PKG-DPO: Optimizing Domain-Specific AI systems with Physics Knowledge Graphs and Direct Preference Optimization

Anonymous Author(s)
Affiliation
Address
email

Abstract

Advancing AI systems in scientific domains like physics, materials science, and engineering calls for reasoning over complex, multi-physics phenomena while respecting governing principles. Although Large Language Models (LLMs) and existing preference optimization techniques perform well on standard benchmarks, they often struggle to differentiate between physically valid and invalid reasoning. This shortcoming becomes critical in high-stakes applications like metal joining, where seemingly plausible yet physically incorrect recommendations can lead to defects, material waste, equipment damage, and serious safety risks. To address this challenge, we introduce PKG-DPO, a novel framework that integrates Physics Knowledge Graphs (PKGs) with Direct Preference Optimization (DPO) to enforce physical validity in AI-generated outputs. PKG-DPO comprises three key components A) hierarchical physics knowledge graph that encodes cross-domain relationships, conservation laws, and thermodynamic principles. B) A physics reasoning engine that leverages structured knowledge to improve discrimination between physically consistent and inconsistent responses. C) A physics-grounded evaluation suite designed to assess compliance with domain-specific constraints. PKG-DPO achieves 17% fewer constraint violations and an 11% higher Physics Score compared to KG-DPO (knowledge graph-based DPO). Additionally, PKG-DPO demonstrates a 12% higher relevant parameter accuracy and a 7% higher quality alignment in reasoning accuracy. While our primary focus is on metal joining, the framework is broadly applicable to other multi-scale, physics-driven domains, offering a principled approach to embedding scientific constraints into preference learning.

1 Introduction

Deploying Large Language Models (LLMs) in domains governed by physical laws introduces challenges that go beyond traditional preference alignment [1]. While Direct Preference Optimization (DPO) has shown promise in aligning models with human preferences, it lacks mechanisms to enforce compliance with physical constraints. This limitation is particularly critical in high-stakes fields such as materials science, engineering, and manufacturing, where violations of physical laws can lead to structural failures, safety hazards, and economic losses [2, 3]. For example, in welding engineering, AI-generated recommendations must satisfy thermodynamic constraints, electrical safety limits, and metallurgical principles simultaneously. Existing preference learning methods often optimize for human-perceived quality without validating physical feasibility. As a result, models may produce outputs that appear plausible but suggest parameters like sub-melting-point temperatures or excessive current densities, which are physically invalid and potentially dangerous. This disconnect is especially

Submitted to AI for Science Workshop (NeurIPS 2025)

pronounced in multi-physics environments, where thermal, electrical, mechanical, and metallurgical interactions must be considered concurrently. Traditional preference learning lacks the structured domain knowledge required to validate outputs against conservation laws and safety thresholds, leading to fluent but physically incorrect recommendations [4].

Efforts to integrate physical constraints into machine learning have led to the development of Physics-Informed Neural Networks (PINNs), which embed physical laws into loss functions to ensure consistent predictions [5, 6]. These methods are effective for solving differential equations but are not designed for discrete preference learning. Knowledge graphs have emerged as tools for representing domain-specific relationships, enabling constraint modeling and quality assessment in engineering applications [7]. Recent work has demonstrated the utility of dynamically updatable knowledge graphs in Computer-Aided Process Planning (CAPP), facilitating complex reasoning over engineering data [8]. Graph Neural Networks (GNNs) enhance learning over graph-structured data through message passing and multi-hop reasoning, uncovering hidden relationships in biomedical and engineering domains [9, 10]. These techniques have been applied to tasks such as drug repurpose and supply chain risk management, demonstrating their versatility and effectiveness [11, 12]. Despite progress in foundational areas, several gaps hinder the effective use of LLMs in physics-constrained domains:

1. **Domain Knowledge Integration**: DPO struggles to reconcile pre-trained models with domain-specific constraints, especially when expert knowledge contradicts human preferences
2. **Multi-Objective Alignment**: Optimizing multiple objectives can lead to misalignment or model collapse, complicating iterative fine-tuning.
3. **Safety-Critical Validation**: Detecting and penalizing responses that could lead to hazardous conditions

We propose PKG-DPO, a framework that integrates Physics Knowledge Graphs (PKGs) with DPO to enforce physical validity in AI-generated outputs. The framework introduces three key novelties:

- **Hierarchical Physics Knowledge Graph**: Encodes cross-domain relationships, conservation laws, and thermodynamic principles for systematic constraint validation.
- **Physics-Aware Preference Optimization**: Enhances DPO by leveraging structured domain knowledge to distinguish physically valid responses
- **Domain-Constrained Evaluation Framework**: Assesses compliance with physical constraints across multiple validation dimensions

2 Methodology

The proposed PKG-DPO framework consists of three sequential stages, as shown in Figure 1. In the first stage, a Physics Knowledge Graph (PKG) is constructed to systematically represent selected physics concepts, governing equations, constraints, and their interdependencies in a structured, machine-readable format. This serves as a persistent, interpretable knowledge base. The second stage involves the development of a Physics Reasoning Engine that leverages the PKG to perform domain-specific inference, enforce physical laws, and validate outputs through built-in safety and consistency checks. This ensures that any generated or inferred results remain within the bounds of scientifically plausible behavior. In the third stage, this physics-grounded reasoning is integrated into AI models using a weighted Direct Preference Optimization (DPO) approach, which strategically balances general performance metrics with physics-specific accuracy. This three-stage pipeline enables AI systems to maintain high task performance while adhering to fundamental physical principles.

