotomy in hallucination suppression across semantic roles. *1) Quantity Group Stability:* SFT is already near-deterministic ($H \approx 0.0071$). GRPO further compresses residual uncertainty ($H \approx 0.0034$) with negligible spike fluctuations ($R_{sample} \approx 1.54\%$). *2) Relation Group Sharpening:* GRPO reduces the spike ratio from 33.85% to 14.62% and lowers mean entropy by 42.3%. These results indicate that relational reasoning, which is highly ambiguous under SFT, becomes substantially more stable under GRPO. We attribute this improvement to the directed collapse induced by GRPO, which truncates long-tail uncertainty and enforces convergence toward deterministic facts.

Subsequently, to illustrate the effect of GRPO training on reasoning stability more intuitively, we select high-entropy points in the reasoning process for a case study. Details are in Appendix B.

Application for Embodied AI Tasks To validate the practical value of our fine-grained extraction for embodied AI, we adapt OpenExp (Liu et al., 2024) to a text-to-action generation task, where models generate executable chemical action sequences (e.g., *ADD . . . (100 mg)*) from unstructured experimental text, mimicking real-world automated laboratory scenarios.

We construct *OpenExp-Action-100*, a dataset of 100 diverse instances, by using unstructured experimental narratives as inputs and OpenExp's linearized action sequences as gold-standard outputs. To enable controlled comparison, we further define three experimental settings: Baseline (no augmentation), Gemini-Aug (quantity relations extracted by Gemini) and MeasEval-Aug (quantity relations extracted by MeasHalu).

Model	Source	Val	BLEU		LEV	ROUGE		
			B-2	B-4	50%	R-1	R-2	R-L
Gemini-2.5-Pro	MeasHalu	**14.67**	**58.72**	**44.23**	**60.00**	**71.71**	**54.24**	**66.51**
	Gemini	12.67	58.26	43.99	59.67	71.28	54.21	66.49
	None	0.33	51.78	37.23	37.33	62.92	46.25	58.90
DeepSeek-R1	MeasHalu	**10.67**	**58.99**	**43.01**	**61.33**	**71.62**	**52.79**	**65.85**
	Gemini	10.33	58.28	42.37	56.67	71.13	52.31	65.65
	None	8.67	38.32	26.21	22.00	59.12	42.32	53.56
GPT-5	MeasHalu	**16.33**	**51.55**	37.43	**52.33**	**71.00**	51.62	**65.47**
	Gemini	13.33	50.61	36.56	48.33	70.73	51.31	65.06
	None	1.35	50.39	**39.39**	50.17	66.31	**52.53**	62.76

Table 5: Performance on OpenExp-Action-100 with MeasEval-formatted quantity–relation context from different sources (MeasHalu vs. Gemini). Best scores per model are in bold.

Table 5 shows that injecting structured quantity relations significantly improves Structural Validity (Val, executable/logical consistency), with MeasEval-Aug (82.3%) outperforming Gemini-Aug and Baseline. The modest BLEU improvement (19.8 vs. 16.3) stems from gold-standard granularity mismatch. Specifically, OpenExp's minimalist annotations omit critical details (e.g., *anhydrous*) that our extraction retains. For embodied AI, structural validity—rather than textual overlap—is pivotal; MeasEval-formatted extraction ensures this validity by capturing critical details, providing constraints for executable instructions and practical utility for perception-to-action pipelines. Table 11 in the Appendix shows the full table with standard deviations.

5 Conclusion

The proposal of MEASHALU marks a significant step forward in systematically characterizing measurement hallucinations in large language models within the scientific extraction domain. Experimental results on both in-distribution and newly annotated out-of-distribution benchmarks (*MeasEval-Ext*) show that MEASHALU substantially improves robustness and consistently outperforms strong supervised baselines and state-of-the-art large language models. Ultimately, MEASHALU proves to be a reliable external tool that drives significant gains in downstream applications, validating its utility for embodied AI and AI4Science.

Limitations

Despite advances achieved in this paper, MeasHalu has notable limitations. First, even though MeasHalu outperforms all existing baselines, the extraction performance for sparse components (e.g., qualifiers, F1 = 0.170) remains low, hindered by limited annotated data and ambiguous semantic dependencies in scientific text. Second, the framework's generalization to low-resource languages or domain-specific jargon (e.g., niche engineering units) is untested, as current training data focuses on English scientific literature. Third, processing ultra-long documents with nested measurement relations may introduce computational inefficiencies, as the sentence-based reasoning strategy requires contextual localization for each quantity.

References

Satya Almasian, Vivian Kazakova, Philipp Göldner, and Michael Gertz. 2023a. Cqe: a comprehensive quantity extractor. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12845–12859.

Satya Almasian, Vivian Kazakova, Philipp Göldner, and Michael Gertz. 2023b. CQE: A comprehensive quantity extractor. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12845–12859, Singapore. Association for Computational Linguistics.

Soumia Lilia Berrahou, Patrice Buche, Juliette Dibie-Barthelemy, and Mathieu Roche. 2013. How to extract unit of measure in scientific documents? In *Special Session on Text Mining*, volume 2, pages 249–256. SCITEPRESS.

