Unsupervised Post-Training of Foundation Models: A Survey

Yijie Xu[1] **Qianyi Cai**[1] **Huizai Yao**[1] **Yili Wang**[1] **Tianfu Wang**[1] **Cehao Yang**[1]
Xingbo Yao[1,3] **Zhiyu Guo**[3] **Aiwei Liu**[4] **Xuming Hu**[1,2,†] **Weiyu Guo**[5,6,†] **Hui Xiong**[1,2,†]

[1]HKUST(GZ) [2]HKUST [3]Xiaohongshu Inc. [4]WeChat, Tencent [5]CUHK [6]AI[2] Robotics
[†] Corresponding authors.

Abstract

Foundation-model post-training usually relies on human labels, preference data, stronger teachers, or executable verifiers. We study *Unsupervised Post-Training* (UPT): update-bearing adaptation on unlabeled inputs whose learning signal is derived from same-lineage model artifacts rather than an external oracle. We catalog 80 strict UPT methods and organize them by the object that supplies the update signal: a prediction statistic, a sample relation, a self-generated target, or an internal evaluator. Beyond inventory, we show how the choice of internal signal and task structure determines whether post-training improves the model or recursively amplifies error. An orthogonal Input Visibility × Update Persistence view maps deployment regimes and defines a unified framework for UPT selection and evaluation.

1 Introduction

Foundation-model post-training has so far followed two waves of external supervision.[1] The first wave is *human labels*, including supervised fine-tuning and preference-based RL. The second is *external verifiers*: math checkers, unit tests, and executable environments that license RL with verifiable rewards (Shao et al., 2024; Zhao et al., 2025; Liao et al., 2026). A third wave, accumulating since 2023 and accelerating through 2025–2026, abandons external supervision and updates the model using *only* unlabeled prompts, text, or target inputs, drawing every update signal from the model's own samples, distributions, judges, or curricula (Huang et al., 2023; Yuan et al., 2024; Zuo et al., 2025; Huang et al., 2025).

UPT enables adaptation when labels and task-specific verifiers cannot be obtained or transferred. This setting covers newly arrived domain corpora and open-ended generation tasks such as dialogue and summarization, where exact-answer checkers are unavailable.

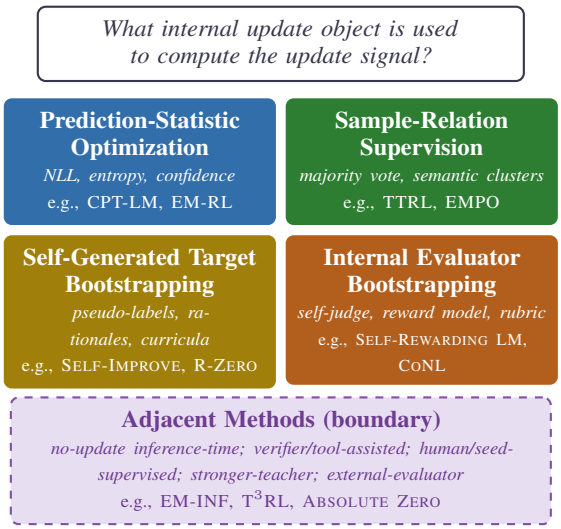

Figure 1: Update-object taxonomy of *Unsupervised Post-Training*. Strict UPT methods are grouped by the internal update object. Dashed boxes denote boundary-adjacent methods that fail at least one strict UPT check.

We use *UPT* for update-bearing adaptation whose learning signal comes from same-lineage model artifacts; §2 operationalizes this scope with four boundary checks. Existing surveys cover neighboring territory: LLM self-improvement (Tao et al., 2024; Kumar et al., 2025), test-time adaptation (Liang et al., 2023), and reinforced reasoning (Xu et al., 2025b; Chen et al., 2025b). None of them, however, enforces the three constraints that define our scope: a real update, no external supervision, and a classification axis defined by the internal object that produces the update signal. Appendix E provides the dimension-wise comparison.

We organize UPT by the internal object consumed by the update: a prediction statistic, a sample relation, a self-generated target, or an internal evaluator. Figure 1 shows the four resulting families: Prediction-Statistic Optimization, Sample-

[1]We use *foundation models* for both text-only and multimodal large models; a recurring finding is that the dominant unsupervised mechanisms are modality-agnostic.

