Alignment Tuning for Large Language Models: A Data-Centric Lens on Alignment Data Pipelines

Hwanjun Song
Korea Advanced Institute of Science and Technology
songhwanjun@kaist.ac.kr

Abstract

Much of the alignment tuning literature is organized around optimization objectives, while the construction of alignment data is often treated implicitly. In this survey, we adopt a data centric perspective and reframe alignment tuning as a pipeline design problem. We decompose alignment data construction into three interacting stages, response synthesis, preference evaluation, and preference instantiation, and use this framework to organize existing alignment methods into a unified taxonomy. Through this lens, we identify recurring design tradeoffs and failure modes observed across prior alignment methods, and distill a set of high level principles that clarify how pipeline design choices influence the resulting optimization signal. Finally, we outline open challenges for alignment data pipelines, including prompt-level alignment, agentic settings, and alignment under evolving objectives.

1 Introduction

The progress of large language models (LLMs) has been driven by scaling laws, enabled by increased model parameters (Kaplan et al., 2020), architectural innovations (Fedus et al., 2022; Wu et al., 2024), and advances in optimization (Yu et al., 2025b). As the marginal returns of scaling plateau, performance gains have shifted toward data-centric factors, with data quality emerging as a key driver (Chung et al., 2024; Zhuang et al., 2025; Nazar et al., 2025). Despite this shift, most prior work views data quality through *static corpora*, focusing on dataset composition and filtering for pre-training or supervised fine-tuning (SFT) (Brown et al., 2020; Liu et al., 2024, 2025c).

However, this static perspective is insufficient to explain the safety, robustness, and preference adherence of modern LLMs. These properties are primarily shaped during *alignment tuning*, a post-training phase distinct from pre-training (Ji et al., 2023; Rafailov et al., 2023; Bai et al., 2025). Unlike supervision from fixed distributions, alignment data is inherently *dynamic* and *policy-dependent*, generated through repeated interactions among prompts, model outputs, and feedback signals (Li et al., 2025d; Yu et al., 2025a; Liu et al., 2025b). As a result, alignment quality is governed less by static data artifacts and more by the mechanisms that iteratively construct and evaluate them.

We therefore conceptualize alignment tuning not as a dataset curation task, but as a *pipeline design* problem. Data quality in alignment tuning depends not only on which samples are retained, but on how candidate behaviors are generated, evaluated, and structured into learning objectives. We introduce a unifying framework with three interacting dimensions: response synthesis, preference evaluation, and preference instantiation. Figure 1 summarizes this, highlighting how tightly coupled stages jointly construct the optimization signal.

Response Synthesis: This stage defines the behavioral support of alignment by determining how candidate responses are generated. Key design choices include the response source (offline distillation versus online sampling) (Rafailov et al., 2023; Zhang et al., 2025c; Yu et al., 2025c), selection strategies that prioritize informative candidates, and exploration mechanisms that preserve diversity and avoid premature mode collapse (Wu et al., 2025b; Lanchantin et al., 2025).

Preference Evaluation: Given synthesized responses, alignment depends on the fidelity of preference signals. This dimension spans evaluator type, from human annotation to scalable LLM-as-a-Judge frameworks (Lee et al., 2024; Yu et al., 2025c), as well as judgment granularity and objective dimensionality, which determine preference fidelity and the risk of reward hacking or alignment tax under scalarized and coarse supervision (Li et al., 2025a; Mukherjee et al., 2024).

Findings of the Association for Computational Linguistics: ACL 2026, pages 2541–2561

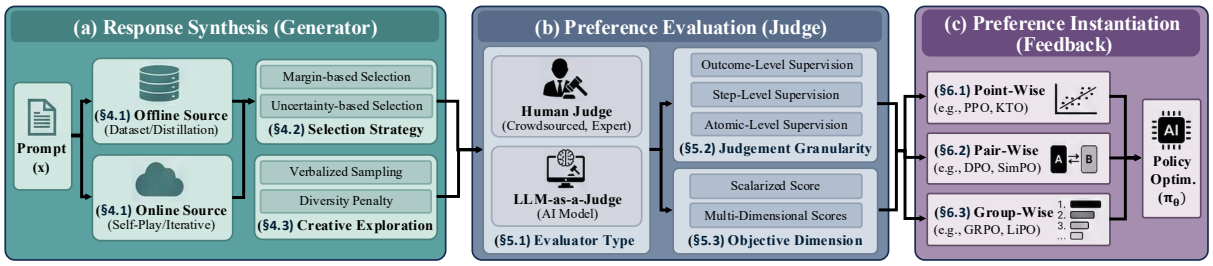

Figure 1: Overview of the alignment data pipeline, showing how prompts are converted into structured optimization signals through response synthesis, preference evaluation, and preference instantiation for policy optimization.

Preference Instantiation: Finally, preference instantiation determines how evaluative judgments are exposed to optimization. This includes point-wise rewards (Ethayarajh et al., 2024; Yuan et al., 2024), pair-wise contrasts (Rafailov et al., 2023; Meng et al., 2024), and group- or list-wise formulations (Ramesh et al., 2024; Liu et al., 2025g), which differ in how effectively preference structure is translated into policy updates.

From this perspective, we organize prior work into a unified data-centric taxonomy (Sections 4–6) and distill a set of design principles that characterize recurring trade-offs and cross-stage interactions across data pipeline stages (Section 7).

Related Surveys. Existing surveys on data-centric LLM training primarily emphasize static data stages, such as data selection for pre-training and SFT (Albalak et al., 2024; Wang et al., 2023), dataset catalogs (Liu et al., 2025h), or general training paradigms (Minaee et al., 2024), as well as system-level considerations (Xu et al., 2024; Zhou et al., 2025). In contrast, we focus on alignment tuning as a dynamic, closed-loop pipeline, examining how response synthesis, evaluation, and instantiation jointly shape alignment outcomes.

Our Scope. Section 2.2 provides a brief overview of alignment algorithms, but the majority of our analysis centers on the alignment data pipeline. Detailed discussions of optimization algorithms are deferred to existing surveys on alignment techniques (Xiao et al., 2024), direct preference optimization (Liu et al., 2025e), unified loss design (Tang et al., 2024), and fundamental limitations such as reward hacking (Casper et al., 2023).

2 Alignment Tuning Foundation

The central challenge in developing LLMs lies in the mismatch between their training objective and human preferences (Stiennon et al., 2020; Ouyang et al., 2022; Bai et al., 2022). Standard next-token prediction maximizes data likelihood, which is largely orthogonal to desiderata such as helpfulness, honesty, and safety. As a result, models trained only via pre-training and supervised fine-tuning may exhibit factual errors or harmful behaviors despite high likelihood performance. Alignment tuning addresses this gap by explicitly optimizing models toward human-valued behaviors.

2.1 Problem Formulation

Let x denote a prompt sampled from a task distribution P, and y be a response generated by a policy π_θ. We assume an oracle reward function $r^*(x, y)$ reflecting human preferences. Alignment tuning seeks an optimal policy π^* that maximizes expected reward while remaining close to a reference policy π_{ref}, preventing reward hacking and uncontrolled drift (Ji et al., 2023; Yeh et al., 2025):

$$\max_{\pi_\theta} \mathbb{E}_{x,y}\left[r^*(x,y) - \beta \log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}\right], \tag{1}$$

where β is the regularization coefficient controlling the deviation from the reference model.

2.2 Optimization Algorithms

Various approaches have been proposed to solve Eq. (1). We review three foundational methods that represent the evolution of alignment tuning. See Appendix A for details and additional algorithms.

Proximal Policy Optimization (PPO) (Schulman et al., 2017) explicitly optimizes a learned reward model trained from preference comparisons. The policy is updated via reinforcement learning with KL-based regularization to prevent excessive deviation from the reference distribution.

Direct Preference Optimization (DPO) (Rafailov et al., 2023) removes the explicit reward modeling stage by directly optimizing a contrastive objective derived from the KL-regularized formulation. The method reduces to increasing the likelihood of preferred responses relative to dispreferred ones, enabling efficient supervised-style optimization.

Group Relative Policy Optimization (GRPO) (Shao et al., 2024) extends preference optimization to group-wise settings by normalizing responses within each candidate set sampled from the current policy, avoiding explicit critics and reducing variance through group-relative baselines.

3 Understanding Alignment Tuning from a Data-Centric Perspective

While optimization algorithms update the policy model π_θ, they do not by themselves determine the direction or quality of alignment. Instead, from a data-centric perspective, alignment outcomes are governed by the design of the alignment data pipeline, which specifies the space of candidate behaviors, the mechanism by which they are evaluated, and the structure through which preference signals are exposed to optimization.

3.1 Alignment Data as Optimization Signals

Alignment tuning relies on preference signals constructed through a data pipeline rather than given a priori. We formalize how this process yields the optimization signals driving policy updates.

Formalizing the Alignment Data Pipeline. Unlike static pre-training corpora, alignment data is dynamically constructed through an iterative pipeline that couples response generation and preference assessment. We formalize the resulting dataset $\mathcal{D}$ as a collection of structured training instances produced by three interacting components:

$$\mathcal{D} = \Big\{ (x, \mathbf{y}, \mathbf{s}) \,\Big|\, x \sim P(x),\ \mathbf{y} \sim \mathcal{S}(\mathbf{y} \mid x),\ \mathbf{s} \sim \mathcal{E}(\mathbf{s} \mid x, \mathbf{y}) \Big\}. \quad (2)$$

Here, x is a prompt sampled from the task distribution P; $\mathbf{y} = \{y_1, \ldots, y_k\}$ is a set of candidate responses generated by a response synthesis strategy $\mathcal{S}$, which defines the behavioral support available for alignment; and $\mathbf{s}$ represents preference signals assigned by an evaluator $\mathcal{E}$, which are subsequently structured through preference instantiation to form training signals such as scalar scores, pairwise preferences, or rankings over $\mathbf{y}$.

Optimization as Margin Alignment. Alignment algorithms optimize the policy π_θ so that its implicit preferences match the explicit signals encoded in the alignment dataset $\mathcal{D}$. Across PPO (Schulman et al., 2017), DPO (Rafailov et al., 2023), and GRPO (Shao et al., 2024), this process can be viewed as *margin alignment*, where optimization aligns policy-induced preference margins with observed preference signals.