Jiarun Cao, Yuejia Xiang, Yunyan Zhang, Zhiyuan Qi, Xi Chen, and Yefeng Zheng. 2021. CONNER: A cascade count and measurement extraction tool for scientific discourse. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1239–1244, Online. Association for Computational Linguistics.

Qiguang Chen, Mingda Yang, Libo Qin, Jinhao Liu, Zheng Yan, Jiannan Guan, Dengyun Peng, Yiyan Ji, Hanjing Li, Mengkang Hu, Yimeng Zhang, Yihao Liang, Yuhang Zhou, Jiaqi Wang, Zhi Chen, and Wanxiang Che. 2025. Ai4research: A survey of artificial intelligence for scientific research. *Preprint*, arXiv:2507.01903.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*.

Ganqu Cui and Ning Ding. 2025. The entropy mechanism of reinforcement learning for reasoning language models. *Computing Magazine of the CCF*, 1(7):26–33.

John Dagdelen, Alexander Dunn, Sanghoon Lee, Nicholas Walker, Andrew S Rosen, Gerbrand Ceder, Kristin A Persson, and Anubhav Jain. 2024. Structured information extraction from scientific text with large language models. *Nature communications*, 15(1):1418.

Adis Davletov, Denis Gordeev, Nikolay Arefyev, and Emil Davletov. 2021. LIORI at SemEval-2021 task 8: Ask transformer for measurements. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1249–1254, Online. Association for Computational Linguistics.

Luca Foppiano, Guillaume Lambard, Toshiyuki Amagasa, and Masashi Ishii. 2024. Mining experimental data from materials science literature with large language models: an evaluation study. *Science and Technology of Advanced Materials: Methods*, 4(1):2356506.

Akash Gangwar, Sabhay Jain, Shubham Sourav, and Ashutosh Modi. 2021. Counts@IITK at SemEval-2021 task 8: SciBERT based entity and semantic relation extraction for scientific data. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1232–1238, Online. Association for Computational Linguistics.

Mark A. Hanson, Pablo Gómez Barreiro, Paolo Crosetto, and Dan Brockington. 2024. The strain on scientific publishing. *Quantitative Science Studies*, 5(4):823–843.

Corey Harper, Jessica Cox, Curt Kohler, Antony Scerri, Ron Daniel Jr., and Paul Groth. 2021a. SemEval-2021 task 8: MeasEval – extracting counts and measurements and their related contexts. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 306–316, Online. Association for Computational Linguistics.

Corey Harper, Jessica Cox, Curt Kohler, Antony Scerri, Ron Daniel Jr, and Paul Groth. 2021b. Semeval-2021 task 8: Measeval–extracting counts and measurements and their related contexts. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 306–316.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38.

Olga Kononova, Tanjin He, Haoyan Huo, Amalie Trewartha, Elsa A. Olivetti, and Gerbrand Ceder. 2021. Opportunities and challenges of text mining in materials research. *iScience*, 24(3).

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Chaozhuo Li, Pengbo Wang, Chenxu Wang, Litian Zhang, Zheng Liu, Qiwei Ye, Yuanbo Xu, Feiran Huang, Xi Zhang, and Philip S Yu. 2025. Loki's dance of illusions: A comprehensive survey of hallucination in large language models. *arXiv preprint arXiv:2507.02870*.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*.

Zhiyuan Liu, Yaorui Shi, An Zhang, Sihang Li, Enzhi Zhang, Xiang Wang, Kenji Kawaguchi, and Tat-Seng Chua. 2024. ReactXT: Understanding molecular "reaction-ship" via reaction-contextualized molecule-text pretraining. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5353–5377, Bangkok, Thailand. Association for Computational Linguistics.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. *arXiv preprint arXiv:1808.09602*.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Maciej P Polak and Dane Morgan. 2024. Extracting accurate materials data from research papers with conversational language models and prompt engineering. *Nature Communications*, 15(1):1569.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Tarek Saier, Mayumi Ohta, Takuto Asakura, and Michael Färber. 2024. Hyperpie: Hyperparameter information extraction from scientific publications. In *Advances in Information Retrieval*, pages 254–269, Cham. Springer Nature Switzerland.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, and 1 others. 2023. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36:55006–55021.

A Prompt template

Prompt for $\mathcal{P}_{\text{trace}}$

Instruction:
You are an expert in extracting structured annotations from text. I have an text input and you need to extract all the quantities within it. I need you to strictly follow the format with six specific sections: ARABIC-QUANTITY, NUMERIC-QUANTITY, TIME-QUANTITY, CHANGE-QUANTITY, CHANGE-QUANTITY, FORMULA-QUANTITY, CONCLUSION.
To explain further: In ARABIC-QUANTITY, outline a step-by-step thought process you use to extract quantity in arabic form. In NUMERIC-QUANTITY, outline a step by step thought process . . . In CONCLUSION, give the final answer in a tsv format explained below.
I will provide you with the quantities extracted using the quantulum library for your reference, the information provided by Quantulum is standardized. You need to find the original text in the passage and fill in the tsv form. Also, the quantulum information maybe incorrect, You can't follow it completely.