Relation Supervision, Self-Generated Target Bootstrapping, and Internal Evaluator Bootstrapping. A parallel adjacent track maps methods that introduce verifier, tool, seed, teacher, or external-evaluator signals. Together, the four families expose a common error chain: an imperfect proxy selects or rewards outputs, the update concentrates the model on them, and the next round reads an even more biased proxy. The family determines where this loop begins and which safeguard can interrupt it.

Our contributions are: (i) an operational boundary protocol that distinguishes signal provenance from task structure (§2, Appendix A); (ii) an update-object taxonomy of four families and an orthogonal Input Visibility × Update Persistence view (§3–§8), with the hierarchy and inventory cross-sections in Appendix C; and (iii) a cross-family synthesis of applicability, error propagation, and deployment checks, supported by empirical and task-structure audits (§7–§9, Appendix B).[2]

2 Scope, Survey Protocol, and Definitions

Survey protocol. We cover text-only and multimodal foundation-model methods from January 2023 to May 2026 that perform a post-pretraining update on unlabeled prompts, text, or target inputs. Seed-and-snowball search covered continued pretraining, test-time adaptation, internal consensus, self-training, self-rewarding, and multimodal UPT across ACL Anthology, arXiv, Semantic Scholar, and Google Scholar. The frozen inventory contains 94 method records from 91 papers: 80 strict rows from 78 papers, 8 adjacent rows, and 6 prose-only boundary or antecedent records. A paper can contribute more than one method record. Appendix A gives database-specific query templates, criteria, and counts.

Definition. We define *Unsupervised Post-Training* (UPT) as any procedure that (a) begins from a finetuned foundation model, (b) uses unlabeled prompts, text, or target inputs, (c) modifies model parameters, adapters, memories, or persistent local state, and (d) computes the update signal without external supervision. External supervision includes ground-truth answers, verifier feedback, executable or tool verdicts, human labels, and labels from a stronger teacher.

Update objects. We organize UPT with an *update-object taxonomy*. An internal update object is the model-derived object that produces the update signal: a prediction statistic, a relation among model samples, a self-generated target, or an internal evaluator. The taxonomy is orthogonal to optimizer, task, modality, and training schedule.

Boundary checks and update-object rule. A method is *strict UPT* if all four checks hold:

- **B1.** It performs an explicit update to parameters, adapters, memories, or persistent local state.
- **B2.** The update signal is only from unlabeled inputs & same-lineage samples or judgments.
- **B3.** No external supervision enters the update.
- **B4.** Any judge, scorer, or reward model used in the update derives from the same model lineage.

Family assignment follows the object consumed by the gradient. A multi-sample majority statistic used as a reward belongs to Sample-Relation Supervision, whereas one used to build pseudo-labels, curricula, or preference pairs belongs to Self-Generated Target Bootstrapping. The four checks rule out *explicit* external supervision; the separate audit below records structural properties of the task and evaluation protocol.

Task-structure audit. We separately record whether a result relies on an answer extractor or canonicalizer, a finite answer alphabet, a code signature, full-cohort transductive access, or an open-ended output space. These features define equivalence classes that make model samples easier to group and compare; correctness-bearing executions are treated as external verdicts under (B3). Table 6 separates these structural priors from correctness-bearing supervision in the representative-evidence audit presented in Appendix B.

Adjacent methods. The adjacent track retains five neighboring paradigms for comparison: no-update inference-time optimization; verifier- or tool-assisted self-training; human- or seed-supervised bootstrapping; stronger-teacher or cross-model distillation; and external reward or evaluator methods (Liao et al., 2026; Zhao et al., 2025; Wang et al., 2023; Li et al., 2023). This parallel organization preserves a precise internal-signal core while retaining the broader design space.

Formal setup. Let f_θ be a foundation model and let $\mathcal{D}_x$ be a distribution of unlabeled prompts, text, or target inputs. θ denotes the updatable state, including parameters, adapters, memories,

[2] Companion inventory.