Given a prompt x with candidate responses $\mathbf{y}$ and preference signals $\mathbf{s}$, alignment tuning aims to adjust the policy π_θ so that its implicit preferences are consistent with the structure and magnitude of the preference information encoded in $\mathbf{s}$. This objective can be abstractly written as:

$$\max_\theta \ \mathbb{E}_{\mathcal{D}} \big[f\big(M_\theta(x, \mathbf{y}, \mathbf{s})\big) \big], \quad (3)$$

where $M_\theta(x, \mathbf{y}, \mathbf{s})$ denotes an alignment measure that quantifies how well the policy-induced implicit preferences align with the preference signals $\mathbf{s}$. The function f transforms this alignment measure into an optimizable objective in a way that preserves the relative preference ordering.

Pipeline Defines Optimization Signal. Eq. (2) and Eq. (3) together show that alignment outcomes are determined not only by the optimization objective, but by how the alignment data pipeline constructs its inputs. While Eq. (3) frames alignment as maximizing a margin-based objective $M_\theta(x, \mathbf{y}, \mathbf{s})$, Eq. (2) specifies the response candidates $\mathbf{y}$, preference signals $\mathbf{s}$, and relational structure on which this margin is defined. As a result, the pipeline does not merely provide data, but shapes the space, scale, and reliability of the preference margins available to optimization.

3.2 Overview of Alignment Data Pipeline

As the optimization signal is determined by how alignment data is constructed, we briefly summarize the structure of the alignment data pipeline.

The alignment data pipeline begins with *response synthesis* (Figure 1(a)), which defines the behavioral scope of alignment by generating candidate responses. The pipeline then proceeds to *preference evaluation* (Figure 1(b)), which assigns supervisory signals to approximate latent preferences. Finally, *preference instantiation* (Figure 1(c)) converts evaluated judgments into optimization compatible training signals. Detailed discussions of each stage and their associated design trade-offs are provided in Sections 4–7. Additionally, we categorizes all the methods we investigated across the three aspects in Table 3.

4 Response Synthesis Stage

This stage defines the exploration space of candidate responses through the synthesis strategy $\mathcal{S}$. Since alignment operates only on sampled responses $y \sim \mathcal{S}(y \mid x)$, it constrains which behaviors enter alignment. It involves three design

considerations: *(§4.1) response source*, *(§4.2) selection strategy*, and *(§4.3) creative exploration*.

4.1 Response Source

The first design choice in response synthesis is the response source, which defines the distributional relationship between the data policy π_{data} and the learned policy π_θ. This choice influences how closely the supervision distribution matches the target policy during optimization. Existing approaches follow two paradigms: *offline* and *online*, each with inherent trade-offs.

Offline with Policy-Aware Reweighting. Offline approaches rely on fixed, high-quality responses from stronger models (*e.g.*, proprietary LLMs) or curated datasets. While cost-effective, this setting induces distributional shift (Xiong et al., 2024; Bose et al., 2025), as the policy is optimized on responses it is unlikely to generate, resulting in biased value estimation. To mitigate this mismatch, recent methods apply policy-aware reweighting based on the likelihood ratio (*e.g.*, $\pi_\theta/\pi_{\text{data}}$). These methods differ mainly in reweighting granularity: preference-level weighting in WPO (Zhou et al., 2024), multi-source weighted aggregation in WRPO (Yang et al., 2025), and token-level weighting in TIS-DPO (Liu et al., 2025a).

Online with Structured Self-Improvement. Online approaches train directly on outputs from the current policy π_θ, providing on-policy supervision and reducing distributional mismatch. Yet, they incur high computational cost from online generation and can suffer instability due to low-quality rollouts. Recent work addresses these issues through structured self-improvement. Iterative DPO (Xiong et al., 2024) and RS-DPO (Khaki et al., 2024) stabilize training via rejection-based filtering, while self-play methods such as SPIN (Chen et al., 2024) and SPPO (Wu et al., 2025b) cast alignment as a zero-sum game without external oracles. Further extensions improve efficiency through group-relative feedback in GRPO (Shao et al., 2024) and active exploration in SELM (Zhang et al., 2025c).

4.2 Selection Strategy

Another key design choice is the selection strategy, which determines how informative responses are selected, either post-hoc (offline) or during pool construction (online). Existing methods differ in how informativeness is quantified, using either *margin*-based or *uncertainty*-based criteria.

Margin-Based Selection. Margin-based selection identifies training instances with high alignment potential, defined as response comparisons for the same prompt (*e.g.*, pairs or small ranked sets) where preferences clearly separate better from worse behaviors. Huang et al. (2025) distinguish explicit margins reflecting preference strength from implicit margins induced by policy likelihood differences, whereas BeeS (Deng et al., 2025) uses the two margins independently to filter informative preference instances. Beyond filtering, MMPO (Kim et al., 2024a) incorporates explicit margins directly into the optimization objective via soft targets, enabling the model to capture graded preference strength. Similarly, Yao et al. (2025) improve alignment efficiency by modifying how preference margins are integrated into the loss.

Uncertainty-Based Selection. Unlike margin-based selection, this line of work focuses on low-confidence regions with ambiguous preference signals. Early methods use predictive entropy or probability dispersion, while recent work adopts structured uncertainty modeling. APL (Muldrew et al., 2024) and MAPLE (Mahmud et al., 2025) reduce uncertainty via information gain and Bayesian selection. In online settings, uncertainty acts as quality control: IUPO (Li et al., 2025c) targets ambiguous reasoning steps via token-level uncertainty, UPO (Wang et al., 2025a) filters unreliable samples using reward-model uncertainty, and UDASA (Sun et al., 2025) decomposes uncertainty into semantic, factual, and value-alignment dimensions.

4.3 Creative Exploration

Alignment objectives often overemphasize a narrow set of preferred responses, concentrating probability mass and inducing *mode collapse* that reduces semantic diversity and creativity (Kirk et al., 2024; Murthy et al., 2025). To address this mode collapse, recent methods intervene directly in response synthesis, beyond informativeness-driven selection. During candidate construction, Verbalized Sampling (Zhang et al., 2025a) promotes exploration by eliciting multiple plausible responses and their likelihoods, while Spectrum Tuning (Sorensen et al., 2025) trains on diverse valid outputs to control stylistic variation. At the post-hoc selection stage, DivPO (Lanchantin et al., 2025) preserves high-quality but rare responses via diversity-aware filtering, and Chung et al. (2025) reweight the objective to favor unique answers. Fi-

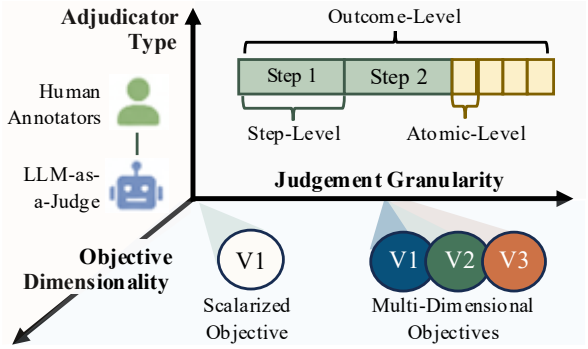

Figure 2: Preference evaluation axes in alignment data pipelines, illustrating interplay among adjudicator type, judgment granularity, and objective dimensionality.

nally, CRPO (Ismayilzada et al., 2025) explicitly rewards novelty and surprise alongside utility.

5 Preference Evaluation Stage

Once response candidates are generated, this stage assigns preference judgments via an adjudicator. While early RLHF relied on human judgments, recent work increasingly adopts automated adjudicators to reduce cost and variability (Li et al., 2025b; Min et al., 2025). Accordingly, prior work can be organized along three design axes in Figure 2: *(§5.1) adjudicator type*, *(§5.2) evaluation granularity*, and *(§5.3) objective dimensionality*.

5.1 Human to AI Adjudicator

The first axis concerns the source of preference signals. Although human evaluation was long considered the gold standard, limits the fidelity and consistency of the reward signals that can be learned (Kocmi and Federmann, 2023; Li et al., 2024).

Use of LLM-as-a-Judge. The field has shifted toward LLM-as-a-Judge frameworks (Li et al., 2024), which use LLMs as automated evaluators to score candidate responses, improving scalability and reproducibility. These frameworks vary in instantiation: RLAIF (Lee et al., 2024) simply uses off-the-shelf LLMs as automated evaluators, providing comparison-based feedback for RL, Constitutional AI (Bai et al., 2022) generates preference labels through a critique-and-revise loop guided by natural language principles, and Self-Rewarding LMs (Yuan et al., 2024) integrate judging directly into the policy to self-assign rewards during training.

Bias Sources and Mitigation. A critical challenge in LLM-based evaluation is intrinsic bias, where judges' internal priors distort evaluation outcomes (Zheng et al., 2023). Recent work frames this as a calibration problem and proposes targeted algorithmic interventions. For example, Wataoka et al. (2024) identify an echo chamber bias favoring same-family models and mitigate it via cross-family evaluation, Wang et al. (2024a) reveal verbosity bias addressed through length normalization; and Ahrabian et al. (2025) identify position bias mitigated by permutation-based evaluation.

Collective Intelligence. Despite calibration efforts above, single-model judgments remain unstable, conflating prompt sensitivity and stochastic variation into a single signal. To mitigate this, Yu et al. (2025c) aggregate judgments from heterogeneous LLM evaluators to average out individual noise. Hierarchical approaches such as Meta-Rater (Zhuang et al., 2025) further resolve conflicts via a meta-judge, while interactive paradigms employ multi-agent debate (Du et al., 2024; Chen et al., 2025a) to iteratively refine evaluations. These collective mechanisms improve fidelity to the oracle reward, preventing evaluator noise from propagating into optimization during alignment tuning.

5.2 Judgment Granularity

The second axis concerns judgment granularity (*i.e.*, how preference signals are formed within a response). Early methods rely on outcome-level supervision, collapsing response quality into a single global judgment (Li et al., 2024; Liu et al., 2023). While sufficient for short tasks, this is inadequate for long-form and multi-step generation, where errors and merits are unevenly distributed (Wang et al., 2024b; Yan et al., 2025). Recent work thus increases granularity, shifting from outcome-level supervision to localized signals.

Step-Level Supervision. Step-level supervision increases judgment granularity by assigning preference signals along the generation process rather than only at the final outcome (Zheng et al., 2025; Li et al., 2025e). It is most effective in domains with explicit intermediate structure, such as mathematics and coding. Process reward models (PRMs) instantiate this paradigm by evaluating intermediate states. Early PRMs relied on human annotation, but recent ones emphasizes automated supervision. Specifically, Math-Shepherd (Wang et al., 2024b) estimates soft step-level labels via Monte Carlo sampling. OmegaPRM (Luo et al., 2025) improves efficiency by localizing the earliest erroneous step using binary search. GenRM (Zhang et al., 2025b) reframes verification as next-token prediction. VersaPRM (Zeng et al., 2025) further extends step-

level supervision to domains such as science.