Here's how the format should look: <ARABIC-QUANTITY> [Provide a chain-of-thought explanation of how you extract all quantities in the arabic forms] </ARABIC-QUANTITY> <NUMERIC-QUANTITY> . . . <CONCLUSION>[State the final answer in a tsv format explained below format. . .] </CONCLUSION>

Task Definition: Extract Quantities
1. Annotation of Quantities: . . .
2. Example Process: . . .
Output Format (TSV Fields): . . .
Final Output Example: . . .
The reference answer from quantulum: . . .

Prompt for $\mathcal{P}_{\text{aug}}$

Instruction:
You are an expert in extracting structured annotations from text. I have an text input and you need to extract all the quantities within it. I need you to strictly follow the format with six specific sections: ARABIC-QUANTITY, NUMERIC-QUANTITY, TIME-QUANTITY, CHANGE-QUANTITY, CHANGE-QUANTITY, FORMULA-QUANTITY, CONCLUSION.
To explain further: In ARABIC-QUANTITY, outline a step-by-step thought process you use to extract quantity in arabic form. In NUMERIC-QUANTITY, outline a step by step thought process . . . In CONCLUSION, give the final answer in a tsv format explained below. It is crucial that you adhere to this structure exactly as outlined and that the final answer in the CONCLUSION matches the standard correct answer precisely.

Here's how the format should look: <ARABIC-QUANTITY> [Provide a chain-of-thought explanation of how you extract all quantities in the arabic forms] </ARABIC-QUANTITY> <NUMERIC-QUANTITY> . . . <CONCLUSION>[State the final answer in a tsv format explained below format. . . It is crucial that you adhere to this structure exactly as outlined and that the final answer] </CONCLUSION>

Task Definition: Extract Quantities
1. Annotation of Quantities: . . .
2. Example Process: . . .
Output Format (TSV Fields): . . .
Final Output Example: . . .
The gold answers: . . .

B Case Study: High-Entropy Token Suppression by GRPO

To better understand the effect of GRPO, we examine a representative sample:

Input:

> Samples were then annealed in air in a pre-heated furnace at temperatures up to 798 °C for times chosen to ensure complete iron diffusion through the sample.

Gold Standard Annotation:

- **Quantity:** `up to 798 °C`
- **Unit:** `°C`
- **Modifiers:** `["IsRange"]`
- **MeasuredProperty:** `annealed in air in a pre-heated furnace at temperatures`
- **MeasuredEntity:** `Samples`

SFT Output:

> We can find the quantity with surface form [up to 798 °C], it has unit [°C]. The modifier for the quantity are [IsRange]. This quantity is used to describe the entity [furnace]. The entity has the following property [temperatures].

Here, the red token indicates a high-entropy token (entropy = 1.39), and the top-5 candidate tokens with probabilities are listed in the first column of Table 6.

GRPO Output:

> We can find the quantity with surface form [up to 798 °C], it has unit [°C]. The modifier for the quantity are [IsRange]. This quantity is used to describe the entity [Samples]. The entity has the following property [temperatures].

There is no high-entropy token in the model's output after the GRPO training. The top-5 candidate tokens with probabilities at the [Samples] position are shown in the second column of Table 6.

Analysis: GRPO successfully suppresses the high-entropy token observed in SFT, assigning the correct token `Samples` with high confidence and eliminating uncertainty, demonstrating improved reasoning stability and more deterministic output.

Rank	w/o GRPO Candidate	w/o GRPO Prob	GRPO Candidate	GRPO Prob
1	f	0.547	**Samples**	0.847
2	Samples	0.376	‘ Samples	0.115
3	‘ Samples	0.051	f	0.037
4	samples	0.015	samples	0.0013
5	pre	0.011	_samples	0.0001

Table 6: Comparison of Top-5 candidate tokens at the target position between SFT and GRPO outputs.

C MeasEval-Ext and its Annotation Details

The annotations are drawn from recent research papers that postdate the original MeasEval corpus. A distinct advantage of this data source is its adversarial selection strategy: unlike the randomized distribution in the original dataset, we deliberately curated 135 text segments(the same as the MeasEval evaluation dataset) containing **novel units and complex quantity expressions absent from the training distribution**. This design ensures that the dataset strictly tests the model's ability to identify and ground quantities based on semantic context rather than memorized vocabulary.

To ensure high data quality, we enlisted researchers from our laboratory as annotators. The annotation process strictly followed the official *MeasEval Annotation Guidelines*. All samples were **independently labeled by two annotators** to capture the dense quantity-centric information. Following the initial annotation, results were reviewed and reconciled during an **adjudication meeting** to resolve disagreements and reach a final consensus.

The consistency of the dataset is validated by the Inter-Annotator Agreement (IAA). As shown in Table 7, the Krippendorff's Alpha scores (e.g., 0.921 for Quantity) indicate strong agreement, comparable to the original MeasEval benchmarks.