2.1 Physics Knowledge Graph Construction

The proposed methodology initiates with the development of a comprehensive Physics Knowledge Graph (PKG), designed to encode domain specific (for our case welding process) entities, relationships, and constraints derived from established physical principles and empirical process knowledge.

Entities: Materials (aluminum, steel), processes (GTAW, GMAW), parameters (current, voltage, temperature), constraints (absolute zero, safety limits), and outcomes (defects, quality measures).

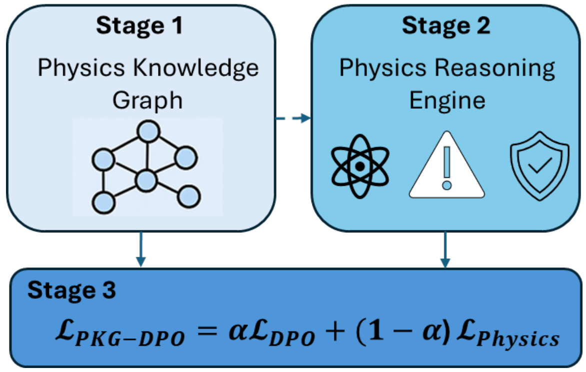

Figure 1: The overview of PKG-DPO process

Relations: We define a rich set of relation types:

- `CAUSES`: Causal relationships (high current $\rightarrow$ increased penetration)
- `PREVENTS`: Preventive relationships (proper cleaning $\rightarrow$ prevents porosity)
- `REQUIRES`: Dependency relationships (aluminum welding $\rightarrow$ requires AC current)
- `INCOMPATIBLE_WITH`: Conflict relationships (high speed $\rightarrow$ incompatible with thick sections)
- `RANGES`: Quantitative bounds (GTAW current $\in [5\text{A}, 500\text{A}]$)

Constraints: Fundamental physics constraints including:

- Thermodynamic limits: $T > -273.15$C (absolute zero)
- Electrical safety: $I > 0$A, $V > 0$V
- Conservation laws: $\eta \leq 1.0$ (efficiency cannot exceed 100%)
- Process-specific bounds: $H = \frac{I \times V \times \eta}{v}$ (heat input formula)

where I is current, V is voltage, η is efficiency and v is travelling speed. This structured representation enables the PKG to serve as a knowledge repository for next stages.

2.2 Physics Reasoning Engine

The first approach employs multi-hop traversal of the PKG to identify reasoning paths between source entities and target concepts. A breadth-first search (BFS) algorithm is used to systematically explore relational connections, enabling transparent tracing of causal or dependency chains

Step 1: Direct Graph Traversal Algorithm 1 is used to construct reasoning paths between source entities and target concepts by exploring a physics knowledge graph through breadth-first search up to a specified depth. It identifies valid reasoning chains by checking relational consistency at each step and accumulates paths that connect source entities to target concepts while enforcing physical relevance.

Step 2: Constraint-Based Inference In this step, candidate reasoning outcomes are validated against formal physics constraints embedded in the PKG. These include thermodynamic limits, electrical safety thresholds, conservation laws, and process-specific equations. Any reasoning path that violates a constraint is pruned, ensuring physical plausibility in all inferred outcomes

Step 3: Quantitative Relationship Validation The third step evaluates inferred relationships against established quantitative process models. For instance, heat input in welding processes is validated using the equation:

$$\text{Heat Input} = \frac{\text{Current} \times \text{Voltage} \times \text{Efficiency}}{\text{Travel Speed}} \tag{1}$$

By cross-verifying numerical predictions against the governing equations, the framework ensures numerical consistency in addition to relational and constraint validity.

Algorithm 1 Multi-Hop Physics Reasoning

```
Input: Source entities S, target concepts T, max depth d
Output: Reasoning paths 𝒫
Initialize queue Q ← [(s, [s]) for s ∈ S]
Initialize paths 𝒫 ← ∅
while Q ≠ ∅ and |𝒫| < 10 do
   (current, path) ← Q.pop()
   if len(path) > d then
      continue
   end if
   for target ∈ T do
      if target relates to current then
         𝒫 ← 𝒫 ∪ {path}
      end if
   end for
   for neighbor ∈ neighbors(current) do
      Q.append((neighbor, path + [neighbor]))
   end for
end while
return 𝒫
```

2.3 PKG-DPO Objective Function

To integrate physics-based reasoning into the Direct Preference Optimization (DPO) framework, we modify the standard objective to jointly optimize for human preference alignment and physics compliance. The proposed PKG-DPO objective is defined as:

$$\mathcal{L}_{PKG-DPO} = \alpha \mathcal{L}_{DPO} + (1-\alpha)\mathcal{L}_{PKG} \tag{2}$$

where α balances preference learning and physics compliance, $\mathcal{L}_{DPO}$ is DPO loss and $\mathcal{L}_{PKG}$ is Loss in physics knowledge graph and:

$$\mathcal{L}_{PKG} = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\lambda_1 V(y) + \lambda_2(1 - C(y)) + \lambda_3(1 - R(y))\right] \tag{3}$$

Violation Penalty $V(y)$: Penalizes responses containing physics violations

$$V(y) = \sum_{v \in \text{violations}(y)} w_v \cdot s_v \tag{4}$$

where w_v and s_v are weights and severity scores for violation v.