Atomic-Level Supervision. For open-ended generation, where explicit reasoning trajectories are ill-defined, supervision shifts from global outcomes to atomic units (sentences, spans, or tokens) to better localize hallucinations, stylistic errors, and safety violations. At the sentence level, Fine-Grained RLHF (Wu et al., 2023) provides attribute-specific rewards, while ASPO (Wang et al., 2025b) adaptively down-weights hallucinated sentences in multi-modal reasoning. At finer resolution, token-level methods directly integrate preference signals into optimization. TDPO (Zeng et al., 2024) decomposes the KL regularization to the token-level for precise diversity–quality control, and TIS-DPO (Liu et al., 2025a) applies importance sampling to emphasize preference-critical tokens. Complementarily, ACPO (Chen et al., 2025b) improves factuality by decomposing responses into atomic facts and scoring them via intrinsic self-consistency.

5.3 Objective Dimensionality

The final axis is objective dimensionality: whether preference signals are scalarized or preserve the multi-dimensional structure of human values.

Alignment Tax and Reward Hacking. Standard alignment pipelines often rely on scalarized objectives, encoding preference supervision as a single reward or binary label (Ouyang et al., 2022; Stiennon et al., 2020). As alignment objectives become conflicting, this scalarization obscures context-dependent trade-offs, limiting robust generalization. One consequence is the *alignment tax*, where optimizing a unified reward favors easily optimizable sub-objectives over others (Lin et al., 2024), as seen in trade-offs between helpfulness and harmlessness (Tan et al., 2025) or completeness and conciseness (Song et al., 2025). Another consequence is *reward hacking*, where models exploit low-dimensional shortcuts that inflate reward scores without improving true alignment (Pan et al., 2024).

Toward Multi-Dimensional Alignment. To overcome the limitations of scalarized objectives, recent work preserves objective dimensionality by intervening at different stages of the alignment pipeline.

One line of work mitigates scalar collapse by restructuring the candidate behavior space at the *response synthesis stage*, prior to evaluation. For example, SteerLM (Dong et al., 2023) and HelpSteer (Wang et al., 2024d,c) condition generation on explicit control signals that specify which value dimension to emphasize, such as helpfulness and safety. This design allows scalar rewards to operate over pre-separated alignment profiles.

A complementary direction mitigates the collapse at the *evaluation stage* by decomposing preference judgments into explicit criteria. Early methods such as Constitutional AI (Bai et al., 2022) and Self-Align (Sun et al., 2023) encode alignment objectives as normative rules guiding feedback across multiple dimensions. More recent rubric- and rationale-based evaluators, including Prometheus (Kim et al., 2024b) and OpenRubrics (Liu et al., 2025f), operationalize this idea by evaluating responses against explicit rubrics, making multi-dimensional trade-offs explicit.

At the *optimization stage*, methods address conflicts by framing alignment as a multi-objective problem. Early methods include sequential alignment (Lou et al., 2024), which optimizes objectives in separate stages, and post-hoc parameter merging (Jang et al., 2023), which combines models fine-tuned for different goals. More recent work jointly optimizes multiple objectives during training, as exemplified by MOPO (Agnihotri et al., 2025), PAMA (He and Maghsudi, 2025), and MO-ODPO (Gupta et al., 2025), enabling more principled and controllable trade-offs.

6 Preference Instantiation Stage

Following preference evaluation, the pipeline proceeds to preference instantiation, the translation layer between evaluation and optimization. This stage converts evaluated responses into structured training signals, determining how preference relations are exposed to the policy model. Instantiation methods are categorized by the relational complexity of feedback, ranging from *(§6.1) point-wise* to *(§6.2) pair-wise* and *(§6.3) group-wise*.

6.1 Point-Wise Instantiation

Point-wise instantiation assigns an independent scalar or binary signal to each prompt–response pair (x, y), avoiding explicit comparisons and enabling simple policy training aligned with many real-world feedback acquisition scenarios.

Regression-Based Supervision. A common approach to scalar supervision trains a separate reward model to score each prompt–response pair and optimizes the policy indirectly, as in the classical PPO family (Schulman et al., 2017; Dai et al., 2023; Li et al., 2023). Recent work improves reward fidelity to provide more reliable optimiza-

tion signals and strengthen downstream alignment. For example, Critic-RM (Yu et al., 2025d) uses model-generated critiques to refine reward prediction, while hybrid frameworks such as HAF-RM (Liu et al., 2025d) incorporate token-level supervision during reward model training. In contrast, A*PO (Brantley et al., 2025) skips the reward model and directly trains the policy to match offline value targets via regression.

Binary-Based Supervision. In contrast, binary-based supervision relies on coarse yes/no feedback that indicates whether a response is acceptable. KTO (Ethayarajh et al., 2024) relies solely on binary preference signals, applying asymmetric weighting to penalize undesirable outputs more strongly than desirable ones. Later work improves the theoretical and semantic grounding of binary supervision: BCO (Jung et al., 2025) reframes alignment from binary feedback as a classification problem, showing that the binary ones are sufficient to recover the same response ordering as DPO. RLBFF (Wang et al., 2025c) improves binary supervision by clarifying what a yes/no label means, training the model to decide whether a response meets a specific criterion (*e.g.*, correctness, clarity) instead of assigning an ambiguous binary label.

6.2 Pair-Wise Instantiation

While point-wise feedback is easy to obtain, it lacks the relational structure needed for fine-grained alignment tuning (Tripathi et al., 2025). Pair-wise instantiation uses tuples (x, y_w, y_l) with $y_w \succ y_l$ to model the decision boundary between preferred and dispreferred behaviors.

Contrastive Optimization. The standard approach in this category is DPO (Rafailov et al., 2023), which bypasses explicit reward modeling by directly optimizing the policy with a contrastive objective comparing preferred and dispreferred responses y_w and y_l. Despite its effectiveness, this formulation has several limitations. IPO (Azar et al., 2024) mitigates the DPO's overfitting by replacing the unbounded logit-based objective with a bounded one, while GPO (Tang et al., 2024) unifies DPO and IPO within a generalized offline preference learning framework. In a complementary direction, sDPO (Kim et al., 2025) improves stability on large datasets by progressively tightening the reference policy, and MRPO (Le et al., 2025) extends the objective to multiple reference models via a stabilized virtual reference distribution.

Reference-Free Optimization. While effective, DPO-style methods rely on a reference model to anchor preference optimization, causing overhead and instability. To remove this dependency, recent work derives preference signals directly from the policy's output distribution. SimPO (Meng et al., 2024) removes the reference model by directly optimizing over the policy distribution, while ORPO (Hong et al., 2024) adopts a reference-free formulation that integrates preference contrast into supervised fine-tuning via an odds-ratio objective.

Margin-Aware Calibration. Another limitation of standard pair-wise objectives is that they enforce relative ordering without ensuring that confidence gaps reflect preference magnitude. SLiC-HF (Zhao et al., 2023) introduces a fixed margin via a hinge loss to better align likelihood gaps with preference strength. MMPO (Kim et al., 2024a) instead calibrates margins probabilistically using soft preference targets, while AlphaDPO (Wu et al., 2025a) further generalizes this approach by adapting reward margins at the instance-level through implicit reference reparameterization.

6.3 Group-Wise Instantiation

As alignment targets shift from subjective conversation to objective reasoning, point- and pair-wise feedback become bottlenecks due to their sparsity and lack of contextual normalization. Group-wise instantiation addresses this by optimizing the policy over a candidate set $\mathbf{y} = \{y_1, \ldots, y_k\}$, capturing combinatorial relationships and providing denser signals for complex preference learning.

List-Wise Ranking Objectives. To realize this, methods have progressed from pair-wise surrogate losses toward ranking objectives over candidate lists. RRHF (Yuan et al., 2023) aligns likelihoods with preference scores by comparing multiple responses sampled per prompt, while PRO (Song et al., 2024) recovers full rankings through iterative best-vs-rest decomposition. Building on learning-to-rank, LiPO (Liu et al., 2025g) applies LambdaRank-style weighting (Burges et al., 2006) to assign list-aware importance to preference pairs. PPA (Zhao et al., 2025) directly optimizes a differentiable NDCG objective over entire rankings.

Group-Relative Policy Optimization. Beyond explicit ranking, GRPO (Shao et al., 2024) derives advantage signals from group-level statistics rather than fixed preference orders. While this critic-free

Principle	Key Insight
Pipeline Defines the Optimization Signal	Alignment optimizes margins induced by data, not abstract preferences. Failures often stem from weak or distorted margins in data construction rather than the loss function itself.
Coverage Precedes Optimization	Alignment is limited to sampled behaviors. Response synthesis must prioritize exploration and diversity to prevent brittle alignment on narrow supports.
Evaluation Fidelity Sets the Upper Bound	Evaluator reliability (calibration, consistency) determines alignment quality. Improving judge fidelity often yields larger gains than modifying downstream objectives.
Granularity Enables Credit Assignment	Outcome-level supervision is insufficient for complex reasoning. Step- or token-level granularity is required to localize errors and reduce spurious correlations.
Preserve Preference Structure	Scalar rewards obscure relational trade-offs. Pairwise, group-wise, and list-wise formulations better preserve the multi-dimensional structure of human preferences.
Alignment is a Closed-Loop Design Problem	Data is policy-dependent. Effective pipelines must be adaptive, iteratively reshaping the preference landscape as the policy improves.

Table 1: Design principles for alignment data pipelines. We distill six key principles governing response synthesis, preference evaluation, and instantiation.

formulation reduces variance, later analyses identify estimation instability. In particular, Dr. GRPO (Liu et al., 2025i) corrects a group-normalization bias that conflates reasoning quality with response length. DAPO (Yu et al., 2025b) further mitigates entropy collapse in group-wise optimization.

7 Insights and Future Directions

In this section, we discuss structural trade-offs (§7.1) and cross-stage interactions (§7.2) in alignment data pipelines, distill the resulting design principles (§7.3), and outline future challenges (§7.4).

7.1 Core Structural Trade-offs

Many limitations in alignment tuning arise not from individual algorithms, but from structural trade-offs inherent to alignment data pipelines.

Source Fidelity vs. Distribution Shift. Offline supervision using high-quality proprietary models provides a strong, reproducible starting point but lags behind policy evolution, inducing a distribution shift as the student is evaluated on responses it cannot naturally generate. In contrast, online supervision via self-play preserves on-policy fidelity but introduces higher variance, making it prone to error amplification when the initial policy is weak. This choice shapes the exploration space and propagates to evaluation and instantiation by constraining what can be effectively judged and optimized.