Annotation Class	Krippendorff's α
Quantity	0.921
MeasuredEntity	0.639
MeasuredProperty	0.584
Qualifier	0.416

Table 7: Inter-Annotator Agreement (Krippendorff's Alpha) for MeasEval-Ext

D Quantity Phase Reward

The total reward $R(y)$ is a weighted sum of four components for mitigating distinct Quantity hallucination types, which includes the Format Reward (r_{fmt}), the Out-of-Scope Penalty (r_{scope}), the Fabrication Penalty (r_{fab}) and the Misclassification Penalty (r_{mis}):

$$R(y) = r_{\text{fmt}}(y) + r_{\text{scope}}(y) + r_{\text{fab}}(y) + r_{\text{mis}}(y) \quad (7)$$

r_{fmt} To enforce output schema compliance, we validate the sequential semantic tags $\mathcal{S}_{\text{tags}} = \{\texttt{<ARABIC>}, \ldots, \texttt{<CONCLUSION>}\}$ via regex pattern $\mathcal{P}_{\text{struct}}$. The binary reward is:

$$r_{\text{fmt}}(y) = \mathbb{I}\Big(y \equiv \mathcal{P}_{\text{struct}}\Big) \quad (8)$$

r_{scope} Constrains out-of-scope entities(e.g., "Fig. 1") via local patterns $\mathcal{C}(e)$ and global precision P_{ans}:

$$r_{\text{scope}} = -\lambda_{\text{loc}} \sum_{e} \mathcal{C}(e) + \beta_{\text{scope}} P_{\text{ans}} \quad (9)$$

r_{fab} Prohibiting invalid quantity fabrication via parsers combining CQE(Almasian et al., 2023b) and Quantulum[4] $\mathcal{T}_{\text{parse}}$, the penalty includes grounding constraints:

$$r_{\text{fab}}(y) = -\lambda_{\text{fab}} \sum_{e \in \mathcal{E}_y} \mathbb{I}\Big(\mathcal{T}_{\text{parse}}(e) = \emptyset\Big) \quad (10)$$

r_{mis} Mitigating span boundary errors via token-level precision P_{tok}, the reward is:

$$r_{\text{mis}}(y) = \bar{F}1_{\text{tok}} - \lambda_{\text{mis}} \cdot (1 - P_{\text{tok}}) \quad (11)$$

E Relation Phase Reward

While the sentence-based extraction excels at local entity identification (e.g., Units, Modifiers), it suffers from failures in capturing long-range dependency chains (e.g., *MeasuredEntity*, *MeasuredProperty*, *Qualifier*), inference bias, and under-extraction of sparse components. To address these issues, enforce logical completeness, suppress Quantity hallucinations, and incentivize sparse component retrieval, we design a composite reward function $R(y)$ optimized via GRPO. The total reward is a weighted sum of three dedicated components (Format & Grounding Reward (r_{fmt})), Relation Completeness & Exploration Reward (r_{comp}), and Misclassification Penalty (r_{mis})), that target distinct quantitative extraction flaws:

$$R(y) = r_{\text{fmt}}(y) + r_{\text{comp}}(y) + r_{\text{mis}}(y) \quad (12)$$

The reward components are elaborated as follows with explicit optimization objectives and mathematical formulations:

r_{fmt} To enforce structural consistency, we validate the existence of analysis sections $\mathcal{S}_y$ and adherence to the SFT schema $\mathcal{F}_{\text{SFT}}$. The binary reward is:

$$r_{\text{fmt}}(y) = \mathbb{I}\Big(\mathcal{S}_y \neq \emptyset \wedge y \models \mathcal{F}_{\text{SFT}}\Big) \quad (13)$$

r_{comp} Drives **comprehensive exploration** by aligning predicted groups p with gold groups g. To prevent partial extraction, we incentivize full recovery via stepwise matching, closure bonuses, and weighted component bonuses:

$$r_{\text{comp}}(y) = \sum_{p \sim g} \Big(\underbrace{\lambda_{\text{step}} |p \cap g|}_{\text{Stepwise}} + \underbrace{\beta_{\text{full}} \mathbb{I}(g \subseteq p)}_{\text{Closure}} \Big) + \underbrace{\lambda_{\text{exp}} \sum\nolimits_{c} w_c F1_c^{\text{ans}}}_{\text{Weighted Exploration}} \quad (14)$$

where weights $\mathbf{w}$ prioritize harder-to-predict dependencies to ensure no critical node is missed.

r_{mis} Suppresses over-broad spans by penalizing token-level precision loss $(1 - P_{\text{tok}})$:

$$r_{\text{mis}}(y) = F1_{\text{tok}} - (1 - P_{\text{tok}}) \quad (15)$$

[4] https://github.com/nielstron/quantulum3

Type	Fabrication	Out-of-Scope	Misclassification	Inference Bias	Role Definition
Quantity	Generates fictitious values absent from the source text, or yields extracted strings that contain no valid numerical content.	Extracts invalid numerical tokens, including figure citations (e.g., "Fig. 4") or scientific nomenclature containing digits (e.g., "4S RNA").	Generates excessively long spans that erroneously incorporate surrounding components, such as the *MeasuredEntity*.	–	–
Relation	–	–	Generates excessively long spans that erroneously incorporate surrounding context or unrelated text segments.	Propagates errors from preceding components (e.g., incorrect *MeasuredEntity* will result in cascading hallucinations in properties and qualifiers).	Fails to distinguish semantic roles, frequently inverting the *MeasuredEntity* and *MeasuredProperty*.