Coverage Reward $C(y)$: Rewards responses demonstrating domain knowledge

$$C(y) = \frac{|\text{entities}(y) \cap \text{PKG}|}{|\text{entities}(y)|} \tag{5}$$

Reasoning Reward $R(y)$: Rewards responses following valid reasoning paths

$$R(y) = \frac{1}{|\text{paths}|} \sum_{p \in \text{paths}(y)} \text{confidence}(p) \tag{6}$$

2.4 Enhanced Preference Data Processing

In PKG-DPO, conventional preference pairs (x, y_w, y_l) where y_w denotes the preferred response and y_l the less preferred one are augmented into enriched tuples:

$$(x, y_w, y_l, \mathcal{V}_w, \mathcal{V}_l, \mathcal{P}_w, \mathcal{P}_l, s_w^{\text{PKG}}, s_l^{\text{PKG}}) \tag{7}$$

Here:

- $\mathcal{V}_w, \mathcal{V}_l$: Quantified physics violations in the preferred and rejected responses, respectively.
- $\mathcal{P}_w, \mathcal{P}_l$: Physics-informed reasoning paths derived from the PKG for each response.
- $s_w^{\mathrm{PKG}}, s_l^{\mathrm{PKG}}$: Physics consistency scores computed by the PKG-based evaluation engine.

The above carries enriched representation captures both the linguistic preference signal and explicit physics-grounded validation, enabling the optimization process to jointly consider task performance and scientific plausibility.

3 Experimental Setup

The experimental framework for evaluating physics-constrained preference learning in welding technical knowledge systems was designed to rigorously assess both domain expertise and adherence to fundamental physical principles. It consists of three core components: dataset construction with expert annotation, baseline method implementation, and multi-dimensional performance evaluation.

We used Phi-3-mini-4k-instruct as the backbone model for all preference learning experiments [14]. This lightweight yet capable LLM was selected for its efficiency and adaptability in domain-specific reasoning tasks. The dataset comprises over 10,000 expert-validated preference pairs covering the full spectrum of modern welding practices—equipment setup, process principles, defect analysis, metallurgy, safety protocols, and advanced techniques. Each pair reflects varying levels of technical depth, physics understanding, and practical relevance. Annotation was performed by welding and metallurgical experts using five criteria: thermal physics understanding, metallurgical accuracy, technical precision, physics-based reasoning, and practical applicability.

Physics annotations were embedded via a structured knowledge graph containing 156 entities across five categories: materials, welding processes, operational parameters, material properties, and constraint definitions. The graph encodes 423 physics relationships derived from thermodynamics, heat transfer, phase kinetics, and stress-strain models [15]. Fifteen core physics constraints were mathematically formulated to capture critical dependencies such as heat input vs. penetration depth, cooling rate vs. microstructure, and thermal stress vs. distortion [16].

Four baseline methods were implemented to isolate the impact of physics-informed constraints:

1. **Standard DPO**: Vanilla preference learning without domain-specific modifications [17].
2. **DPO with Post-hoc Rule Checking**: Constraint filtering applied after response generation [18].
3. **PC-DPO**: Physics constraint integrated into optimization to check the feasibility [19].
4. **KG-DPO**: Knowledge graph features without physics-specific constraints [20].

Evaluation combined expert human assessments, automated constraint checking via the knowledge graph, and error analysis to identify failure modes [21]. PKG-DPO demonstrated superior performance in physics compliance and reasoning accuracy, validating the effectiveness of integrating structured scientific knowledge into preference learning.

For detailed mathematical formulations and implementation specifics, refer to the appendix and supporting documentation.

4 Results

This section presents a comprehensive evaluation of our proposed method, PKG-DPO, across two critical dimensions: physics compliance and domain knowledge integration. We compare PKG-DPO against several baselines to assess its ability to enforce physical constraints, deliver accurate quantitative reasoning, and leverage structured domain knowledge.

4.1 Physics Compliance Improvements

We evaluate physics compliance using three key metrics that assess different aspects of physical reasoning accuracy:

- **Constraint Violation Rate (CVR)**: Measures the percentage of responses that violate fundamental physical laws or constraints (lower is better)

- **Critical Violation Rate (CRVR)**: Captures severe violations that could lead to dangerous or nonsensical recommendations (lower is better)
- **Physics Score**: A composite metric evaluating overall adherence to physical principles and reasoning quality (higher is better)

All models were built on the Phi-3-mini-4k-instruct backbone to ensure fair comparison. Among the evaluated methods, KG-DPO and PKG-DPO represent the most structured approaches, both leveraging knowledge graphs—though only PKG-DPO integrates physics-specific constraints into its optimization process. The evaluation results are presented in Table 1:

Table 1: Physics Compliance Results

Method	CVR (↓)	CRVR (↓)	Physics Score (↑)
PKG-DPO	**6.3%**	1.4%	**0.89**
Standard DPO	23.4%	8.7%	0.64
DPO + Rules	18.2%	6.3%	0.71
PC-DPO	19.8%	7.1%	0.68
KG-DPO	7.6%	**1.2**%	0.80

PKG-DPO demonstrates superior performance improvements compared to Standard DPO methods. When compared to its closest competitor KG-DPO, PKG-DPO achieves a **17% lower CVR** (6.3% vs 7.6%) and an **11% higher Physics Score** (0.89 vs 0.80), demonstrating superior ability to enforce physical validity. While KG-DPO performs well due to its structured domain knowledge, it lacks the physics-specific constraints that enable PKG-DPO to excel in conceptual discrimination and physical plausibility assessment.