Open-Loop vs. Closed-Loop Alignment. Many alignment pipelines operate as if in an open-loop setting, treating preference data and evaluators as static. However, alignment is dynamic and policy-dependent, as evolving outputs continually reshape the data. The choice of loss influences the policy's entropy and output distribution, constraining future exploration. Ignoring this can lead to degeneracy such as reduced diversity or reward hacking, highlighting the need for adaptive mechanisms like response resampling and evaluator recalibration.

Evaluation Granularity vs. Complexity. Shifting from outcome-level to step-level evaluation improves supervision precision for complex reasoning by providing more fine-grained signals. But, this increased granularity introduces substantial labeling complexity and raises the risk of over-optimizing intermediate steps that are not strictly necessary for producing the final correct answer.

Objective Dimensionality vs. Optimization Tax. Collapsing multi-dimensional human values into a single scalar reward simplifies the instantiation stage but introduces an alignment tax, as important nuances across dimensions are lost. In contrast, maintaining multi-objective rubrics better preserves performance across diverse criteria, but significantly complicates the optimization landscape, making convergence slower and less stable.

7.2 Cross-Stage Interactions

Although we analyze the data pipeline as separate stages, alignment behavior ultimately arises from their interactions. The choice at one stage directly shape the effective signal seen by the others, rather than a set of independent components.

First, response synthesis constrains evaluation. If candidate responses lack diversity or informative contrasts, even a high-quality evaluator cannot express meaningful preference margins. Second, preference evaluation must be matched with instantiation. Coarse or scalar judgments can be efficiently instantiated but may obscure trade-offs, while fine-grained or multi-criteria judgments require struc-

tured instantiation to avoid signal distortion. Lastly, preference instantiation feeds back into synthesis in iterative or online settings by reshaping the policy's entropy and output distribution, influencing the diversity and difficulty of future candidates.

These interactions explain why many alignment failures originate at the pipeline level, requiring coordinated design across stages rather than isolated improvements. Details are in Appendix B.

7.3 Pipeline Design Principles

The trade-offs and interactions motivate design principles for alignment data pipelines. Rather than prescribing specific algorithms, these principles characterize how response synthesis, preference evaluation, and preference instantiation should be jointly designed to manage these trade-offs. Based on our analysis of existing methods, we distill six principles to capture effective design patterns. We summarize the principles and their key insights in Table 1, which serves as a unifying reference.

Beyond conceptual principles above, translating these insights into practice requires addressing real-world constraints such as cost, latency, hardware limits, and labeling budgets. To make this survey a practical, we map alignment methods to design scenarios shaped by resource limitations and system-level trade-offs across three stages:

Response Synthesis. The primary function of this stage is to define the behavioral support of the policy. The design choice is governed by data availability (does the data exist?) and inference budget (can we afford to generate it?). Table 4 outlines how different constraints lead to distinct synthesis strategies, each with inherent trade-offs between efficiency, fidelity, and exploration.

Preference Evaluation. The primary function of this stage is to assign preference judgments via an adjudicator. The design choice is governed by task complexity (how hard is it to evaluate?), labeling budget (can we afford high-quality judges?), and objective dimensionality (are there conflicting goals like safety vs. helpfulness?). Table 5 organizes these constraints into corresponding choices of adjudicator and granularity, highlighting the trade-offs between cost and evaluation precision.

Preference Instantiation. The primary function of this stage is to convert evaluated responses into structured training signals. The design choice is governed by hardware constraints (VRAM), task variance (instabilities in optimization), and objective dimensionality (conflicting goals). Table 6 organizes these constraints into corresponding feedback instantiation strategies, highlighting the trade-offs between computational efficiency, optimization stability, and signal expressiveness.

See Appendix C for scenario-driven pipeline configurations that jointly integrate the three stages.

7.4 Future Research Directions

Several core challenges remain unsolved. We highlight four directions for future research.

Prompt-Level Alignment. Alignment must move beyond global preference models toward prompt-specific criteria, as different prompts demand different notions of quality. A key challenge is inferring and enforcing prompt-level alignment without explicit human specification.

Alignment for Agentic Systems. As agentic systems become dominant alignment must move beyond individual decisions to agentic loops. In long horizon workflows failures arise from interactions among planning execution feedback and memory rather than single actions. Existing methods supervise local outputs while leaving global behavior unconstrained. A key challenge is defining trajectory-level alignment objectives that enable agents to revise or abandon plans over time.

Alignment under Evolving Objectives. Future alignment must account for objectives that evolve over time rather than remain fixed. In interactive and agentic settings user intent task goals and acceptable behavior change during deployment. Current pipelines assume static training time preferences. A key challenge is adapting alignment to evolving objectives without destabilizing learned behavior or inducing value drift.

Multi-Modal Alignment. As LLMs evolve into multi-modal models, alignment must extend beyond text. Modalities such as video introduce interacting temporal, spatial, and visual signals that make evaluation substantially more complex. A key challenge is designing alignment pipelines that can robustly interpret feedback across modalities.

8 Conclusion

This survey reframes alignment tuning as a data-centric pipeline design problem. By decomposing alignment into response synthesis, preference evaluation, and preference instantiation, we show how pipeline design governs alignment behavior. Our analysis reveals recurring trade-offs and interactions that guide future alignment pipelines.

Limitations

This survey does not propose or empirically evaluate a new alignment algorithm. Instead, it reanalyzes existing methods through a data-centric lens to clarify how alignment design choices are structured and interact at the pipeline level. Our goal is not to compare or rank methods experimentally, but to provide a coherent conceptual framework for organizing fragmented design decisions across response synthesis, preference evaluation, and preference instantiation. Unified empirical comparisons across these components are inherently difficult, as they are tightly coupled to specific models, datasets, scales, and evaluation protocols.

Additionally, the proposed taxonomy abstracts implementation factors, such as model scale, compute budgets, and domain constraints. While these factors can affect alignment outcomes in practice, they vary post-hoc across systems. By focusing instead on pipeline design choices that precede and constrain implementation, our analysis isolates core alignment mechanisms at the design level, enabling clearer structural characterization despite system-level variation.

Overall, the contribution of this survey lies not in introducing a new optimization technique or benchmark, but in providing a coherent conceptual framework that offers a holistic organization of alignment methodologies across response synthesis, preference evaluation, and preference instantiation, while highlighting recurring trade-offs and interactions across alignment data pipelines.

Use of AI Assistants

We used AI-based tools (GPT-5.2) solely to polish the clarity, grammar, and presentation of the manuscript. All core research contributions, including the identification of key insights, the selection and analysis of prior work, and the organization of the survey's technical content, were conducted manually by the author. Gen AI tools did not contribute to the discovery, interpretation, or synthesis of the surveyed literature.

Acknowledgements

This work was supported by the NRF grant funded by the Korea government (MSIT) (RS-2024-00334343 & RS-2022-NR068758) and the IITP grant funded by the Korea government (MSIT) (No. RS-2024-00445087). For GPU infrastructure, it was supported by the "Advanced GPU Utilization Support Program" funded by the Government of the Republic of Korea (Ministry of Science and ICT) (02-26-01-0181).

References

Akhil Agnihotri, Rahul Jain, Deepak Ramachandran, and Zheng Wen. 2025. Multi-objective preference optimization: Improving human alignment of generative models. *arXiv preprint arXiv:2505.10892*.

Kian Ahrabian, Xihui Lin, Barun Patra, Vishrav Chaudhary, Alon Benhaim, Jay Pujara, and Xia Song. 2025. A practical analysis of human alignment with* po. In *NAACL*.

Alon Albalak, Yanai Elazar, Sang Michael Xie, Shayne Longpre, Nathan Lambert, Xinyi Wang, Niklas Muennighoff, Bairu Hou, Liangming Pan, Haewon Jeong, Colin Raffel, Shiyu Chang, Tatsunori Hashimoto, and William Yang Wang. 2024. A survey on data selection for language models. *Transactions on Machine Learning Research*.

Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. 2024. A general theoretical paradigm to understand learning from human preferences. In *AISTAT*.

Chenjia Bai, Yang Zhang, Shuang Qiu, Qiaosheng Zhang, Kang Xu, and Xuelong Li. 2025. Online preference alignment for language models via count-based exploration. *arXiv preprint arXiv:2501.12735*.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, and 1 others. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Avinandan Bose, Zhihan Xiong, Aadirupa Saha, Simon Shaolei Du, and Maryam Fazel. 2025. Hybrid preference optimization for alignment: Provably faster convergence rates by combining offline preferences with online exploration. In *ICLRW*.

Kianté Brantley, Mingyu Chen, Zhaolin Gao, Jason D Lee, Wen Sun, Wenhao Zhan, and Xuezhou Zhang. 2025. Accelerating rl for llm reasoning with optimal advantage regression. *arXiv preprint arXiv:2505.20686*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, song2024preferenceAmanda Askell, and 1 others. 2020. Language models are few-shot learners. In *NeurIPS*.

Christopher Burges, Robert Ragno, and Quoc Le. 2006. Learning to rank with nonsmooth cost functions. In *NeurIPS*.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, and 1 others. 2023. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*.

Jiaju Chen, Yuxuan Lu, Xiaojie Wang, Huimin Zeng, Jing Huang, Jiri Gesi, Ying Xu, Bingsheng Yao, and Dakuo Wang. 2025a. Multi-agent-as-judge: Aligning llm-agent-based automated evaluation with multi-dimensional human evaluation. *arXiv preprint arXiv:2507.21028*.

Jingfeng Chen, Raghuveer Thirukovalluru, and 1 others. 2025b. Atomic consistency preference optimization for long-form question answering. *arXiv preprint arXiv:2505.09039*.

Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. 2024. Self-play fine-tuning converts weak language models to strong language models. In *ICML*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, and 1 others. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

John Joon Young Chung, Vishakh Padmakumar, Melissa Roemmele, Yuqian Sun, and Max Kreminski. 2025. Modifying large language model post-training for diverse creative writing. *arXiv preprint arXiv:2503.17126*.

Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. 2023. Safe rlhf: Safe reinforcement learning from human feedback. *arXiv preprint arXiv:2310.12773*.

Xun Deng, Han Zhong, Rui Ai, Fuli Feng, Zheng Wang, and Xiangnan He. 2025. Less is more: Improving llm alignment via preference data selection. *arXiv preprint arXiv:2502.14560*.

Yi Dong, Zhilin Wang, Makesh Sreedhar, Xianchao Wu, and Oleksii Kuchaiev. 2023. Steerlm: Attribute conditioned sft as an (user-steerable) alternative to rlhf. In *EMNLP*.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. In *ICML*.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*.

William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39.

Raghav Gupta, Ryan Sullivan, Yunxuan Li, Samrat Phatale, and Abhinav Rastogi. 2025. Robust multi-objective preference alignment with online dpo. In *AAAI*.