Table 8: Taxonomy of Hallucinations in Information Extraction

Model	Overall	Quantities			Entities		Properties		Qualifiers	
		Quantity	Unit	Modifier	ME	HasQuantity	MP	HasProperty	Qualifier	Qualifies
Top Ranked Systems from the MeasEval Competition										
Baseline	0.239	0.827	0.561	0.000	0.053	0.075	0.064	0.007	0.005	0.000
Counts	0.432	0.861	0.804	0.614	0.406	0.311	0.245	0.183	0.077	0.064
CONNER	0.473	0.855	0.719	0.523	0.398	0.424	0.437	0.257	0.000	0.000
LIORI	0.519	0.861	0.722	0.642	0.437	0.482	0.467	0.318	0.163	0.092
Rule-based Prompting										
Qwen2.5-7b-inst	$0.171_{\pm 0.028}$	$0.491_{\pm 0.052}$	$0.478_{\pm 0.075}$	$0.106_{\pm 0.011}$	$0.088_{\pm 0.021}$	$0.045_{\pm 0.015}$	$0.057_{\pm 0.008}$	$0.000_{\pm 0.000}$	$0.040_{\pm 0.012}$	$0.017_{\pm 0.006}$
Qwen2.5-72b-inst	$0.286_{\pm 0.028}$	$0.644_{\pm 0.035}$	$0.826_{\pm 0.026}$	$0.236_{\pm 0.115}$	$0.196_{\pm 0.038}$	$0.147_{\pm 0.028}$	$0.164_{\pm 0.042}$	$0.001_{\pm 0.002}$	$0.076_{\pm 0.019}$	$0.021_{\pm 0.011}$
DeepSeek-R1	$0.253_{\pm 0.008}$	$0.569_{\pm 0.002}$	$0.586_{\pm 0.002}$	$0.240_{\pm 0.018}$	$0.216_{\pm 0.017}$	$0.163_{\pm 0.014}$	$0.163_{\pm 0.003}$	$0.024_{\pm 0.010}$	$0.085_{\pm 0.015}$	$0.029_{\pm 0.002}$
DeepSeek-V3	$0.271_{\pm 0.008}$	$0.657_{\pm 0.018}$	$0.768_{\pm 0.010}$	$0.214_{\pm 0.003}$	$0.239_{\pm 0.009}$	$0.135_{\pm 0.014}$	$0.113_{\pm 0.011}$	$0.001_{\pm 0.002}$	$0.085_{\pm 0.004}$	$0.014_{\pm 0.004}$
Gemini-2.5-Pro	$0.359_{\pm 0.008}$	$0.712_{\pm 0.007}$	$0.784_{\pm 0.035}$	$0.464_{\pm 0.009}$	$0.306_{\pm 0.009}$	$0.266_{\pm 0.007}$	$0.287_{\pm 0.021}$	$0.090_{\pm 0.009}$	$0.146_{\pm 0.018}$	$0.076_{\pm 0.014}$
GPT-5	$0.371_{\pm 0.004}$	$0.804_{\pm 0.013}$	$0.742_{\pm 0.018}$	$0.395_{\pm 0.020}$	$0.361_{\pm 0.003}$	$0.270_{\pm 0.005}$	$0.355_{\pm 0.014}$	$0.020_{\pm 0.004}$	$0.152_{\pm 0.019}$	$0.052_{\pm 0.006}$
Sentence-based Prompting										
Qwen2.5-7b-inst	$0.073_{\pm 0.006}$	$0.151_{\pm 0.008}$	$0.160_{\pm 0.007}$	$0.027_{\pm 0.001}$	$0.066_{\pm 0.009}$	$0.059_{\pm 0.010}$	$0.044_{\pm 0.010}$	$0.031_{\pm 0.010}$	$0.003_{\pm 0.004}$	$0.005_{\pm 0.006}$
Qwen2.5-72b-inst	$0.212_{\pm 0.002}$	$0.403_{\pm 0.006}$	$0.516_{\pm 0.005}$	$0.232_{\pm 0.010}$	$0.204_{\pm 0.005}$	$0.137_{\pm 0.007}$	$0.113_{\pm 0.008}$	$0.087_{\pm 0.005}$	$0.038_{\pm 0.011}$	$0.012_{\pm 0.007}$