	Signal				Mechanism					Regime		
Method	NLL	Ent.	Conf.	Geom./Rule	CPT	TTT	EM-min	EM-RL	State	Train	Test	**LC**
CPT-LM (Ke et al., 2023)	✓				✓					✓		
SIMPLE CPT (Ibrahim et al., 2024)	✓				✓					✓		
LANGADAPT CPT (Elhady et al., 2025)	✓				✓					✓		
STABILITY-GAP CPT (Guo et al., 2025)	✓				✓					✓		
REPLAYALIGN CPT (Abbes et al., 2026)	✓				✓					✓		
E2-LLM (Liu et al., 2024)	✓				✓					✓		✓
DATA ENG 128K (Fu et al., 2024)	✓				✓					✓		✓
LONGCONTEXT SCALING (Xiong et al., 2024)	✓				✓					✓		✓
TLM (Hu et al., 2025a)	✓					✓					✓	
TTT-NN (Hardt and Sun, 2024)	✓					✓					✓	
LONG TTT (Bansal et al., 2025)	✓					✓					✓	✓
IN-PLACE TTT (Feng et al., 2026)	✓					✓					✓	
EM-FT (Agarwal et al., 2025)		✓					✓			✓		
ONE-SHOT EM (Gao et al., 2025b)		✓					✓				✓	
EM-RL(SEQ) (Agarwal et al., 2025)		✓						✓		✓		
EM-RL(TOK) (Agarwal et al., 2025)		✓						✓		✓		
RENT (Prabhudesai et al., 2025)		✓						✓		✓		
RLSC (Li et al., 2025)			✓					✓		✓		
SLOT (Hu et al., 2025b)	✓								✓		✓	
SYTTA (Xu et al., 2025c)			✓						✓		✓	
MODEL WHISPER (Kang et al., 2025)			✓						✓		✓	
ULDTTA (Xu et al., 2026)	✓								✓		✓	
VIGOR (Wen et al., 2026b)				✓				✓		✓		
LATENT-GRPO (Zhang et al., 2026a)				✓				✓		✓		
SSL-R1 (Xie et al., 2026)				✓				✓		✓		
SUDER‡ (Hong et al., 2025)				✓				✓		✓		

Table 1: Strict UPT methods in Family I (§3). *Signal*: NLL, entropy (Ent.), self-confidence (Conf.), or other geometric/rule-based statistic (Geom./Rule). *Mechanism*: continued pretraining (CPT), test-time training (TTT), entropy/confidence minimization (EM-min), entropy/confidence or other internal-statistic policy-gradient reward (EM-RL), sample-local state update (State). *LC*: long-context target. ‡ Family I/IV bridge case (§6).

or persistent local state. A UPT procedure has three components. First, a sampling or aggregation operator $\mathcal{S}$ produces an internal update object $z = \mathcal{S}(f_\theta, x), x \sim \mathcal{D}_x$, such as predictive distribution, a set of rollouts, candidate answers, preference pairs, generated targets, or judge verdicts. Second, a signal extractor σ maps this object to an update signal $s = \sigma(z)$, which may be scalar, pairwise, token or sequence-level. Third, an update operator $\mathcal{U}$ maps these to an updated state:

$$\theta' = \mathcal{U}(\theta; x, z, s).$$

We call z, not $\sigma \circ \mathcal{S}$, the *internal update object*. The update-object taxonomy classifies a method by the type of z used to compute the update signal.

3 Prediction-Statistic Optimization

The first family is direct: σ reads a scalar from f_θ at a single observation, and $\mathcal{U}$ optimizes that scalar. This makes it the historical baseline, with continued pretraining and test-time training predating the other three families. By construction, it generates no pseudo-labels, constructs no preference pairs, and trains no internal judge: the update object is the predictive quantity (token NLL, sequence likelihood, entropy, confidence) or an analogous geometric or rule-based statistic. Table 1 catalogs 26 strict UPT methods along Signal × Mechanism.

Predictive likelihood minimization. The basic instantiation is CPT: minimizing LM loss on unlabeled text. CPT recipes (Ke et al., 2023; Ibrahim et al., 2024) need no annotation when domain shift is the bottleneck; refinements address failure modes (Guo et al., 2025; Abbes et al., 2026; Elhady et al., 2025), and a parallel sub-line scales context windows by corpus engineering alone (Liu et al., 2024; Fu et al., 2024; Xiong et al., 2024). The NLL objective also reaches into TTA: TLM (Hu et al., 2025a) treats the test stream as a CPT corpus; TTT-NN (Hardt and Sun, 2024) restricts each update to a retrieved neighborhood; LONG TTT (Bansal et al., 2025) extends it to long context; IN-PLACE TTT (Feng et al., 2026) interleaves updates with generation.