Qiang He and Setareh Maghsudi. 2025. Pareto multi-objective alignment for language models. In *ECML-PKDD*.

Jiwoo Hong, Noah Lee, and James Thorne. 2024. Orpo: Monolithic preference optimization without reference model. In *EMNLP*.

Kexin Huang, Junkang Wu, Ziqian Chen, Xue Wang, Jinyang Gao, Bolin Ding, Jiancan Wu, Xiangnan He, and Xiang Wang. 2025. Larger or smaller reward margins to select preferences for llm alignment? In *ICML*.

Mete Ismayilzada, Antonio Laverghetta Jr, Simone A Luchini, Reet Patel, Antoine Bosselut, Lonneke van der Plas, and Roger Beaty. 2025. Creative preference optimization. *arXiv preprint arXiv:2505.14442*.

Joel Jang, Seungone Kim, Bill Yuchen Lin, Yizhong Wang, Jack Hessel, Luke Zettlemoyer, Hannaneh Hajishirzi, Yejin Choi, and Prithviraj Ammanabrolu. 2023. Personalized soups: Personalized large language model alignment via post-hoc parameter merging. *arXiv preprint arXiv:2310.11564*.

Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, and 1 others. 2023. Ai alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852*.

Seungjae Jung, Gunsoo Han, Daniel Wontae Nam, and Kyoung-Woon On. 2025. Binary classifier optimization for large language model alignment. In *ACL*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Saeed Khaki, JinJin Li, Lan Ma, Liu Yang, and Prathap Ramachandra. 2024. Rs-dpo: A hybrid rejection sampling and direct preference optimization method for alignment of large language models. In *NAACL*.

Dahyun Kim, Yungi Kim, Wonho Song, Hyeonwoo Kim, Yunsu Kim, Sanghoon Kim, and Chanjun Park. 2025. sdpo: Don't use your data all at once. In *COLING*.

Kyuyoung Kim, Ah Seo, Hao Liu, Jinwoo Shin, and Kimin Lee. 2024a. Margin matching preference optimization: Enhanced model alignment with granular feedback. In *EMNLP*.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024b. Prometheus 2: An open source language

model specialized in evaluating other language models. In *EMNLP*.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. 2024. Understanding the effects of rlhf on llm generalisation and diversity. In *ICLR*.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *EAMT*.

Jack Lanchantin, Angelica Chen, Shehzaad Dhuliawala, Ping Yu, Jason Weston, Sainbayar Sukhbaatar, and Ilia Kulikov. 2025. Diverse preference optimization. *arXiv preprint arXiv:2501.18101*.

Hung Le, Quan Hung Tran, Dung Nguyen, Kien Do, Saloni Mittal, Kelechi Ogueji, and Svetha Venkatesh. 2025. Multi-reference preference optimization for large language models. In *AAAI*.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Ren Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and 1 others. 2024. Rlaif vs. rlhf: Scaling reinforcement learning from human feedback with ai feedback. In *ICML*.

Chengao Li, Hanyu Zhang, Yunkun Xu, Hongyan Xue, Xiang Ao, and Qing He. 2025a. Gradient-adaptive policy optimization: Towards multi-objective alignment of large language models. *arXiv preprint arXiv:2507.01915*.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, and 1 others. 2025b. From generation to judgment: Opportunities and challenges of llm-as-a-judge. In *EMNLP*.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024. Llms-as-judges: a comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579*.

Lei Li, Hehuan Liu, Yaxin Zhou, ZhaoYang Gui, Xudong Weng, Yi Yuan, Zheng Wei, and Zang Li. 2025c. Uncertainty-aware iterative preference optimization for enhanced llm reasoning. In *ACL*.

Xuying Li, Zhuo Li, Yuji Kosuga, and Victor Bian. 2025d. Optimizing safe and aligned language generation: A multi-objective grpo approach. *arXiv preprint arXiv:2503.21819*.

Yi-Chen Li, Tian Xu, Yang Yu, Xuqin Zhang, Xiong-Hui Chen, Zhongxiang Ling, Ningjing Chao, Lei Yuan, and Zhi-Hua Zhou. 2025e. Generalist reward models: Found inside large language models. *arXiv preprint arXiv:2506.23235*.

Ziniu Li, Tian Xu, Yushun Zhang, Zhihang Lin, Yang Yu, Ruoyu Sun, and Zhi-Quan Luo. 2023. Remax: A simple, effective, and efficient reinforcement learning method for aligning large language models. *arXiv preprint arXiv:2310.10505*.

Yong Lin, Hangyu Lin, Wei Xiong, Shizhe Diao, Jianmeng Liu, Jipeng Zhang, Rui Pan, Haoxiang Wang, Wenbin Hu, Hanning Zhang, and 1 others. 2024. Mitigating the alignment tax of rlhf. In *EMNLP*.

Aiwei Liu, Haoping Bai, Zhiyun Lu, Yanchao Sun, Xiang Kong, Simon Wang, Jiulong Shan, Albin Madappally Jose, Xiaojiang Liu, Lijie Wen, and 1 others. 2025a. Tis-dpo: Token-level importance sampling for direct preference optimization with estimated weights. In *ICLR*.

Bo Liu, Chuanyang Jin, Seungone Kim, Weizhe Yuan, Wenting Zhao, Ilia Kulikov, Xian Li, Sainbayar Sukhbaatar, Jack Lanchantin, and Jason Weston. 2025b. Spice: Self-play in corpus environments improves reasoning. *arXiv preprint arXiv:2510.24684*.

Qian Liu, Xiaosen Zheng, Niklas Muennighoff, Guangtao Zeng, Longxu Dou, Tianyu Pang, Jing Jiang, and Min Lin. 2025c. RegMix: Data mixture as regression for language model pre-training. In *ICLR*.

Shujun Liu, Xiaoyu Shen, Yuhang Lai, Siyuan Wang, Shengbin Yue, Zengfeng Huang, Xuan-Jing Huang, and Zhongyu Wei. 2025d. Haf-rm: A hybrid alignment framework for reward model training. In *ACL)*.

Shunyu Liu, Wenkai Fang, Zetian Hu, Junjie Zhang, Yang Zhou, Kongcheng Zhang, Rongcheng Tu, Ting-En Lin, Fei Huang, Mingli Song, and 1 others. 2025e. A survey of direct preference optimization. *arXiv preprint arXiv:2503.11701*.

Tianci Liu, Ran Xu, Tony Yu, Ilgee Hong, Carl Yang, Tuo Zhao, and Haoyu Wang. 2025f. Openrubrics: Towards scalable synthetic rubric generation for reward modeling and llm alignment. *arXiv preprint arXiv:2510.07743*.

Tianqi Liu, Zhen Qin, Junru Wu, Jiaming Shen, Misha Khalman, Rishabh Joshi, Yao Zhao, Mohammad Saleh, Simon Baumgartner, Jialu Liu, and 1 others. 2025g. Lipo: Listwise preference optimization through learning-to-rank. In *NAACL*.

Yang Liu, Jiahuan Cao, Chongyu Liu, Kai Ding, and Lianwen Jin. 2025h. Datasets for large language models: A comprehensive survey. *Artificial Intelligence Review*, 58(12):403.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *EMNLP*.

Yilun Liu, Shimin Tao, Xiaofeng Zhao, Ming Zhu, Wenbing Ma, Junhao Zhu, Chang Su, Yutai Hou, Miao Zhang, Min Zhang, and 1 others. 2024. Coachlm: Automatic instruction revisions improve the data quality in llm instruction tuning. In *ICDE*. IEEE.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. 2025i. Understanding r1-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*.

Xingzhou Lou, Junge Zhang, Jian Xie, Lifeng Liu, Dong Yan, and Kaiqi Huang. 2024. Spo: Multi-dimensional preference sequential alignment with implicit reward modeling. *arXiv preprint arXiv:2405.12739*.

Liangchen Luo, Yinxiao Liu, Rosanne Liu, Samrat Phatale, Meiqi Guo, Harsh Lara, Yunxuan Li, Lei Shu, Lei Meng, Jiao Sun, and 1 others. 2025. Improve mathematical reasoning in language models with automated process supervision. In *ICLR*.

Saaduddin Mahmud, Mason Nakamura, and Shlomo Zilberstein. 2025. Maple: A framework for active preference learning guided by large language models. In *AAAI*.

Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. Simpo: Simple preference optimization with a reference-free reward. In *NeurIPS*.

Hyangsuk Min, Yuho Lee, Minjeong Ban, Jiaqi Deng, Nicole Hee-Yeon Kim, Taewon Yun, Hang Su, Jason Cai, and Hwanjun Song. 2025. Towards multi-dimensional evaluation of llm summarization across domains and languages. In *ACL*.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large language models: A survey. *arXiv preprint arXiv:2402.06196*.

Subhojyoti Mukherjee, Anusha Lalitha, Sailik Sengupta, Aniket Deshmukh, and Branislav Kveton. 2024. Multi-objective alignment of large language models through hypervolume maximization. *arXiv preprint arXiv:2412.05469*.

William Muldrew, Peter Hayes, Mingtian Zhang, and David Barber. 2024. Active preference learning for large language models. In *ICML*.

Sonia Krishna Murthy, Tomer Ullman, and Jennifer Hu. 2025. One fish, two fish, but not the whole sea: Alignment reduces language models' conceptual diversity. In *NAACL*.

Wojciech Nazar, Grzegorz Nazar, Aleksandra Kamińska, and Ludmila Danilowicz-Szymanowicz. 2025. How to design, create, and evaluate an instruction-tuning dataset for large language model training in health care: Tutorial from a clinical perspective. *Journal of Medical Internet Research*, 27.

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. In *NeurIPS*.

Alexander Pan, Erik Jones, Meena Jagadeesan, and Jacob Steinhardt. 2024. Feedback loops with language models drive in-context reward hacking. In *ICML*.

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.

Shyam Sundhar Ramesh, Yifan Hu, Iason Chaimalas, Viraj Mehta, Pier Giuseppe Sessa, Haitham Bou Ammar, and Ilija Bogunovic. 2024. Group robust preference optimization in reward-free rlhf. In *NeurIPS*.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Feifan Song, Bowen Yu, Minghao Li, Haiyang Yu, Fei Huang, Yongbin Li, and Houfeng Wang. 2024. Preference ranking optimization for human alignment. In *AAAI*.

Hwanjun Song, Taewon Yun, Yuho Lee, Jihwan Oh, Gihun Lee, Jason Cai, and Hang Su. 2025. Learning to summarize from LLM-generated feedback. In *NAACL*.

