

Appendix A. DPO Experiment Configuration

This section provides a detailed breakdown of the configuration used for the Direct Preference Optimization (DPO) training experiments. The core training logic is adapted from the example `dpo.py` script provided by the Hugging Face TRL library.¹⁹

A.1. Model, Data, and Preprocessing

The core assets and data handling steps for the experiment are detailed below.

- **Base Model:** The starting point for DPO training was Llava-TCxYZL-Instruct.
- **Preference Dataset Construction:** The preference dataset is not publicly available due to licensing restrictions on the source data. It was constructed using the ground truth instruction-response pairs for the training and validation splits as the `prompt` and `chosen` fields, respectively. For each prompt, a `rejected` response was generated using the base model.
- **Data Preprocessing:** Each example in the dataset was processed to create `prompt`, `chosen`, and `rejected` fields. The prompt was formatted using the model’s chat template, while the chosen and rejected responses were stripped of any leading assistant-role headers.
- **Sequence Lengths:** The maximum sequence length was set to **8192** tokens, with the prompt limited to **8171** tokens.²⁰

A.2. DPO Training Hyperparameters

Table 5 lists the key DPO parameters. The effective global batch size was 256 (1 per-device batch \times 4 GPUs \times 64 accumulation steps).

Table 5: DPO Training Hyperparameters.

Parameter	Value
DPO Beta (β)	0.01
Loss Type	sigmoid
Learning Rate	5×10^{-7}
LR Scheduler	Cosine
Optimizer	8-bit Paged AdamW (<code>adamw_bnb_8bit</code>)
Training Epochs	3
Warmup Ratio	0.1
Gradient Accumulation	64 steps
Gradient Clipping	1.0
Seed	42

19. We use TRL v0.9.6; <https://github.com/huggingface/trl/blob/v0.9.6/examples/scripts/dpo.py>

20. `max_prompt_length` had to be set to use default collator.

A.3. Execution Environment and Optimizations

The training was executed in a distributed environment using a DeepSpeed configuration adapted from a TRL template²¹.

- **Distributed Training:** DeepSpeed ZeRO Stage 3 with optimizer state offloaded to CPU.
- **Hardware:** Single machine with 4 GPUs.
- **Mixed Precision:** FP8 via Accelerate (`mixed_precision: fp8`).
- **Model Data Type:** `bfloat16`.
- **Attention:** Flash Attention 2.
- **Compiler:** `torch.compile`.
- **Memory:** Gradient checkpointing.

Appendix B. ORPO Experiment Configuration

This section details the configuration for the Odds Ratio Preference Optimization (ORPO) experiment. The training script is based on the TRL library’s example²².

B.1. Model, Data, and Execution Environment

The base model, dataset construction, data preprocessing, and execution environment (including all DeepSpeed and hardware settings) were identical to those used in the DPO experiment described in Appendix A.

B.2. ORPO Training Hyperparameters

Table 6 lists the key hyperparameters for the ORPO run. The effective global batch size was also 256.

Appendix A: Model Performance Comparison Statistics

This section provides detailed statistical analysis supporting the statement: *“Statistical validation using Mann-Whitney U tests revealed significant performance differences between most model pairs (13/15 comparisons, $p < 0.05$) with large effect sizes (9/15 with $d > 0.8$).”* All analyses were conducted on final grades from direct answer evaluations (n=40 per model).

B.3. Pairwise Model Comparisons

We performed Mann-Whitney U tests for all possible pairwise comparisons between the six evaluated models. This non-parametric test was chosen because it makes no assumptions about the underlying distribution of grades and is robust to outliers.

21. https://github.com/huggingface/trl/blob/v0.9.6/examples/accelerate_configs/deepspeed_zero3.yaml

22. <https://github.com/huggingface/trl/blob/v0.9.6/examples/scripts/orpo.py>

Table 6: ORPO Training Hyperparameters.

Parameter	Value
ORPO Beta (λ)	0.1
Loss Type	orpo
Learning Rate	5×10^{-7}
LR Scheduler	Cosine
Optimizer	8-bit Paged AdamW (adamw_bnb_8bit)
Training Epochs	3
Warmup Ratio	0.1
Gradient Accumulation	64 steps
Gradient Clipping	1.0
Seed	42

Table 7: Model Performance Summary Statistics (n=40 each).