Entropy and confidence objectives. A second sub-line generalizes the read-out to *predictive confidence*, exploiting that high-quality reasoning traces are low-entropy. Entropy minimization (Agarwal et al., 2025; Gao et al., 2025b) treats per-token entropy as a training loss or single-prompt gradient. The statistics wrap into policy-optimization loops: EM-RL (Agarwal et al., 2025) recasts entropy as intrinsic reward in GRPO; RENT (Prabhudesai et al., 2025) and RLSC (Li et al., 2025) reward self-confidence with no answer key. Zhang et al. (2025e) show that initialization and update duration

Method	Consensus signal						Regime		Multimodal
	Majority vote	Semantic cluster	Self-certainty	Path consist.	Pairwise agreement	Soft cluster	Train	Test	
EMPO (Zhang et al., 2025c)		✓					✓		
INTUITOR (Zhao et al., 2026b)			✓				✓		
COVO (Zhang et al., 2025b)				✓			✓		
CO-REWARDING (Zhang et al., 2026b)					✓		✓		
EVOLMM (Thawakar et al., 2025)	✓						✓		✓
TTRL (Zuo et al., 2025)	✓							✓	
ETTRL (Liu et al., 2025)	✓							✓	
ECHO (Zhao et al., 2026a)	✓†							✓	
SPINE (Wu et al., 2025)	✓†							✓	
SELF-HARMONY (Wang et al., 2026b)					✓			✓	
DARE (Du et al., 2026a)						✓		✓	
SCOPE (Wang et al., 2025a)	✓						✓		
COMPASS (Xing et al., 2025)	✓						✓		
SCRL (Yan et al., 2026)	✓						✓		
RLCCF (Yuan et al., 2025)					✓		✓		
ROIRL (Arzhantsev et al., 2025)	✓						✓		
EVOL-RL (Zhou et al., 2025)	✓						✓		
TTRV (Singh et al., 2025)	✓							✓	✓
MM-UPT (Wei et al., 2025)	✓						✓		✓
DUAL CONSENSUS (Du et al., 2026b)	✓						✓		
CSRS (Yu et al., 2026)						✓	✓		✓
EVOQUALITY (Wen et al., 2026a)					✓		✓		✓

Table 2: Strict UPT methods in Family II (§4). *Consensus signal*: which multi-sample statistic over multiple internal samples drives the gradient. † marks methods where the consensus is the *selection* mask and a separate intrinsic term shapes advantages or tokens.

shape entropy- and confidence-based optimization, motivating stratified evaluation in §9.

Beyond entropy and confidence. A recent line broadens Family I to richer internal statistics (*Geom./Rule* column of Table 1). VIGOR (Wen et al., 2026b) uses policy's teacher-forced gradient norm as intrinsic GRPO reward; LATENT-GRPO (Zhang et al., 2026a) replaces external judges with terminal hidden-state geometry. In multimodal settings, SSL-R1 (Xie et al., 2026) derives rewards from visual self-supervised puzzles, and SUDER (Hong et al., 2025) uses reverse-task likelihood as a prediction statistic bridging Families I and IV (§6).

Sample-local state update. A third sub-line moves the update target from full parameters to a small, sample-local state: a per-prompt optimization vector (Hu et al., 2025b), a test-time LoRA (Xu et al., 2025c), a steering vector (Kang et al., 2025), or layer-wise dynamic adaptation (Xu et al., 2026). These methods optimize the same prediction-statistic objectives but with a smaller update footprint and shorter persistence; we trace this timing axis explicitly in §8.

4 Sample-Relation Supervision

The second family changes what σ computes. Methods in Family I read a scalar from the model's distribution at a single observation; Family II reads a *relation* across multiple internal update objects, such as rollouts, paraphrases, candidate answers, or multiple agents. The internal update object is a *multi-sample statistic*: a cluster mass, a consistency score, a vote, or a contrastive agreement. Table 2 shows that 13/22 methods reduce the relation to a binary majority vote; semantic-cluster, self-certainty, pairwise-agreement, and softened-frequency variants populate the long tail.