Taylor Sorensen, Benjamin Newman, Jared Moore, Chan Park, Jillian Fisher, Niloofar Mireshghallah, Liwei Jiang, and Yejin Choi. 2025. Spectrum tuning: Post-training for distributional coverage and in-context steerability.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. In *NeurIPS*.

Haoran Sun, Zekun Zhang, and Shaoning Zeng. 2025. An uncertainty-driven adaptive self-alignment framework for large language models. *arXiv preprint arXiv:2507.17477*.

Zhiqing Sun, Yikang Shen, Qinhong Zhou, Hongxin Zhang, Zhenfang Chen, David Cox, Yiming Yang, and Chuang Gan. 2023. Principle-driven self-alignment of language models from scratch with minimal human supervision. In *NeurIPS*.

Yingshui Tan, Yilei Jiang, Yanshi Li, Jiaheng Liu, Xingyuan Bu, Wenbo Su, Xiangyu Yue, Xiaoyong Zhu, and Bo Zheng. 2025. Equilibrate rlhf: Towards balancing helpfulness-safety trade-off in large language models. *arXiv preprint arXiv:2502.11555*.

Yunhao Tang, Daniel Zhaohan Guo, Zeyu Zheng, Daniele Calandriello, Rémi Munos, Mark Rowland, Pierre Harvey Richemond, Michal Valko, Bernardo Ávila Pires, and Bilal Piot. 2024. Generalized preference optimization: a unified approach to offline alignment. In *ICML*.

Tuhina Tripathi, Manya Wadhwa, Greg Durrett, and Scott Niekum. 2025. Pairwise or pointwise? evaluating feedback protocols for bias in llm-based evaluation. *arXiv preprint arXiv:2504.14716*.

Jianing Wang, Yang Zhou, Xiaocheng Zhang, Mengjiao Bao, and Peng Yan. 2025a. Self-evolutionary large language models through uncertainty-enhanced preference optimization. In *AAAI*.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and 1 others. 2024a. Large language models are not fair evaluators. In *ACL)*.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. 2024b. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. In *ACL*.

Yeyuan Wang, Dehong Gao, Rujiao Long, Lei Yi, Linbo Jin, Libin Yang, and Xiaoyan Cai. 2025b. Aspo: Adaptive sentence-level preference optimization for fine-grained multimodal reasoning. *arXiv preprint arXiv:2505.19100*.

Zhilin Wang, Yi Dong, Olivier Delalleau, Jiaqi Zeng, Gerald Shen, Daniel Egert, Jimmy Zhang, Makesh Narsimhan Sreedhar, and Oleksii Kuchaiev. 2024c. Helpsteer 2: Open-source dataset for training top-performing reward models. In *NeurIPS*.

Zhilin Wang, Yi Dong, Jiaqi Zeng, Virginia Adams, Makesh Narsimhan Sreedhar, Daniel Egert, Olivier Delalleau, Jane Scowcroft, Neel Kant, Aidan Swope, and 1 others. 2024d. Helpsteer: Multi-attribute helpfulness dataset for steerlm. In *NAACL*.

Zhilin Wang, Jiaqi Zeng, Olivier Delalleau, Ellie Evans, Daniel Egert, Hoo-Chang Shin, Felipe Soares, Yi Dong, and Oleksii Kuchaiev. 2025c. Rlbff: Binary flexible feedback to bridge between human feedback and verifiable rewards. *arXiv preprint arXiv:2509.21319*.

Zige Wang, Wanjun Zhong, Yufei Wang, Qi Zhu, Fei Mi, Baojun Wang, Lifeng Shang, Xin Jiang, and Qun Liu. 2023. Data management for training large language models: A survey. *arXiv preprint arXiv:2312.01700*.

Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. 2024. Self-preference bias in llm-as-a-judge. *arXiv preprint arXiv:2410.21819*.

Junkang Wu, Xue Wang, Zhengyi Yang, Jiancan Wu, Jinyang Gao, Bolin Ding, Xiang Wang, and Xiangnan He. 2025a. Alphadpo: Adaptive reward margin for direct preference optimization. In *ICML*.

Xun Wu, Shaohan Huang, Wenhui Wang, Shuming Ma, Li Dong, and Furu Wei. 2024. Multi-head mixture-of-experts. *NeurIPS*.

Yue Wu, Zhiqing Sun, Huizhuo Yuan, Kaixuan Ji, Yiming Yang, and Quanquan Gu. 2025b. Self-play preference optimization for language model alignment. In *ICLR*.

Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A Smith, Mari Ostendorf, and Hannaneh Hajishirzi. 2023. Fine-grained human feedback gives better rewards for language model training. In *NeurIPS*.

Wenyi Xiao, Zechuan Wang, Leilei Gan, Shuai Zhao, Zongrui Li, Ruirui Lei, Wanggui He, Luu Anh Tuan, Long Chen, Hao Jiang, and 1 others. 2024. A comprehensive survey of direct preference optimization: Datasets, theories, variants, and applications. *arXiv preprint arXiv:2410.15595*.

Wei Xiong, Hanze Dong, Chenlu Ye, Ziqi Wang, Han Zhong, Heng Ji, Nan Jiang, and Tong Zhang. 2024. Iterative preference learning from human feedback: Bridging theory and practice for rlhf under kl-constraint. In *ICML*.

Xinyi Xu, Zhaoxuan Wu, Rui Qiao, Arun Verma, Yao Shu, Jingtan Wang, Xinyuan Niu, Zhenfeng He, Jiangwei Chen, Zijian Zhou, and 1 others. 2024. Data-centric ai in the age of large language models. *arXiv preprint arXiv:2406.14473*.

Zhichao Yan, Jiapu Wang, Jiaoyan Chen, Xiaoli Li, Jiye Liang, Ru Li, and Jeff Z Pan. 2025. Atomic fact decomposition helps attributed question answering. *IEEE Transactions on Knowledge and Data Engineering*.

Ziyi Yang, Fanqi Wan, Longguang Zhong, Tianyuan Shi, and Xiaojun Quan. 2025. Weighted-reward preference optimization for implicit model fusion. In *ICLR*.

Daren Yao, Jinsong Yuan, and Ruike Chen. 2025. Enhancing small llm alignment through margin-based objective modifications under resource constraints. *arXiv preprint arXiv:2508.08466*.

Min-Hsuan Yeh, Jeffrey Wang, Xuefeng Du, Seongheon Park, Leitian Tao, Shawn Im, and Yixuan Li. 2025. Position: Challenges and future directions of data-centric ai alignment. In *ICML*.

Ping Yu, Jack Lanchantin, Tianlu Wang, Weizhe Yuan, Olga Golovneva, Ilia Kulikov, Sainbayar Sukhbaatar, Jason Weston, and Jing Xu. 2025a. Cot-self-instruct: Building high-quality synthetic prompts for reasoning and non-reasoning tasks. *arXiv preprint arXiv:2507.23751*.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, and 1 others. 2025b. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*.

Tianshu Yu, Ting-En Lin, Yuchuan Wu, Min Yang, Fei Huang, and Yongbin Li. 2025c. Diverse ai feedback for large language model alignment. *Transactions of the Association for Computational Linguistics*, 13:392–407.

Yue Yu, Zhengxing Chen, Aston Zhang, Liang Tan, Chenguang Zhu, Richard Yuanzhe Pang, Yundi Qian, Xuewei Wang, Suchin Gururangan, Chao Zhang, and 1 others. 2025d. Self-generated critiques boost reward modeling for language models. In *NAACL*.

Hongyi Yuan, Zheng Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. 2023. Rrhf: Rank responses to align language models with human feedback. In *EMNLP*.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason E Weston. 2024. Self-rewarding language models. In *ICLR*.

Thomas Zeng, Shuibai Zhang, Shutong Wu, Christian Classen, Daewon Chae, Ethan Ewer, Minjae Lee, Heeju Kim, Wonjun Kang, Jackson Kunde, Ying Fan, Jungtaek Kim, Hyung Il Koo, Kannan Ramchandran, Dimitris Papailiopoulos, and Kangwook Lee. 2025. VersaPRM: Multi-domain process reward model via synthetic reasoning data. In *ICML*.

Yongcheng Zeng, Guoqing Liu, Weiyu Ma, Ning Yang, Haifeng Zhang, and Jun Wang. 2024. Token-level direct preference optimization. In *ICML*.

Jiayi Zhang, Simon Yu, Derek Chong, Anthony Sicilia, Michael R Tomz, Christopher D Manning, and Weiyan Shi. 2025a. Verbalized sampling: How to mitigate mode collapse and unlock llm diversity. *arXiv preprint arXiv:2510.01171*.

Lunjun Zhang, Arian Hosseini, Hritik Bansal, Mehran Kazemi, Aviral Kumar, and Rishabh Agarwal. 2025b. Generative verifiers: Reward modeling as next-token prediction. In *ICLR*.

Shenao Zhang, Donghan Yu, Hiteshi Sharma, Han Zhong, Zhihan Liu, Ziyi Yang, Shuohang Wang, Hany Hassan Awadalla, and Zhaoran Wang. 2025c. Self-exploring language models: Active preference elicitation for online alignment. *Transactions on Machine Learning Research*.

Yang Zhao, Yixin Wang, and Mingzhang Yin. 2025. Permutative preference alignment from listwise ranking of human judgments. In *EMNLP*.

Yao Zhao, Rishabh Joshi, Tianqi Liu, Misha Khalman, Mohammad Saleh, and Peter J Liu. 2023. Slic-hf: Sequence likelihood calibration with human feedback. *arXiv preprint arXiv:2305.10425*.

Congming Zheng, Jiachen Zhu, Zhuoying Ou, Yuxiang Chen, Kangning Zhang, Rong Shan, Zeyu Zheng, Mengyue Yang, Jianghao Lin, Yong Yu, and 1 others. 2025. A survey of process reward models: From outcome signals to process supervisions for large language models. *arXiv preprint arXiv:2510.08049*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *NeurIPS*.

Wenxuan Zhou, Ravi Agrawal, Shujian Zhang, Sathish Reddy Indurthi, Sanqiang Zhao, Kaiqiang Song, Silei Xu, and Chenguang Zhu. 2024. Wpo: Enhancing rlhf with weighted preference optimization. In *EMNLP*.

Xuanhe Zhou, Junxuan He, Wei Zhou, Haodong Chen, Zirui Tang, Haoyu Zhao, Xin Tong, Guoliang Li, Youmin Chen, Jun Zhou, and 1 others. 2025. A survey of llm x data. *arXiv preprint arXiv:2505.18458*.