Model	Mean	Std Dev	Min	Max
GPT-4-Turbo	80.9	5.6	68.8	89.5
Llama3-TAIDE	77.6	7.1	60.5	89.8
Mistral-Nemo	74.3	8.4	53.8	88.0
Meta-Llama-3-8B	68.3	11.3	41.0	87.8
Blawstral	62.0	21.0	7.4	92.5
Bllawa	51.4	22.8	6.0	85.25

B.4. Mann-Whitney U Test Results

Table 8 presents the complete results of all 15 pairwise comparisons. Effect sizes were calculated using Cohen’s d approximation for non-parametric tests.

Table 8: Pairwise Mann-Whitney U Test Results.

Comparison	Mean Diff	p-value	Significant	Effect Size (d)
GPT-4-Turbo vs Bllawa	29.4	5.06×10^{-10}	Yes	1.77
Llama3-TAIDE vs Bllawa	26.2	4.99×10^{-8}	Yes	1.55
Blawstral vs GPT-4-Turbo	-18.8	1.55×10^{-5}	Yes	1.23
Meta-Llama-3-8B vs GPT-4-Turbo	-12.6	1.23×10^{-7}	Yes	1.41
Bllawa vs Mistral-Nemo	-22.8	2.47×10^{-6}	Yes	1.33
Llama3-TAIDE vs Blawstral	15.6	7.97×10^{-4}	Yes	0.99
Llama3-TAIDE vs Meta-Llama-3-8B	9.4	7.95×10^{-5}	Yes	0.99
Bllawa vs Meta-Llama-3-8B	-16.8	6.35×10^{-4}	Yes	0.94
GPT-4-Turbo vs Mistral-Nemo	6.6	3.83×10^{-4}	Yes	0.92
Blawstral vs Mistral-Nemo	-12.2	0.009	Yes	0.77
Meta-Llama-3-8B vs Mistral-Nemo	-6.0	0.016	Yes	0.60
Llama3-TAIDE vs GPT-4-Turbo	-3.2	0.041	Yes	0.50
Bllawa vs Blawstral	-10.6	0.034	Yes	0.48
Llama3-TAIDE vs Mistral-Nemo	3.4	0.070	No	0.43
Blawstral vs Meta-Llama-3-8B	-6.2	0.299	No	0.37

B.5. Effect Size Analysis

Of the 15 pairwise comparisons:

- **13 comparisons (86.7%)** reached statistical significance ($p < 0.05$)
- **9 comparisons (60.0%)** demonstrated large effect sizes (Cohen’s $d > 0.8$)
- **4 comparisons (26.7%)** showed medium effect sizes ($0.5 \leq d \leq 0.8$)
- **2 comparisons (13.3%)** had small effect sizes ($d < 0.5$)

The largest effect sizes were observed for comparisons involving the weakest performing model (Bllawa) versus stronger models, indicating substantial practical differences in legal reasoning capability.

Appendix C. Question Length Effects Analysis

This section provides detailed analysis supporting the statement: *“Question length showed minimal systematic effects within models (ANOVA $p > 0.05$ for all), but robustness varied substantially (GPT-4-Turbo: 3.2% vs. Bllawa: 18.9% performance drop).”*

C.1. Question Length Distribution

Questions were divided into quartiles based on character length:

- **Q1 (Short):** 34-86 characters (n=60 total, 10 per model)
- **Q2:** 87-119 characters (n=60 total, 10 per model)
- **Q3:** 120-164 characters (n=60 total, 10 per model)
- **Q4 (Long):** 165-438 characters (n=60 total, 10 per model)

C.2. Within-Model ANOVA Results

One-way ANOVA was performed within each model to test for systematic effects of question length quartiles on performance.

Table 9: Within-Model ANOVA Results for Question Length Effects.

Model	F-statistic	p-value	Significant
Bllawa	0.67	0.578	No
Llama3-TAIDE	0.87	0.465	No
Blawstral	0.85	0.478	No
Meta-Llama-3-8B	1.53	0.222	No
GPT-4-Turbo	0.73	0.543	No
Mistral-Nemo	1.30	0.290	No

Result: All models showed non-significant ANOVA results ($p > 0.05$), indicating no systematic within-model effects of question length on performance.

C.3. Model Robustness Analysis

Despite the lack of systematic effects, models showed varying degrees of robustness to question length, measured as performance drop from shortest (Q1) to longest (Q4) quartile.

Table 10: Question Length Robustness by Model.