DeepSeek-R1	$0.304_{\pm 0.004}$	$0.589_{\pm 0.016}$	$0.711_{\pm 0.021}$	$0.356_{\pm 0.030}$	$0.260_{\pm 0.011}$	$0.198_{\pm 0.012}$	$0.182_{\pm 0.006}$	$0.118_{\pm 0.008}$	$0.113_{\pm 0.030}$	$0.057_{\pm 0.019}$
DeepSeek-V3	$0.320_{\pm 0.006}$	$0.567_{\pm 0.013}$	$0.726_{\pm 0.005}$	$0.355_{\pm 0.016}$	$0.303_{\pm 0.006}$	$0.225_{\pm 0.018}$	$0.226_{\pm 0.022}$	$0.149_{\pm 0.002}$	$0.019_{\pm 0.009}$	$0.000_{\pm 0.000}$
Gemini-2.5-Pro	$0.440_{\pm 0.011}$	$0.782_{\pm 0.003}$	$0.882_{\pm 0.003}$	$0.486_{\pm 0.011}$	$0.436_{\pm 0.017}$	$0.376_{\pm 0.020}$	$0.386_{\pm 0.028}$	$0.280_{\pm 0.025}$	$0.143_{\pm 0.019}$	$0.056_{\pm 0.010}$
GPT-5	$0.406_{\pm 0.008}$	$0.724_{\pm 0.007}$	$0.817_{\pm 0.017}$	$0.500_{\pm 0.027}$	$0.397_{\pm 0.002}$	$0.351_{\pm 0.006}$	$0.355_{\pm 0.009}$	$0.226_{\pm 0.002}$	$0.138_{\pm 0.026}$	$0.042_{\pm 0.031}$
MeasHalu-0.5B	$0.333_{\pm 0.006}$	$0.649_{\pm 0.004}$	$0.734_{\pm 0.013}$	$0.277_{\pm 0.013}$	$0.292_{\pm 0.014}$	$0.254_{\pm 0.013}$	$0.249_{\pm 0.010}$	$0.175_{\pm 0.011}$	$0.240_{\pm 0.293}$	$0.048_{\pm 0.002}$
w/o GRPO	$0.312_{\pm 0.008}$	$0.649_{\pm 0.001}$	$0.717_{\pm 0.019}$	$0.239_{\pm 0.006}$	$0.263_{\pm 0.004}$	$0.241_{\pm 0.017}$	$0.238_{\pm 0.017}$	$0.140_{\pm 0.004}$	$0.069_{\pm 0.009}$	$0.022_{\pm 0.007}$
MeasHalu-3B	$0.448_{\pm 0.007}$	$0.806_{\pm 0.012}$	$0.861_{\pm 0.010}$	$0.446_{\pm 0.010}$	$0.380_{\pm 0.005}$	$0.399_{\pm 0.010}$	$0.384_{\pm 0.009}$	$0.262_{\pm 0.008}$	$0.093_{\pm 0.004}$	$0.038_{\pm 0.010}$
w/o GRPO	$0.433_{\pm 0.010}$	$0.782_{\pm 0.014}$	$0.850_{\pm 0.003}$	$0.449_{\pm 0.009}$	$0.370_{\pm 0.006}$	$0.377_{\pm 0.017}$	$0.365_{\pm 0.019}$	$0.245_{\pm 0.002}$	$0.084_{\pm 0.006}$	$0.043_{\pm 0.015}$
MeasHalu-7B	$0.512_{\pm 0.004}$	$0.848_{\pm 0.001}$	$0.860_{\pm 0.008}$	$0.607_{\pm 0.006}$	$0.455_{\pm 0.008}$	$0.472_{\pm 0.009}$	$0.442_{\pm 0.012}$	$0.310_{\pm 0.005}$	$0.170_{\pm 0.005}$	$0.100_{\pm 0.009}$
w/o GRPO	$0.479_{\pm 0.005}$	$0.846_{\pm 0.002}$	$0.863_{\pm 0.004}$	$0.610_{\pm 0.015}$	$0.429_{\pm 0.004}$	$0.433_{\pm 0.016}$	$0.397_{\pm 0.016}$	$0.272_{\pm 0.012}$	$0.155_{\pm 0.012}$	$0.063_{\pm 0.010}$