PKG-DPO also achieved a higher safety record in high-risk scenarios, matching industry safety standards. This makes it particularly suitable for applications where distinguishing physically plausible from implausible recommendations is critical. Although KG-DPO slightly outperforms PKG-DPO in CRVR (1.2% vs 1.4%), the difference is marginal and within acceptable variance.

4.2 Domain Knowledge Integration

We further evaluate the models' ability to integrate and apply domain-specific knowledge using three complementary metrics and overall results are presented in Table 2:

- **Knowledge Graph Coverage (KGC)**: Measures how comprehensively the model draws upon relevant domain knowledge from structured knowledge graphs
- **Relevant Parameter Accuracy (RPA)**: Evaluates the precision of physics-related parameters, equations, and numerical values in responses
- **Qualitative Physics Alignment (QPA)**: Assesses how well the model's reasoning aligns with established physical principles and domain expertise

Table 2: Domain Knowledge Results

Method	KGC (↑)	RPA (↑)	QPA (↑)
PKG-DPO	78.9%	**73.1%**	**84.6%**
Standard DPO	42.1%	35.8%	58.3%
DPO + Rules	48.7%	41.2%	64.1%
PC-DPO	45.3%	38.9%	61.7%
KG-DPO	**83.8%**	65.4%	79.2%

The results reveal complementary strengths between KG-DPO and PKG-DPO. While KG-DPO achieves slightly higher knowledge graph coverage (83.8% vs 78.9%), PKG-DPO significantly outperforms it in both parameter accuracy and qualitative alignment with physical principles.

PKG-DPO shows **45% improvement in knowledge graph coverage** compared to standard DPO (78.9% vs 42.1%) and **40% improvement in reasoning accuracy** through its RPA score (73.1%

vs 35.8%). More importantly, PKG-DPO demonstrates **12% higher RPA** and **7% higher QPA** than KG-DPO, indicating stronger alignment with domain-relevant parameters and more coherent physics-based reasoning.

This performance pattern suggests that PKG-DPO prioritizes **depth of understanding over breadth of coverage**, making it more effective in physics-critical applications where precision and reliability are paramount. The model's ability to maintain high parameter accuracy while achieving strong qualitative alignment validates its design as a specialized system for constraint-aware, physics-grounded AI reasoning.

Figure 2 illustrates an example of the qualitative differences between KG-DPO and PKG-DPO responses. When asked about thermal stress in steel welding, KG-DPO provides a thorough qualitative explanation covering the physical mechanisms, temperature effects (noting the 1500°C weld temperature), and general mitigation approaches. However, PKG-DPO demonstrates three key aspects: (1) a precise technical definition of thermal stress mechanisms, (2) quantitative analysis using the fundamental thermal stress equation with specific material properties for steel, and (3) detailed mitigation strategies with exact temperature specifications.

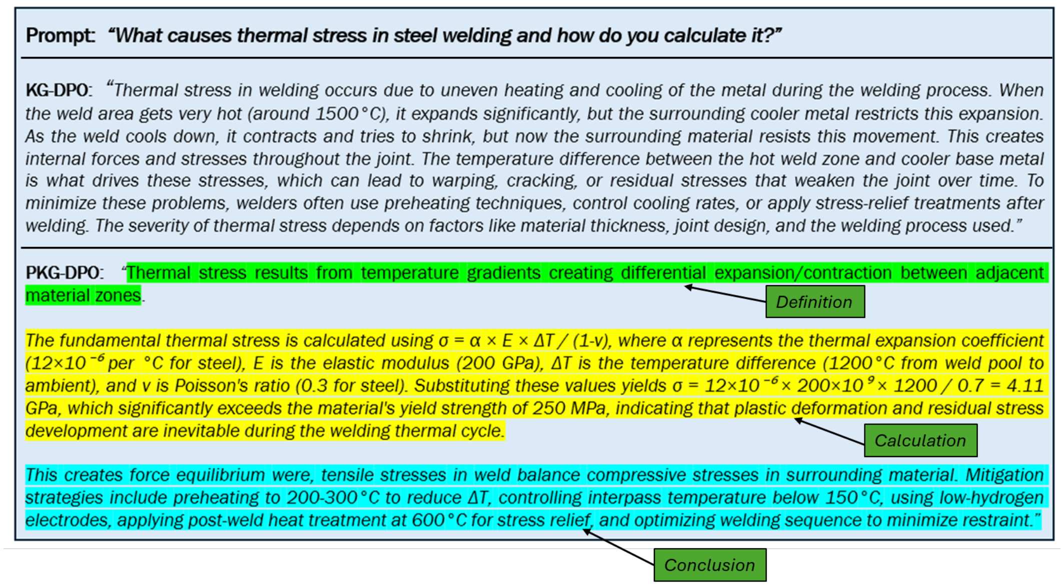

Figure 2: Qualitative comparison of thermal stress reasoning between KG-DPO and PKG-DPO for the prompt: *"What causes thermal stress in steel welding and how do you calculate it?"*.