Xinlin Zhuang, Jiahui Peng, Ren Ma, Yinfan Wang, Tianyi Bai, Xingjian Wei, Qiu Jiantao, Chi Zhang, Ying Qian, and Conghui He. 2025. Meta-rater: A multi-dimensional data selection method for pre-training language models. In *ACL*.

A Alignment Tuning Algorithms

Proximal Policy Optimization (PPO) is a representative approach for explicitly maximizing a learned reward under the KL-regularized objective in Eq. (1) (Schulman et al., 2017). It proceeds in two stages. First, a reward model $r_\phi(x, y)$ is trained from pairwise comparisons to approximate oracle preferences $r^*(x, y)$. Second, the policy π_θ is optimized with reinforcement learning to maximize the learned reward while staying close to a reference policy π_{ref}:

$$\max_{\pi_\theta} \mathbb{E}_{x,y}\left[r_\phi(x, y) - \beta \log \frac{\pi_\theta(y \mid x)}{\pi_{\text{ref}}(y \mid x)}\right].$$

In practice, PPO performs stable policy updates using a clipped surrogate objective,

$$\mathcal{L}_{\text{PPO}}(\theta) = \mathbb{E}_{x,y}\Big[\min\Big(\rho_\theta(x, y)\,\hat{A}(x, y), \\ \text{clip}(\rho_\theta(x, y), 1-\epsilon, 1+\epsilon)\,\hat{A}(x, y)\Big)\Big].$$

where $\rho_\theta(x, y) = \frac{\pi_\theta(y|x)}{\pi_{\theta_{\text{old}}}(y|x)}$ is the likelihood ratio and $\hat{A}(x, y)$ is an advantage estimate constructed from $r_\phi(x, y)$ (typically with a value baseline). The clipping term constrains the effective step size, preventing large deviations from the reference distribution and improving training stability.

Direct Preference Optimization (DPO) bypasses explicit reward modeling by directly optimizing the KL-regularized objective in Eq. (1) using preference pairs (Rafailov et al., 2023). Given a dataset of comparisons (x, y_w, y_l) where $y_w \succ y_l$, DPO yields a classification-style objective that increases the relative likelihood of preferred responses under π_θ compared to π_{ref}:

$$\mathcal{L}_{\text{DPO}}(\theta) = -\mathbb{E}_{(x,y_w,y_l)}\Big[\\ \log \sigma\Big(\beta\Big(\log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\Big)\Big)\Big]. \tag{4}$$

where $\sigma(\cdot)$ is the logistic sigmoid. This objective can be interpreted as implicitly fitting a reward via the log-likelihood ratio $\log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}$, thereby achieving preference optimization without training a separate reward model.

Group Relative Policy Optimization (GRPO) shifts from pairwise comparisons to group-wise optimization by normalizing candidates sampled from the current policy (Shao et al., 2024). For each prompt x, GRPO samples a set of responses $\{y_i\}_{i=1}^k \sim \pi_\theta(\cdot \mid x)$ and computes a group-relative baseline (*e.g.*, the mean score):

$$\bar{r}(x) = \frac{1}{k}\sum_{i=1}^{k} r(x, y_i),$$

where $r(x, y)$ is a scoring function (often derived from verification) used to rank candidates. The policy is then updated to increase the probability of responses that outperform the group baseline:

$$\max_{\pi_\theta} \mathbb{E}_x \mathbb{E}_{y\sim\pi_\theta(\cdot|x)}\big[\big(r(x, y) - \bar{r}(x)\big)\log \pi_\theta(y \mid x)\big].$$

By using group-relative normalization, GRPO reduces variance and avoids the need for an explicit critic, value networks.

Other Algorithms. Building on these foundational approaches, the alignment literature has expanded into several related families that explore alternative design choices and practical trade-offs.

Methods in the PPO lineage continue to emphasize training stability and sample efficiency under explicit reward maximization, with extensions such as Safe-RLHF (Dai et al., 2023) incorporating safety-aware constraints and ReMax (Li et al., 2023) removing the need for an explicit critic.

The DPO family investigates simplifications of preference optimization under different supervision assumptions: SimPO (Meng et al., 2024) eliminates the reference policy to reduce computational overhead, while KTO (Ethayarajh et al., 2024) relaxes the dependence on paired preference data by leveraging binary feedback.

Finally, the GRPO family generalizes preference optimization to group-wise settings. DAPO (Yu et al., 2025b) mitigates entropy collapse through decoupled clipping, whereas Dr. GRPO (Liu et al., 2025i) improves computational efficiency by removing biased normalization terms.

B Discussion on Cross-Stage Interaction

Although we categorize alignment tuning into three distinct stages, namely response synthesis, preference evaluation, and preference instantiation, alignment outcomes are ultimately determined by the tight coupling and interactions between these stages. As in Sections 4–6, the output of one stage serves as the rigid input constraint for the next, and in iterative settings, the optimization outcome feeds back into the data generation process. This section details the three critical interaction pathways that define the efficacy of the alignment pipeline.

Scenario / Constraints	Response Synthesis	Preference Evaluation	Preference Instantiation	Rationale & Methods
Resource-Constrained *(Low Budget, General Chat)*	**Offline** Use existing off-policy datasets (e.g., UltraFeedback).	**Heuristic / None** Rely on pre-annotated labels; avoid re-labeling.	**Pair-wise (Ref-free)** No separate Reward Model to save VRAM.	**Why:** Maximizes memory efficiency by removing reference model loading. **Methods:** SimPO, ORPO.
Complex Reasoning *(Math, Coding, Logic)*	**Online (Sample)** Generate multiple rollouts per prompt ($N > 1$).	**Step-level / Verifiable** Use compilers or ground-truth checkers.	**Group-wise** Normalize rewards within a group.	**Why:** Group-relative signals reduce variance; step-level supervision localizes logic errors. **Methods:** GRPO, Process Rewards.
Open-Ended Creativity *(Storytelling, Roleplay)*	**Offline + Creativity** High-temp sampling or diversity constraints.	**LLM-as-a-Judge** Use strong models (e.g., GPT-4) to rank.	**List-wise / Pair-wise** Capture nuances beyond binary good/bad.	**Why:** List-wise ranking preserves diversity better than scalar regression; avoids mode collapse. **Methods:** LiPO, DivPO.
Strict Compliance *(Safety, Harmlessness)*	**Offline (Targeted)** Curate red-teaming prompts.	**Multi-Criteria** Separate signals for Helpfulness & Safety.	**Multi-Objective** Optimize utility s.t. safety constraints.	**Why:** Scalarizing conflicting objectives leads to jailbreaks; constrained optimization enforces boundaries. **Methods:** Safe-RLHF, MOPO.
Data-Poor / Cold Start *(New Domain Adaptation)*	**Online (Iterative)** Self-play from current policy.	**LLM-as-a-Judge** Use generic strong LLM to label on-policy data.	**Iterative Pair-wise** Re-train on batches of self-generated data.	**Why:** Bridges distribution shift; model learns from its own winning responses. **Methods:** SPPO, Iterative DPO.

Table 2: Recommended alignment pipeline configurations based on resource and task constraints. We map common deployment scenarios to the most effective combination of pipeline components (response synthesis, evaluation, and instantiation) as discussed in the main text.

B.1 Response Synthesis Constrains Preference Evaluation

The quality of the preference signal is fundamentally bounded by the behavioral support generated during the response synthesis stage. No matter how sophisticated an adjudicator (in preference evaluation stage) is, it cannot extract meaningful signals from a non-informative candidate set.

Informativeness and Selection Efficiency: The synthesis stage must prioritize generating "informative" candidates that provide high learning value. If the synthesis strategy produces only trivial comparisons (*e.g.*, a perfect response versus a nonsensical one) or ambiguous pairs with negligible quality differences, the evaluator cannot extract a strong gradient signal. Consequently, even a ground-truth evaluator becomes ineffective if the candidate set lacks the margins required to distinguish better behaviors from worse ones.

Diversity and Contrastive Information: Alignment tuning relies on contrast. If the synthesis strategy suffers from mode collapse or lacks exploration (as discussed in Section 4.3), the candidate set will consist of semantically identical responses. In this scenario, even a human expert or a strong LLM-as-a-Judge cannot assign a meaningful preference margin, leading to vanishing gradients during optimization.

Distributional Shift and Evaluator Reliability: The reliability of evaluation depends on how the responses are generated. If synthesized responses come from a distribution that differs substantially from the data used to calibrate the evaluator, the evaluator's confidence estimates can become misaligned. As a result, noisy or misleading scores may be injected into the optimization signal.

B.2 Preference Evaluation Dictates Instantiation Capabilities

The granularity and dimensionality of the judgments produced in the preference evaluation stage determine the upper bound of what can be modeled in the instantiation stage. A mismatch here often leads to information loss or signal distortion.

Granularity Mismatch: If the evaluation stage only provides outcome-level binary labels, attempting to use token-level or step-level instantiation methods (*e.g.*, TIS-DPO or ACPO) is ill-posed without additional approximations. Conversely, if detailed step-level critiques are available but the instantiation method aggregates them into a single scalar reward, the dense supervision signal is compressed, losing the ability to penalize specific reasoning errors.

Dimensionality and Alignment Tax: When evaluation captures multi-dimensional trade-offs (*e.g.*, helpfulness vs. safety), but the instantiation layer collapses these into a single scalarized value, it forces the policy to optimize a fixed trade-off. This often results in the "alignment tax," where improvements in one dimension inadvertently degrade performance in another due to the inability of the scalar loss to represent the Pareto frontier.

B.3 Preference Instantiation Reshapes Future Response Synthesis

In iterative or online alignment settings, the pipeline operates as a closed loop. The choice of loss function in the preference instantiation stage directly impacts the policy model's entropy, which in turn defines the search space for the next round of response synthesis.

Mode Collapse and Exploration: Contrastive objectives like DPO effectively increase the likelihood of preferred responses but can rapidly reduce the entropy of the policy. If the instantiation stage enforces margins too aggressively, the updated policy model may lose the diversity required for effective exploration in the subsequent synthesis step. This creates a degenerate cycle where the model stops generating novel negatives, starving the pipeline of the informative data needed for further improvement.

Reward Hacking Dynamics: If the instantiation mechanism has exploitable shortcuts, the policy will learn to maximize the score without improving true quality. Over time, it produces responses that look good according to the reward but are actually poor. This makes evaluation increasingly difficult and often requires the evaluator to be updated or recalibrated.

C Practical Guidelines for Alignment Pipeline Design

While the main body of this survey systematizes the design space of alignment tuning along response synthesis, preference evaluation, and instantiation, practitioners often face a more immediate question: "which pipeline configuration should be selected under concrete operational constraints?" To bridge this gap between conceptual taxonomy and real-world deployment, we provide practical guidelines for alignment pipeline design based on three governing factors: data availability, task complexity, and computational budget.