Model	Q1 Mean	Q4 Mean	Drop (Points)	Drop (%)
Meta-Llama-3-8B	65.4	68.6	-3.2	-4.9%
Llama3-TAIDE	78.5	77.0	1.5	1.9%
GPT-4-Turbo	81.5	78.9	2.6	3.2%
Mistral-Nemo	75.4	70.4	5.0	6.6%
Blawstral	71.1	57.9	13.2	18.5%
Bllawa	59.8	48.5	11.3	18.9%

C.4. Robustness Classification

Models were classified based on performance drop magnitude:

- **Robust (5% drop):** Meta-Llama-3-8B, Llama3-TAIDE, GPT-4-Turbo, Mistral-Nemo
- **Vulnerable (15% drop):** Blawstral, Bllawa

The 3.2% vs. 18.9% comparison between GPT-4-Turbo and Bllawa represents a 5.9-fold difference in vulnerability to question length variation, demonstrating substantial architectural differences in robustness.

Appendix D. Correlation Analysis

D.1. Pearson Correlation Results

Pearson correlations between question length and grade components were calculated for each model. Fisher transformation was applied to compute 95% confidence intervals.

Table 11: All Significant Correlations ($p < 0.05$) - 6 out of 36 total.

Model	Component	Correlation	95% CI
Mistral-Nemo	PresentationAndStyle	-0.434	[-0.657, -0.142]
Bllawa	Fragestellung	-0.381	[-0.619, -0.079]
Mistral-Nemo	Obersatz	-0.373	[-0.614, -0.070]
GPT-4-Turbo	PresentationAndStyle	-0.355	[-0.600, -0.049]
GPT-4-Turbo	Fragestellung	-0.328	[-0.580, -0.018]
Bllawa	Final Grade	-0.317	[-0.572, -0.006]

Overall Pattern: Correlations were generally weak ($|r| < 0.3$) with only 6 out of 36 model-component pairs reaching statistical significance, supporting the conclusion of minimal systematic question length effects.

Appendix E. Four Components of Juristisches Gutachten

Juristisches Gutachten is a fundamental tool in German law that provides a structured, written analysis of legal questions based on a given set of facts. It is used to evaluate whether specific legal requirements are met and to determine potential legal consequences.

1. Fragestellung (Problem Statement)

This defines the legal question under investigation as a hypothesis. It does not assert a conclusion but identifies the specific legal issue to be resolved. This step establishes the framework for the subsequent analysis, ensuring that all following stages directly address the core problem.

2. Obersatz (General Principle; Major Premise)

This specifies the relevant legal principle, rule, or statute applicable to the case. This involves citing the relevant law and delineating its constituent elements. The Obersatz provides the abstract criteria that must be satisfied for a legal consequence to arise and serves as the theoretical foundation for the analysis.

3. Subsumtion (Application of Law to Facts)

This constitutes the primary analytical step. It systematically applies the principles outlined in the Obersatz to the specific facts of the case. For each element of the legal rule, the analysis demonstrates whether the facts fulfill the required conditions. This stage establishes the logical link between the abstract legal norm and the concrete case circumstances, providing the evidentiary basis for the final conclusion.

4. Ergebnis (Conclusion)

This presents the conclusion that logically follows from the preceding Subsumtion. It offers a concise response to the initial Fragestellung, confirming or rejecting the hypothesis based on the legal reasoning. The conclusion is the transparent and necessary outcome of the structured application of law to facts.

Appendix F. Prompt for GPT-4o Grading on Task D

F.1. Taiwanese Mandarin (original prompt)

假設你是律師、司法官的閱卷教授，以下是針對答題的鑑定題裁給你參考，並協助以下工作。

A. 將答題內容分成四個主要部份：

1. 案例事實描述（25%）：例如「狗咬傷人」。
2. 選定適用的法律規範（請求權基礎）（15%）：例如「甲得否向乙依民法第 348 條規定請求支付 A 屋並移轉其所有權」。這是法律推理的大前提。
3. 分解請求權基礎的構成要件並進行涵攝之適用（35%）：這是法律推理的小前提。
4. 結論（15%）：即對於大前提做成肯定或否定的結論。

B. 注意事項：法律文案的清楚使用、客觀風格，以及法條、判例與學說的正確引用也佔分10%。

C. 任務：

1. 給定以下申論題題目與答案，請先標示出答案中哪些敘述分別屬於 1, 2, 3, 4（必須依照 A 點所述的四個部分，列出屬於該部分的答案原始敘述，不可使用題目敘述）。
2. 根據各部分的評分比例，針對每個部分分別給予分數。
3. 分解請求權基礎的構成要件並進行涵攝之適用（35%）：這是法律推理的小前提。
4. 最後給予總分。

D. 題目僅作為批改參考，你不需要評價或使用題目敘述。

題目：{problem}

F.2. English translation

Assume you are a professor evaluating exam answers for lawyers and judges. The following is provided as a reference for grading and to assist with the tasks below.