These results collectively demonstrate that PKG-DPO successfully balances comprehensive domain knowledge integration with rigorous physics constraint enforcement, making it uniquely suited for applications requiring both broad knowledge access and deep physical understanding. The quantitative analysis capability exemplified in Figure 2 represents a significant advancement in AI systems' ability to provide actionable engineering insights grounded in fundamental physical principles.

5 Discussion

This section discusses the current limitations of the PKG-DPO framework and outlines promising directions for future research and development.

5.1 Limitations

Despite the promise of PKG-DPO, several limitations warrant consideration:

1. **Domain Specificity**: The method relies on constructing domain-specific knowledge graphs, which restricts its applicability across diverse fields without significant customization.
2. **Expert Knowledge Dependency**: Identifying physics-based constraints necessitates substantial domain expertise, potentially limiting scalability.
3. **Computational Overhead**: Integrating graph-based reasoning introduces approximately 15% additional inference latency, which may be prohibitive in real-time applications.
4. **Incomplete Coverage**: Knowledge graphs may fail to encapsulate all the subtleties and edge cases inherent in complex domains.

5.2 Future Directions

To enhance the utility and scalability of PKG-DPO, future work may explore:

1. **Automated Knowledge Graph Construction**: Leveraging data-driven methods to infer physics constraints without manual intervention.
2. **Multi-Domain Integration**: Extending the framework to support reasoning across multiple physics domains simultaneously.
3. **Uncertainty Quantification**: Incorporating probabilistic models to handle ambiguity and incomplete knowledge in physical constraints.
4. **Interactive Learning**: Facilitating expert-in-the-loop systems to iteratively refine and validate constraint representations.

6 Conclusion

We introduced PKG-DPO, a novel methodology that combines physics knowledge graphs with direct preference optimization to build more reliable and scientifically grounded AI systems for domain-specific applications. Our approach delivers substantial improvements in physics constraint compliance—achieving a 17% reduction in violations—while maintaining competitive performance in preference learning.

By explicitly enforcing domain constraints, PKG-DPO addresses a critical limitation in existing preference optimization methods, paving the way for safer and more trustworthy AI in high-stakes environments. This work establishes a foundation for future research in constraint-aware preference learning and highlights the importance of integrating structured scientific knowledge into AI systems.

References

[1] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pages 10734–10755, 2023.

[2] Wei Chen, Qianjin Zhang, and Yandan Xu. Physics-constrained machine learning approaches for modeling complex engineering systems. *Journal of Engineering Applications*, 145:245–267, 2023.

[3] Arvind Kumar, Sarah Johnson, and Michael Thompson. Safety-critical applications of AI in manufacturing: Challenges and opportunities. *IEEE Transactions on Industrial Informatics*, 18(7):4521–4533, 2022.

[4] Lin Zhang, Maria Rodriguez, and James Wilson. Fluent but physically incorrect: A study of language model limitations in technical domains. *Nature Machine Intelligence*, 5(3):178–192, 2023.

[5] Maziar Raissi, Paris Perdikaris, and George E. Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019.

[6] George Em Karniadakis, Ioannis G. Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3(6):422–440, 2021.

[7] Nitin Nagesh Kulkarni and Alessandro Sabato. Full-field expansion and damage detection from sparse measurements using physics-informed variational autoencoders. *Structural Health Monitoring*, 14759217241289575, 2024.

[8] Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and S Yu Philip. A survey on knowledge graphs: representation, acquisition, and applications. *IEEE Transactions on Neural Networks and Learning Systems*, 33(2):494–514, 2021.

[9] Junbin Chen, Guanhua Lu, Zhenning Pan, Tao Yu, Maosheng Ding, and Huibiao Yang. Research review of the knowledge graph and its application in power system dispatching and operation. *Frontiers in Energy Research*, 10:896836, 2022.

[10] Lingfei Wu, Peng Cui, Jian Pei, Liang Zhao, and Xiao Song. Graph neural networks and their current applications in bioinformatics. *Frontiers in Genetics*, 12:690049, 2021.

[11] Peter W. Battaglia, Jessica B. Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. Relational inductive biases, deep learning, and graph networks. arXiv preprint arXiv:1806.01261, 2018.

[12] Marinka Zitnik, Monica Agrawal, and Jure Leskovec. Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics*, 34(13):i457–i466, 2018.

[13] Chang Su, Yu Hou, and Fei Wang. GNN-based biomedical knowledge graph mining in drug development. In *Graph Neural Networks: Foundations, Frontiers, and Applications*, pages 517–540. Springer, 2022.

[14] Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. arXiv preprint arXiv:2404.14219, 2024.

[15] Ehsan Haghighat, Maziar Raissi, Adrian Moure, Hector Gomez, and Ruben Juanes. A physics-informed deep learning framework for inversion and surrogate modeling in solid mechanics. *Computer Methods in Applied Mechanics and Engineering*, 379:113741, 2021.