Table 9: Experimental results over the MeasEval Benchmark.Comparing MeasHalu with competition leaders and rule/sentence-based LLM baselines. Top ranks are shaded orange (1st), yellow (2nd), and teal (3rd).

Model	Overall	Quantities			Entities		Properties		Qualifiers	
		Quantity	Unit	Modifier	ME	HasQuantity	MP	HasProperty	Qualifier	Qualifies
Rule-based Prompting										
GPT-5	$0.383_{\pm 0.004}$	$0.833_{\pm 0.021}$	$0.706_{\pm 0.017}$	$0.357_{\pm 0.031}$	$0.400_{\pm 0.004}$	$0.282_{\pm 0.003}$	$0.310_{\pm 0.020}$	$0.046_{\pm 0.011}$	$0.135_{\pm 0.017}$	$0.019_{\pm 0.012}$
DeepSeek-R1	$0.252_{\pm 0.006}$	$0.553_{\pm 0.016}$	$0.514_{\pm 0.046}$	$0.217_{\pm 0.018}$	$0.213_{\pm 0.005}$	$0.169_{\pm 0.009}$	$0.141_{\pm 0.015}$	$0.045_{\pm 0.006}$	$0.099_{\pm 0.009}$	$0.052_{\pm 0.018}$
DeepSeek-V3	$0.312_{\pm 0.003}$	$0.724_{\pm 0.014}$	$0.726_{\pm 0.011}$	$0.234_{\pm 0.009}$	$0.270_{\pm 0.014}$	$0.189_{\pm 0.003}$	$0.121_{\pm 0.005}$	$0.012_{\pm 0.003}$	$0.113_{\pm 0.018}$	$0.037_{\pm 0.006}$
Gemini-2.5-Pro	$0.386_{\pm 0.009}$	$0.766_{\pm 0.005}$	$0.707_{\pm 0.012}$	$0.444_{\pm 0.026}$	$0.351_{\pm 0.014}$	$0.312_{\pm 0.014}$	$0.291_{\pm 0.028}$	$0.151_{\pm 0.017}$	$0.140_{\pm 0.017}$	$0.055_{\pm 0.006}$
Qwen2.5-72b	$0.296_{\pm 0.007}$	$0.675_{\pm 0.012}$	$0.747_{\pm 0.010}$	$0.203_{\pm 0.025}$	$0.246_{\pm 0.015}$	$0.187_{\pm 0.015}$	$0.142_{\pm 0.004}$	$0.000_{\pm 0.000}$	$0.091_{\pm 0.014}$	$0.066_{\pm 0.009}$
Qwen2.5-7b	$0.181_{\pm 0.009}$	$0.501_{\pm 0.016}$	$0.353_{\pm 0.031}$	$0.148_{\pm 0.037}$	$0.108_{\pm 0.015}$	$0.078_{\pm 0.010}$	$0.069_{\pm 0.012}$	$0.001_{\pm 0.001}$	$0.038_{\pm 0.025}$	$0.006_{\pm 0.004}$
Sentence-based Prompting										
GPT-5	$0.402_{\pm 0.007}$	$0.750_{\pm 0.002}$	$0.759_{\pm 0.005}$	$0.458_{\pm 0.007}$	$0.435_{\pm 0.013}$	$0.303_{\pm 0.007}$	$0.303_{\pm 0.010}$	$0.239_{\pm 0.018}$	$0.100_{\pm 0.016}$	$0.070_{\pm 0.011}$
DeepSeek-R1	$0.324_{\pm 0.013}$	$0.622_{\pm 0.021}$	$0.648_{\pm 0.018}$	$0.255_{\pm 0.012}$	$0.299_{\pm 0.014}$	$0.235_{\pm 0.019}$	$0.215_{\pm 0.008}$	$0.146_{\pm 0.013}$	$0.075_{\pm 0.015}$	$0.069_{\pm 0.014}$
DeepSeek-V3	$0.299_{\pm 0.012}$	$0.521_{\pm 0.023}$	$0.565_{\pm 0.020}$	$0.229_{\pm 0.007}$	$0.296_{\pm 0.017}$	$0.229_{\pm 0.012}$	$0.209_{\pm 0.016}$	$0.155_{\pm 0.008}$	$0.053_{\pm 0.004}$	$0.048_{\pm 0.004}$
Gemini-2.5-Pro	$0.462_{\pm 0.007}$	$0.827_{\pm 0.009}$	$0.832_{\pm 0.007}$	$0.444_{\pm 0.010}$	$0.472_{\pm 0.014}$	$0.399_{\pm 0.012}$	$0.402_{\pm 0.006}$	$0.332_{\pm 0.015}$	$0.100_{\pm 0.002}$	$0.072_{\pm 0.003}$
Qwen2.5-72b	$0.202_{\pm 0.006}$	$0.343_{\pm 0.008}$	$0.383_{\pm 0.006}$	$0.183_{\pm 0.007}$	$0.207_{\pm 0.012}$	$0.145_{\pm 0.012}$	$0.127_{\pm 0.007}$	$0.106_{\pm 0.010}$	$0.048_{\pm 0.001}$	$0.050_{\pm 0.007}$
Qwen2.5-7b	$0.033_{\pm 0.003}$	$0.065_{\pm 0.002}$	$0.056_{\pm 0.005}$	$0.022_{\pm 0.005}$	$0.030_{\pm 0.001}$	$0.028_{\pm 0.006}$	$0.017_{\pm 0.005}$	$0.008_{\pm 0.003}$	$0.016_{\pm 0.007}$	$0.020_{\pm 0.007}$
MeasHalu-7B	$0.578_{\pm 0.002}$	$0.861_{\pm 0.003}$	$0.832_{\pm 0.012}$	$0.539_{\pm 0.005}$	$0.555_{\pm 0.006}$	$0.551_{\pm 0.008}$	$0.522_{\pm 0.006}$	$0.459_{\pm 0.007}$	$0.159_{\pm 0.021}$	$0.097_{\pm 0.004}$

Table 10: Experimental results over the MeasEval-Ext.