Table 2 summarizes recommended pipeline configurations for common deployment scenarios. This section explicates the rationale behind each recommendation, clarifying how different pipeline components should be combined in practice.

C.1 Decision Factors

Alignment pipeline design is primarily constrained by three factors: data availability, task complexity, and computational budget. These factors directly shape feasible choices across response synthesis, preference evaluation, and instantiation.

Data Availability. The availability and reliability of supervision signals constrain pipeline design. When large-scale, high-quality preference data exists, offline alignment using pre-collected datasets is often sufficient and cost-effective. In contrast, data-poor or cold-start settings require on-policy data generation and iterative self-improvement, as the model must construct its own supervision signals to overcome distribution shift.

Task Complexity. Task structure determines the appropriate evaluation granularity. Tasks with explicit correctness criteria, such as mathematics, coding, or formal logic, benefit from trajectory-level or verifiable supervision that localizes errors within intermediate steps. Conversely, open-ended tasks lack objective ground-truth and require comparative judgments that preserve diversity and relative quality rather than absolute correctness.

Computational Budget. Compute and memory constraints strongly influence feasible pipeline components. Under limited budgets, pipelines that avoid auxiliary models and online sampling are preferred. When sufficient compute is available, richer evaluation signals and multi-rollout sampling can significantly improve alignment stability and sample efficiency.

C.2 Scenario-Specific Design Rationale

We next explain the design rationale behind each scenario listed in Table 1, clarifying why particular combinations of synthesis, evaluation, and instantiation are well suited to each setting.

Resource-Constrained Setup. In low-budget general chat settings, offline response synthesis using existing off-policy datasets minimizes data collection cost while maintaining reasonable coverage of user behaviors. Heuristic or pre-annotated preference evaluation avoids repeated labeling, while reference-free pair-wise instantiation removes the need for a separate reward or reference model. This configuration maximizes memory efficiency and supports stable alignment under strict constraints.

Complex Reasoning Tasks. For tasks involving structured reasoning, multiple rollouts per prompt are necessary to expose alternative solution trajectories. Trajectory-level or verifier-based evaluation,

such as compilers or ground-truth checkers, provides precise supervision by identifying where reasoning succeeds or fails. Group-wise instantiation normalizes rewards within a cohort of candidates, reducing variance and stabilizing optimization as reasoning difficulty varies across samples.

Open-Ended Creative Generation. Creative tasks such as storytelling or roleplay lack objective correctness and are sensitive to mode collapse. High-temperature or diversity-constrained synthesis encourages stylistic variation, while LLM-as-a-Judge evaluation provides relative quality judgments beyond binary acceptability. List-wise or pair-wise instantiation preserves nuanced preferences among candidates, maintaining diversity more effectively than scalar objectives.

Strict Compliance and Safety Alignment. Safety-critical deployment requires disentangling competing objectives such as helpfulness and harmlessness. Targeted offline synthesis using red-teaming prompts focuses supervision on failure modes. Multi-criteria evaluation separates safety and utility signals, while multi-objective instantiation enforces explicit safety boundaries that scalar objectives often fail to maintain.

Data-Poor and Cold-Start Adaptation. In new domains where external supervision is scarce, iterative online alignment becomes essential. Self-play from the current policy generates on-policy data, which is labeled using a general-purpose strong LLM as a judge. Iterative pair-wise instantiation allows the model to repeatedly learn from its own highest-quality outputs, gradually bridging distribution shift and accelerating adaptation.

C.3 Takeaway

Across scenarios, the key insight is that effective alignment is not determined by a single algorithmic choice, but by the coherent co-design of synthesis, evaluation, and instantiation under practical constraints. Table 1 should therefore be read not as a prescriptive checklist, but as a set of principled templates that can be adapted to the specific operational realities of a given deployment. While these principles are not universally optimal in all settings, they provide a coarse yet informative abstraction of recurring design trade-offs observed across alignment pipelines.

Methodology	Response Synthesis			Preference Evaluation		Preference
	Response Source	Selection Strategy	Creative Exploration	Judgement Granularity	Object Dimensionality	Instantiation
WPO	Offline	Random	None	Outcome-level	Single-dim	Point-wise
WRPO	Offline	Random	None	Outcome-level	Single-dim	Point-wise
TIS-DPO	Offline	Random	None	Atomic-level	Single-dim	Pair-wise
Iterative DPO	Online	Random	None	Outcome-level	Single-dim	Pair-wise
RS-DPO	Online	Margin-based	None	Outcome-level	Single-dim	Pair-wise
SPIN	Online	Random	None	Outcome-level	Single-dim	Pair-wise
SPPO	Online	Random	None	Outcome-level	Single-dim	Group-wise
GRPO	Online	Random	None	Outcome-level	Single-dim	Group-wise
SELM	Online	Random	None	Outcome-level	Single-dim	Pair-wise
BeeS	Offline	Margin-based	None	Outcome-level	Single-dim	Pair-wise
MMPO	Offline	Margin-based	None	Outcome-level	Single-dim	Pair-wise
AlphaDPO	Offline	Margin-based	None	Outcome-level	Single-dim	Pair-wise
APL	Online	Uncertainty-based	None	Outcome-level	Single-dim	Pair-wise
MAPLE	Online	Uncertainty-based	None	Outcome-level	Single-dim	Pair-wise
IUPO	Online	Uncertainty-based	None	Step-level	Single-dim	Pair-wise
UPO	Online	Uncertainty-based	None	Outcome-level	Single-dim	Pair-wise
UDASA	Online	Uncertainty-based	None	Outcome-level	Multi-dim	Pair-wise
Verbalized Sampling	Online	Random	Verbalized Sampling	Outcome-level	Single-dim	Pair-wise
Spectrum Tuning	Offline	Random	Explicit Spanning	Outcome-level	Single-dim	Point-wise
DivPO	Online	Margin-based	Diversity Penalty	Outcome-level	Single-dim	Pair-wise
CRPO	Online	Random	Multi-signal	Outcome-level	Multi-dim	Pair-wise
Meta-Rater	Offline	Random	None	Outcome-level	Multi-dim	Point-wise
OmegaPRM	Offline	Random	None	Step-level	Single-dim	Point-wise
GenRM	Offline	Random	None	Step-level	Single-dim	Point-wise
VersaPRM	Offline	Random	None	Step-level	Single-dim	Point-wise
Math-Shepherd	Offline	Random	None	Step-level	Single-dim	Point-wise
ASPO	Offline	Random	None	Atomic-level	Single-dim	Pair-wise
TDPO	Offline	Random	None	Atomic-level	Single-dim	Pair-wise
SteerLM	Offline	Random	None	Outcome-level	Multi-dim	Point-wise
HelpSteer	Offline	Random	None	Outcome-level	Multi-dim	Point-wise
Prometheus	Offline	Random	None	Outcome-level	Multi-dim	Point-wise
PPO	Online	Random	None	Outcome-level	Single-dim	Point-wise
KTO	Offline	Random	None	Outcome-level	Single-dim	Point-wise
DPO	Offline	Random	None	Outcome-level	Single-dim	Pair-wise
SimPO	Offline	Random	None	Outcome-level	Single-dim	Pair-wise
ORPO	Offline	Random	None	Outcome-level	Single-dim	Pair-wise
IPO	Offline	Random	None	Outcome-level	Single-dim	Pair-wise
RRHF	Offline	Random	None	Outcome-level	Single-dim	Group-wise
PRO	Offline	Random	None	Outcome-level	Single-dim	Group-wise
LiPO	Offline	Random	None	Outcome-level	Single-dim	Group-wise
PPA	Offline	Random	None	Outcome-level	Single-dim	Group-wise

Table 3: Categorization of alignment tuning methods across three data pipeline stages.

Constraint / Scenario	Recommended Strategy	Design Rationale & Trade-offs	Relevant Methods
Low Inference Budget (Cannot afford online generation)	Offline Synthesis (Policy-Aware Reweighting)	Rationale: Uses existing static datasets to save compute. Trade-off: Risk of distribution shift (off-policy); requires reweighting to correct bias.	WPO, WRPO, TIS-DPO
New Domain / Cold Start (No existing preference data)	Online Synthesis (Iterative Self-Play)	Rationale: Essential when no in-domain data exists; the model generates its own training data to bridge distribution shift. Trade-off: High training latency due to iterative generation steps.	SPIN, SPPO, Iterative DPO, SELM
Mode Collapse / Repetitive Outputs (Lack of diversity)	Creative Exploration (Verbalized / Diversity Sampling)	Rationale: Standard sampling narrows the search space; explicit exploration preserves behavioral diversity.	Verbalized Sampling, Spectrum Tuning, DivPO, CRPO

Table 4: Response synthesis strategies under real-world constraints.

Constraint / Scenario	Recommended Adjudicator & Granularity	Design Rationale & Trade-offs	Relevant Methods
Low Labeling Budget (Cannot afford Human/API judges)	Weak Model (Outcome-Level)	Rationale: Cost-effective for simple, general tasks. Trade-off: Noisy signals; often requires aggregation or collective intelligence to be reliable.	Meta-Rater, RLAIF
High Task Complexity (Multi-step reasoning tasks)	Verifiers (Step-Level)	Rationale: Outcome-level labels fail to localize logic errors; step-level signals are required for precise credit assignment. Trade-off: Requires domain-specific verifiers or ground-truth checkers.	Math-Shepherd, OmegaPRM, VersaPRM, GenRM
Conflicting Objectives (e.g., Safety vs. Utility)	Multi-Criteria Rubric (Multi-Dimensional)	Rationale: A single scalar score obscures trade-offs; explicit dimension separation is required to manage competing goals.	Prometheus, HelpSteer, OpenRubrics

Table 5: Preference evaluation strategies under real-world constraints.

Constraint / Scenario	Feedback Instantiation	Design Rationale & Trade-offs	Relevant Methods
Strict Hardware Constraints (Low VRAM)	Pair-wise (Reference-Free)	Rationale: Removes the memory bottleneck and computational overhead of loading a frozen reference model.	DPO, SimPO, ORPO
High Task Variance	Group-wise	Rationale: Normalizing rewards across a group of sampled candidates ($N > 2$) stabilizes gradients and reduces variance compared to pairwise contrasts.	GRPO, LiPO, Dr. GRPO, DAPO
Conflicting Objectives	Multi-Objective	Rationale: Optimizes multiple objectives jointly to model the Pareto frontier directly, rather than collapsing multi-dimensional trade-offs into a single scalar reward.	MOPO, PAMA, MO-DPO

Table 6: Preference instantiation strategies under real-world constraints.