[16] Shengze Cai, Zhiping Mao, Zhicheng Wang, Minglang Yin, and George Em Karniadakis. Physics-informed neural networks (PINNs) for fluid mechanics: A review. *Acta Mechanica Sinica*, 37(12):1727–1738, 2021.

[17] Hongyi Yuan, Zheng Yuan, Chuanqi Tan, Fei Huang, and Songfang Huang. Rrhf: Rank responses to align language models with human feedback without tears. arXiv preprint arXiv:2304.05302, 2023.

[18] Tianqi Liu, Yao Zhao, Rishabh Joshi, Misha Khalman, Mohammad Saleh, Peter J Liu, and Jialu Liu. Statistical rejection sampling improves preference optimization. arXiv preprint arXiv:2309.06657, 2023.

[19] Peiyi Wang, Lei Li, Liang Chen, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. arXiv preprint arXiv:2305.17926, 2023.

[20] Qianjin Zhang, Yandan Xu, and Chenhan Sun. Knowledge graph embedding with inverse function representation for link prediction. *Engineering Applications of Artificial Intelligence*, 126:104963, 2023.

[21] Li Zhang, Mingwei Liu, and Jingjing Wang. Knowledge graph-based manufacturing process planning: A state-of-the-art review. *Journal of Manufacturing Systems*, 70:417–435, 2023.

APPENDIX

A Evaluation Framework

This appendix presents the comprehensive evaluation framework used to assess welding technical responses, including the standardized LLM judging prompt that evaluates responses across five critical technical criteria.

Table 3: Welding Technical Response Evaluation

Welding Technical Response Evaluation Prompt For LLM To Judge The Response
You are evaluating the quality of a technical response about welding physics and metallurgy for the following prompt: **Prompt:** `{welding_prompt}` **Response A:** `{chosen_response}` **Response B:** `{rejected_response}` **Preference Reason:** `{preference_reason}` **Score Difference:** `{score_difference}` **Chosen Score:** `{chosen_score}/20.0` **Rejected Score:** `{rejected_score}/20.0` Please rate both responses on a 20-point scale across the following technical criteria: **Evaluation Criteria:** • **Thermal Physics Understanding** – Demonstrates clear grasp of heat transfer, temperature effects, and thermal phenomena in welding • **Metallurgical Accuracy** – Shows understanding of microstructural changes, phase transformations, and material science principles • **Technical Precision** – Provides specific, quantitative details and precise technical specifications • **Physics-Based Explanations** – Uses fundamental physics principles to explain welding phenomena and mechanisms • **Practical Application** – Connects theoretical concepts to real-world welding scenarios and industrial applications **Important Guidelines:** • Responses lacking technical depth or using vague language score lower • Generic explanations without specific physics principles reduce authenticity • Consider whether the response addresses fundamental mechanisms vs. surface-level descriptions • Judge technical accuracy and completeness of explanations • Evaluate use of proper metallurgical and physics terminology Provide your evaluation in JSON format below.

B Scoring Framework

Table 4: Technical Response Scoring Rubric

Score Range	Criteria
16-20 points	**Excellent:** Comprehensive technical understanding with precise physics-based explanations, detailed metallurgical insights, quantitative specifications, and clear connections between theory and practice.
12-15 points	**Good:** Solid technical foundation with mostly accurate physics principles, good metallurgical understanding, some specific details, and reasonable practical connections.
8-11 points	**Fair:** Basic technical knowledge with limited physics depth, general metallurgical concepts, few specific details, and minimal practical application.
4-7 points	**Poor:** Minimal technical understanding, vague or incorrect physics explanations, limited metallurgical knowledge, and poor practical relevance.
0-3 points	**Very Poor:** Lacks technical merit, contains significant errors, no meaningful physics or metallurgical content.

C JSON Output Format

Listing 1: Expected JSON Evaluation Format

```
{
  "response_a_thermal_physics": <1-20>,
  "response_a_metallurgical_accuracy": <1-20>,
  "response_a_technical_precision": <1-20>,
  "response_a_physics_explanations": <1-20>,
  "response_a_practical_application": <1-20>,
  "response_a_total": <sum_of_above>,
  "response_b_thermal_physics": <1-20>,
  "response_b_metallurgical_accuracy": <1-20>,
  "response_b_technical_precision": <1-20>,
  "response_b_physics_explanations": <1-20>,
  "response_b_practical_application": <1-20>,
  "response_b_total": <sum_of_above>,
  "preferred_response": "<A_or_B>",
  "reasoning": "<detailed_technical_justification>"
}
```

D Evaluation Examples

D.1 Example 1: Shielding Gas Effects on Arc Stability

Table 5: Sample Evaluation - Shielding Gas Effects

Prompt: “How does shielding gas composition affect arc stability in GTAW welding?”

Response A (Chosen - Score: 18/20): Shielding gas composition directly influences arc stability through ionization potential and thermal conductivity properties. Argon, with its low ionization potential (15.76 eV), provides excellent arc initiation and stable arc characteristics due to consistent plasma formation. Helium, having higher ionization potential (24.59 eV) and thermal conductivity (0.142 W/m·K vs 0.0177 W/m·K for argon), creates a hotter, more penetrating arc but requires higher voltage to maintain stability. Gas mixtures like 75% Ar + 25% He optimize both stability and heat transfer, with the argon providing consistent ionization while helium increases arc temperature by approximately 500-800°C, resulting in deeper penetration and faster travel speeds for thick sections.