Inference Model	Context Source	Validity	BLEU		LEV	ROUGE		
			B-2	B-4	50%	R-1	R-2	R-L
Gemini-2.5-Pro	MeasHalu	**14.67**$_{\pm 2.31}$	**58.72**$_{\pm 0.23}$	**44.23**$_{\pm 0.29}$	**60.00**$_{\pm 2.00}$	**71.71**$_{\pm 0.18}$	**54.24**$_{\pm 0.32}$	**66.51**$_{\pm 0.26}$
	Gemini	12.67$_{\pm 1.53}$	58.26$_{\pm 0.60}$	43.99$_{\pm 0.67}$	59.67$_{\pm 3.06}$	71.28$_{\pm 0.39}$	54.21$_{\pm 0.55}$	66.49$_{\pm 0.44}$
	None	0.33$_{\pm 0.58}$	51.78$_{\pm 0.32}$	37.23$_{\pm 0.36}$	37.33$_{\pm 6.66}$	62.92$_{\pm 0.39}$	46.25$_{\pm 0.27}$	58.90$_{\pm 0.26}$
DeepSeek-R1	MeasHalu	**10.67**$_{\pm 1.53}$	**58.99**$_{\pm 0.37}$	**43.01**$_{\pm 0.56}$	**61.33**$_{\pm 2.89}$	**71.62**$_{\pm 0.19}$	**52.79**$_{\pm 0.12}$	**65.85**$_{\pm 0.40}$
	Gemini	10.33$_{\pm 0.58}$	58.28$_{\pm 0.73}$	42.37$_{\pm 0.61}$	56.67$_{\pm 2.89}$	71.13$_{\pm 0.38}$	52.31$_{\pm 0.36}$	65.65$_{\pm 0.31}$
	None	8.67$_{\pm 1.15}$	38.32$_{\pm 1.02}$	26.21$_{\pm 0.99}$	22.00$_{\pm 5.29}$	59.12$_{\pm 0.31}$	42.32$_{\pm 0.21}$	53.56$_{\pm 0.41}$
GPT-5	MeasHalu	**16.33**$_{\pm 2.52}$	**51.55**$_{\pm 0.29}$	37.43$_{\pm 0.52}$	**52.33**$_{\pm 3.51}$	**71.00**$_{\pm 0.39}$	51.62$_{\pm 0.27}$	**65.47**$_{\pm 0.32}$
	Gemini	13.33$_{\pm 2.08}$	50.61$_{\pm 0.72}$	36.56$_{\pm 0.81}$	48.33$_{\pm 4.04}$	70.73$_{\pm 0.80}$	51.31$_{\pm 0.64}$	65.06$_{\pm 0.66}$
	None	1.35$_{\pm 1.54}$	50.39$_{\pm 0.58}$	**39.39**$_{\pm 0.78}$	50.17$_{\pm 2.33}$	66.31$_{\pm 0.35}$	**52.53**$_{\pm 0.40}$	62.76$_{\pm 0.44}$

Table 11: Performance comparison on the OpenExp-Action-100 dataset. Models are provided with **MeasEval-formatted quantities and relations** generated by different sources (MeasHalu vs. Gemini) as context. The best scores **for each model** are highlighted in **bold**.

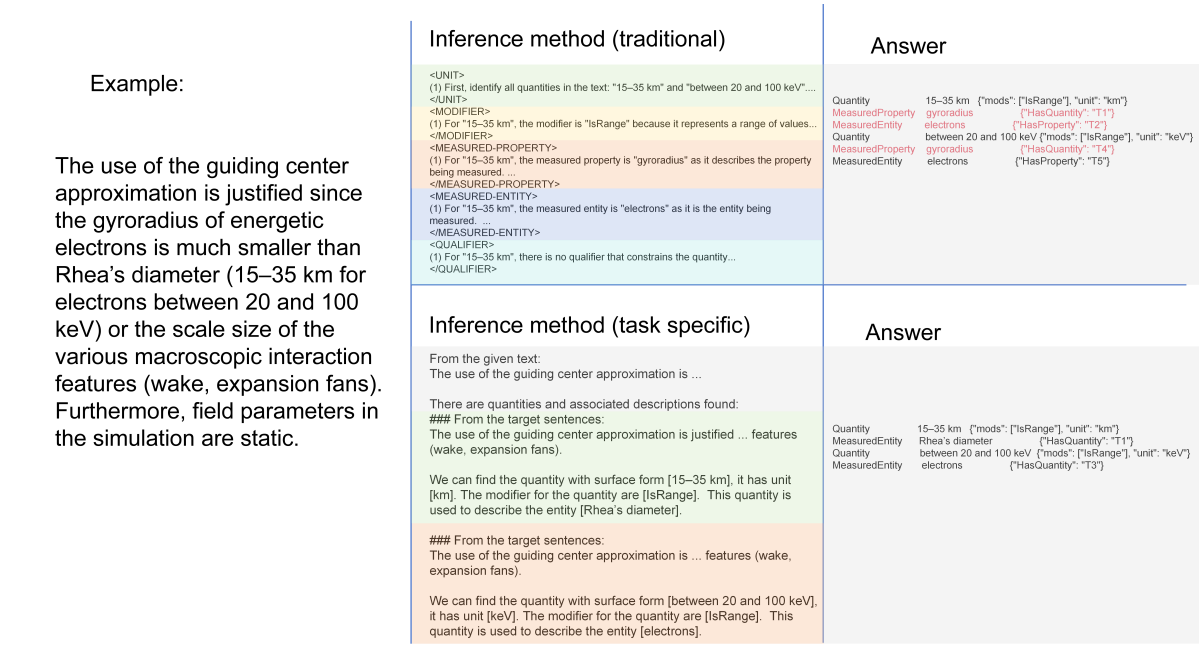

Figure 3: Comparison of Sentence-based and Rule-based Reasoning Approaches