1. Divide the answer into four main parts:
 - (a) Case facts description (25%): e.g., "The dog bit a person."
 - (b) Selection of applicable legal provisions (basis of the claim) (15%): e.g., "Can A request B to pay for house A and transfer its ownership according to Article 348 of the Civil Code?" This is the major premise of legal reasoning.
 - (c) Breakdown of the elements of the claim and application to the facts (35%): This is the minor premise of legal reasoning.
 - (d) Conclusion (15%): That is, a positive or negative conclusion regarding the major premise.
2. Notes: Clear use of legal language, objective style, and correct citation of statutes, precedents, and doctrines account for 10% of the score.
3. Tasks:
 - (a) Given the following essay question and answer, first identify which statements in the answer belong to parts (a), (b), (c), and (d) in 1 (you must follow the four parts above and list the original statements from the answer; do not use the question text).
 - (b) Assign scores for each part according to the specified weighting.
 - (c) Finally, provide the total score.
4. The question is for reference only; you do not need to evaluate or use the question text.

Question: {problem}

Appendix G. Instance of Legal Symposium

Issuance No.: Judicial Yuan (69) Criminal Division No. 059

Issuance Date: December 11, 1980

Conference Agency: Taiwan High Court Taichung Branch and Subordinate Courts

Legal Issue

After the commencement of trial in a juvenile case, may the victim of a crime in such a case file a supplementary civil lawsuit to seek damages?

少年案件之犯罪被害人，在少年法庭開始審理後，可否提起附帶民事訴訟，請求損害賠償？

Discussion Opinions

Opinion A (甲說)

According to Article 1 of the Juvenile Delinquency Act, which provides that “The handling of juvenile corrective measures shall be governed by this Act; where not provided herein, other laws shall apply. Since there is no prohibition against filing a supplementary civil lawsuit, it should be applicable. This is further supported by Article 65, Paragraph 2 of the same Act, which explicitly prohibits the application of private prosecutions in juvenile criminal cases, showing that if prohibition were intended, it would be expressly stated.

依少年事件處理法第一條規定「少年管訓處分之處理，依本法之規定，本法未規定者，適用其他法律」，故既無提起附帶民事之禁止規定，自應適用，此觀之同法第六十五條第二項關於少年刑事案件不適用自訴程序之明文禁止規定益為明確。

Opinion B (乙說)

Juvenile corrective proceedings are completely different in nature from ordinary criminal cases. To implement the spirit of juvenile corrective measures, which emphasize actively guiding juveniles to reform and improve, it is inappropriate to create further complications by allowing supplementary civil lawsuits in such proceedings. Moreover, according to Article 487, Paragraph 1 of the Code of Criminal Procedure, a supplementary civil lawsuit must be filed within the course of criminal proceedings. Therefore, since juvenile corrective proceedings are not criminal proceedings, this provision does not apply.

少年管訓事件與一般刑事案件之性質完全互異，為貫徹少年管訓處分以積極輔導少年改過向善之精神，自不宜在管訓處分審理中再生枝節。況依刑事訴訟法第四百八十七條第一項規定，提起附帶民事訴訟者，須在刑事訴訟程序中，始得提起，從而開始審理自不適用此規定。

Adopted Opinion (研討結果)

Adopting Opinion B: The Juvenile Delinquency Act explicitly enumerates the provisions of the Code of Criminal Procedure that may be applied mutatis mutandis (e.g., Articles 16 and 24 of the Act). Since the provision on supplementary civil lawsuits under the Code of Criminal Procedure is not among those incorporated, it cannot be applied.

採乙說，少年事件處理法關於準用刑事訴訟法者均有列舉之規定（如少年事件處理法第十六條、第二十四條），刑事訴訟法之附帶民事訴訟，於少年事件處理法中既無準用之規定，自不得準用。

Judicial Yuan Second Division Research Opinion (司法院第二廳研究意見)

Agrees with the above conclusion and adopts Opinion B. However, the wording of the issue should be clarified: the phrase “victim of a crime in a juvenile case” is inaccurate and should be amended to “victim in a juvenile corrective case.” As for victims in juvenile criminal cases, they may still file a supplementary civil lawsuit.

同意研討結果，以乙說為當。惟問題意旨係指少年管訓事件，其所謂「少年案件之犯罪被害人」一語，用詞不當，應改為「少年管訓事件之被害人」，至於少年刑事案件之犯罪被害人，仍可提起附帶民事訴訟。