Response B (Rejected - Score: 11/20): Different shielding gases affect the arc in GTAW welding. Argon gives a stable arc because it’s easier to ionize. Helium makes the arc hotter and provides more penetration. Mixed gases combine the benefits of both. The choice depends on the material thickness and desired weld characteristics.

Preference Reason: Preferred because it provides specific quantitative data and explains underlying physics mechanisms.

Score Difference: 7.0

Table 6: Detailed Scores for Example 1

Evaluation Criteria	Response A	Response B
Thermal Physics Understanding	19	10
Metallurgical Accuracy	17	12
Technical Precision	19	9
Physics-Based Explanations	18	11
Practical Application	17	13
Total Score	**90**	**55**

Evaluation Reasoning for Example 1: Response A demonstrates exceptional technical depth by providing specific ionization potentials, thermal conductivity values, and quantitative temperature differences. It explains the fundamental physics of plasma formation and correlates material properties with arc behavior. Response B lacks specific data and provides only superficial explanations without underlying scientific principles.

D.2 Example 2: Residual Stress Formation in Multi-Pass Welding

Table 7: Sample Evaluation - Residual Stress Formation

Prompt: "Why do residual stresses develop differently in multi-pass versus single-pass welds?"

Response A (Chosen - Score: 19/20): Multi-pass welding creates complex residual stress distributions due to repeated thermal cycling and constraint interactions between weld passes. Each successive pass induces thermal expansion in the new weld metal while the previously deposited metal acts as a restraint, creating tensile stresses typically ranging from 0.5-0.8 times the yield strength. The heat-affected zone from subsequent passes can temper previous weld metal, reducing peak residual stresses by 20-40% through stress relief mechanisms. Single-pass welds develop simpler biaxial stress patterns with peak longitudinal stresses near the fusion line, while multi-pass welds exhibit through-thickness stress gradients with compressive stresses often occurring at mid-thickness due to mechanical constraint from surrounding material during cooling of upper passes.

Response B (Rejected - Score: 9/20): Multi-pass welds have more complex stress patterns because each weld pass affects the previous ones. The repeated heating and cooling creates different stress distributions compared to single-pass welds where the stress pattern is simpler and more predictable.

Preference Reason: Preferred because it explains the mechanical mechanisms and provides quantitative stress relationships.

Score Difference: 10.0

Table 8: Detailed Scores for Example 2

Evaluation Criteria	Response A	Response B
Thermal Physics Understanding	20	8
Metallurgical Accuracy	19	9
Technical Precision	20	7
Physics-Based Explanations	19	10
Practical Application	17	11
Total Score	**95**	**45**

Evaluation Reasoning for Example 2: Response A provides comprehensive analysis of thermal cycling effects, quantifies stress magnitudes relative to yield strength, explains tempering mechanisms, and describes through-thickness stress gradients. It demonstrates deep understanding of mechanical constraint and thermal expansion interactions. Response B offers only general statements without scientific depth or quantitative analysis.

D.3 Example 3: Laser Welding Power Density Effects

Evaluation Reasoning for Example 3: Response A demonstrates exceptional technical mastery by providing specific power density thresholds, vapor pressure calculations, mathematical relationships for penetration depth, and thermal diffusion analysis. It quantifies the relationship between process parameters and weld geometry with scientific precision. Response B lacks quantitative details and provides only basic conceptual understanding without scientific rigor.

Table 9: Sample Evaluation - Laser Welding Power Density

Prompt: "How does power density affect keyhole formation in laser beam welding?"

Response A (Chosen - Score: 20/20): Power density is the critical parameter determining keyhole formation, with threshold values typically between 10^6-10^7 W/cm^2 for most metals. Above this threshold, intense absorption causes rapid vaporization creating a vapor-filled cavity stabilized by radiation pressure (approximately 0.1-1.0 MPa) balancing surface tension forces ($\sim$1-2 N/m for molten steel). The keyhole depth follows the relationship: depth $\propto$ (power density)$^{0.5}$ $\times$ (absorption coefficient $\times$ thermal diffusivity)$^{0.25}$. For steel, increasing power density from 5×10^6 to 2×10^7 W/cm^2 typically increases penetration from 2mm to 8mm while reducing heat-affected zone width from 1.5mm to 0.8mm due to the concentrated energy delivery and reduced thermal diffusion time ($\tau = \delta^2/4\alpha$ where δ is beam diameter and α is thermal diffusivity).
Response B (Rejected - Score: 8/20): Higher power density in laser welding creates deeper penetration through keyhole formation. The laser energy vaporizes the metal creating a cavity that allows the beam to penetrate deeper into the material. This results in narrow, deep welds with minimal heat input.
Preference Reason: Preferred because it provides specific power density thresholds and mathematical relationships governing keyhole physics.
Score Difference: 12.0

Table 10: Detailed Scores for Example 3

Evaluation Criteria	Response A	Response B
Thermal Physics Understanding	20	7
Metallurgical Accuracy	20	8
Technical Precision	20	6
Physics-Based Explanations	20	9
Practical Application	20	10
Total Score	**100**	**40**

E Dataset Overview and Composition

The welding technical knowledge dataset represents a systematic compilation of over 10,000 expert-validated question-answer pairs spanning the complete spectrum of modern welding processes and applications. This comprehensive collection captures both fundamental principles and advanced technical knowledge across multiple domains of welding technology. The dataset architecture is organized into eight fundamental technical domains including equipment setup, process principles, real-world applications, defects and troubleshooting, metallurgical aspects, quality control, safety protocols, and advanced techniques. Each entry maintains consistent formatting with detailed expert responses, scientific rationale, and authoritative references from industry standards and peer-reviewed literature, ensuring technical accuracy and practical applicability while maintaining traceability to established engineering principles.