4.1 Self-Consistency Within a Single Prompt

The simplest relation operates among multiple samples drawn from the same prompt. EMPO (Zhang et al., 2025c) clusters rollouts and uses cluster mass or semantic entropy as the reward. INTUITOR (Zhao et al., 2026b) replaces external rewards with *self-certainty*, the forward KL to a uniform distribution. COVO (Zhang et al., 2025b) combines path consistency and volatility, while CO-REWARDING (Zhang et al., 2026b) co-evolves a pair of networks and reads their contrastive agreement as the reward, which sidesteps the reward hacking that arises when one network grades itself.

4.2 Consensus and Test-Time RL

A larger sub-line scales the relation across many samples and feeds it back into a policy-optimization loop. The canonical instance is TTRL (Zuo et al., 2025) (Table 2, majority-vote rows): for each test prompt, it draws N rollouts, takes the majority answer as a pseudo-label, and runs a GRPO step against it. Refinements form a coherent line: entropy-regularized explo-

Method	Generated target				Selection criterion										Regime	
	Instr.	Rationale	Curric.	Pref.	Doc.	KB	Cycle	SC	Confid.	Maj.	Debate	MBR	Foresight	Reflect	Train	Test
SELF-TUNING† (Zhang et al., 2025d)	✓				✓										✓	
KBALIGN (Zeng et al., 2025)	✓					✓									✓	
CYCLE-INSTRUCT† (Shen et al., 2025)	✓						✓								✓	
SELF-IMPROVE (Huang et al., 2023)		✓						✓							✓	
QUIET-STAR (Zelikman et al., 2024)		✓													✓	
CONFIDENT ST (Jang et al., 2025)		✓							✓						✓	
GENIUS (Xu et al., 2025a)		✓											✓		✓	
LRM SELF-TRAIN (Shafayat et al., 2025)		✓								✓					✓	
DTE (Srivastava et al., 2025)		✓									✓				✓	
LONGMAGPIE (Gao et al., 2025a)		✓													✓	
LONG SELF-IMPROVE (Li et al., 2024)		✓										✓			✓	
TTCS (Yang et al., 2026a)			✓					✓								✓
DISCTT (Moradi and Mudur, 2026)			✓					✓								✓
TTSR (He et al., 2026)			✓											✓		✓
R-ZERO (Huang et al., 2025)			✓							✓					✓	
SCPO (Prasad et al., 2025)				✓				✓							✓	
MACA (Samanta et al., 2025)				✓							✓				✓	
LONGPO (Chen et al., 2025a)				✓											✓	
RLSF (van Niekerk et al., 2025)				✓					✓						✓	
G-ZERO (Huang et al., 2026)				✓					✓						✓	
QUEST (Song et al., 2026)	✓							✓								✓
V-ZERO (Wang et al., 2026a)	✓									✓					✓	

Table 3: Strict UPT methods in Family III (§5). *Generated target*: instruction/response (Instr.), rationale, curriculum (Curric.), or preference pair (Pref.). *Selection criterion*: document grounding (Doc.), knowledge-base grounding (KB), cycle consistency, self-consistency (SC), internal confidence (Confid.), majority vote (Maj.), multi-agent debate, minimum-Bayes risk (MBR), stepwise foresight (Foresight), self-reflection (Reflect). † marks a stage- or variant-qualified strict entry.

ration (Liu et al., 2025); paraphrase consistency and self-play (Wang et al., 2026b); soft rewards from rollout statistics (Du et al., 2026a); consensus across model populations (Yuan et al., 2025); entropy-shaped advantages (Zhao et al., 2026a); and entropy-band token masks (Wu et al., 2025). The same recipe also runs at training time on unlabeled prompts (Wang et al., 2025a; Xing et al., 2025; Yan et al., 2026). EVOL-RL (Zhou et al., 2025) isolates semantic novelty as a mechanism for preserving rollout diversity under majority-based optimization, exposing a broader diversity–selection trade-off for sample-relation methods.

Follow-up works generalize the relation beyond a single hard-majority signal. DUAL CONSENSUS (Du et al., 2026b) uses an anchor-explorer voting relation to escape spurious majorities; CSRS (Yu et al., 2026) softens hard majority rewards into frequency signals over retraced multimodal reasoning sets, while EVOQUALITY (Wen et al., 2026a) turns VLM pairwise judgments into majority-voted image-quality pseudo-rankings.