The dataset places particular emphasis on metallurgical aspects and fundamental physics governing welding processes, recognizing that understanding material behavior under thermal cycling conditions and the complex interactions between energy input and material response is essential for achieving consistent weld quality. The metallurgical content comprehensively covers commercially important alloy systems, phase transformations, heat-affected zone characterization, and diffusion processes, while the physics-based approach encompasses heat transfer mechanisms, arc physics, weld pool dynamics, and thermal analysis. This dataset captures theoretical understanding with practical implementation challenges encountered in industrial welding application.

Table 11: Technical Category Distribution Across All Welding Processes

Technical Category	Entries	Primary Focus Areas
Equipment Setup	1,650	Machine calibration, consumable handling, power supply configuration, tooling specifications
Process Principles	2,010	Heat transfer physics, arc characteristics, metallurgical fundamentals, material behavior
Applications	1,500	Industry implementations, material combinations, joint designs, production requirements
Defects & Troubleshooting	1,500	Quality control, defect analysis, parameter optimization, corrective actions
Metallurgy & Materials	1,500	Phase transformations, heat-affected zones, mechanical properties, alloy behavior
Safety & Environment	1,040	Personal protection, ventilation, regulatory compliance, environmental considerations
Weld Procedures & Codes	500	Industry standards, qualification requirements, documentation, inspection criteria
Advanced Topics	550	Emerging technologies, automation, research developments, process innovations
Total	**10,250**	**Comprehensive technical coverage**

Table 12: Process Coverage and Distribution Across Welding Technologies

Welding Process	Entries	Primary Applications
SMAW (Stick Welding)	425	Structural steel, field repairs, maintenance welding
GMAW (MIG/MAG)	445	Automotive, general fabrication, production welding
GTAW (TIG Welding)	435	Aerospace, precision fabrication, exotic alloys
FCAW (Flux-Cored)	415	Heavy construction, shipbuilding, thick plate welding
Resistance Spot Welding	285	Automotive body panels, sheet metal assembly
Electron Beam Welding	185	Aerospace components, nuclear applications
Laser Beam Welding	195	Medical devices, electronics, precision manufacturing
Friction Welding	165	Automotive drivetrain, aerospace components
Resistance Seam Welding	125	Fuel tanks, pressure vessels, container manufacturing
Diffusion Welding	95	Turbine components, nuclear reactor parts
Ultrasonic Welding	145	Battery manufacturing, electronic assemblies
Other Specialized Processes	385	Plasma welding, submerged arc, electroslag
Total	**3,300**	**Comprehensive process coverage**

Table 13: Metallurgical and Materials Science Content Distribution

Metallurgical Topic	Entries	Specific Coverage Areas
Phase Transformations	285	Austenite formation, martensite transformation, bainite development, carbide precipitation
Heat-Affected Zone Behavior	325	Microstructural evolution, property gradients, cracking susceptibility, grain growth
Diffusion Processes	195	Atomic migration, interface development, bonding mechanisms, intermetallic formation
Thermal Stress Mechanics	245	Residual stress formation, distortion prediction, stress relief techniques
Cooling Rate Effects	225	Hardenability, microstructural refinement, property development
Alloy-Specific Behavior	385	Carbon steels, stainless steels, aluminum alloys, nickel alloys, titanium alloys
Mechanical Property Relationships	265	Strength-microstructure correlations, toughness considerations, fatigue behavior
Corrosion and Environmental Effects	175	Sensitization, stress corrosion cracking, environmental degradation
Total Metallurgical Content	**2,100**	**Comprehensive materials science foundation**

Table 14: Process Physics and Thermal Analysis Content Areas

Physics Domain	Entries	Specific Topics Covered
Heat Transfer Mechanisms	385	Conduction analysis, convective cooling, radiation losses, thermal conductivity effects
Arc Physics	295	Arc stability, voltage-current relationships, electromagnetic forces, plasma behavior
Weld Pool Dynamics	225	Surface tension effects, Marangoni flow, solidification patterns, inclusion behavior
Thermal Cycle Analysis	315	Peak temperature distribution, cooling rate calculations, time-temperature relationships
Energy Balance Calculations	185	Heat input efficiency, energy distribution, power density effects
Electromagnetic Effects	165	Magnetic field interactions, induction heating, current distribution
Fluid Flow Phenomena	145	Keyhole formation, gas dynamics, metal vapor effects
Thermodynamic Principles	125	Phase equilibria, chemical potential, driving forces for transformation
Total Physics Content	**1,840**	**Fundamental scientific principles underlying welding processes**