4.3 Cross-Modal Extension

The relational recipe transfers to multimodal foundation models through modality-specific reductions of free-form outputs. TTRV (Singh et al., 2025) ports TTRL to vision-language models with a frequency-plus-entropy reward; on ImageNet, InternVL3-8B reaches 99.31% (vs. GPT-4o's 98.30%) and exceeds GPT-4o by 2.3 points across eight benchmarks. MM-UPT (Wei et al., 2025) runs a training-time variant on Qwen2.5-VL with self-generated prompts; EVOLMM (Thawakar et al., 2025) closes a proposer–solver loop driven by multi-rollout consistency. All three derive visual supervision from the model's own rollouts, without image-text labels, captioners, or external verifiers.

5 Self-Generated Target Bootstrapping

The third family *constructs* a trainable target from the model distribution, such as an instruction, a rationale, a plan, a debate trace, a curriculum, or a preference pair. The update operator $\mathcal{U}$ then applies a standard SFT or DPO step against that object. We organize the family by the type of object that the model bootstraps (§5.1–§5.3). Table 3 shows that rationales form the largest branch with 8 methods, followed by instructions and preferences with 5 each and curricula with 4.

5.1 Self-Curated Instructions and Knowledge

The simplest synthetic targets are prompt–response pairs derived from raw documents. SELF-TUNING (Zhang et al., 2025d) turns documents into a staged memorization–comprehension–reflection curriculum: next-token prediction, automatically constructed document tasks, and closed-book reconstruction. Its document-derived self-teaching component falls within the strict core.

KBALIGN (Zeng et al., 2025) self-annotates short- and long-dependency question–answer pairs from a textual knowledge base, tunes on them, and then uses its own predictions and generated correction rationales in later rounds. CYCLE-INSTRUCT (Shen et al., 2025) instead instantiates two transformer models from the same base: a forward model $M_{Q\to A}$ and a backward model $M_{A\to Q}$. They alternate pseudo-answer generation, backward reconstruction training, pseudo-instruction generation, and forward reconstruction training. The shared-base dual loop anchors both reconstruction directions to the original text and limits drift across successive rounds of self-training.

5.2 Self-Trained Rationales and Curricula

A larger sub-line bootstraps reasoning targets. The seminal recipe is SELF-IMPROVE (Huang et al., 2023) (Table 3, rationale rows): generate multiple CoT traces, filter them by self-consistency, and fine-tune on the survivors. Variants extend the recipe to latent thoughts during CPT (Zelikman et al., 2024), confidence-selected traces (Jang et al., 2025), foresight-resampled sequences (Xu et al., 2025a), multi-agent debate trajectories (Srivastava et al., 2025), long-context extensions (Gao et al., 2025a; Li et al., 2024), and R-ZERO (Huang et al., 2025), which splits a base model into a Challenger (proposing problems at the Solver's ability boundary) and a Solver (training on majority-vote pseudo-labels); both the curriculum and the labels are internally synthesized. A test-time variant produces curricula at inference (Yang et al., 2026a; Moradi and Mudur, 2026; He et al., 2026): the machinery uses consensus (Family II in isolation), but the gradient is computed against the synthesized target, so the update-object rule places it in Family III. LRM SELF-TRAIN (Shafayat et al., 2025) follows the same logic: its reward is a binary majority match, but the stated unit of analysis is the pseudo-target that evolves alongside the solver.

A recent line extends self-bootstrapping beyond math curricula. G-ZERO (Huang et al., 2026) runs a proposer–generator loop creating hint-driven preference pairs for open-ended generation without any external judge. QUEST (Song et al., 2026) moves the loop to test time, generating query-conditioned auxiliary problems and fitting a small LoRA adapter before answering; V-ZERO (Wang et al., 2026a) trains a questioner–solver vision-language loop on unlabeled images with its own questions and majority-vote pseudo-labels.

Method	Update	Issue addressed
SELF-REWARDING LM (Yuan et al., 2024)	DPO	baseline self-judge
CREAM (Wang et al., 2025c)	DPO	bias amplification
META-REWARDING (Wu et al., 2024)	DPO	judge saturation
TEMPORAL SRLM (Wang et al., 2025b)	DPO	gradient vanishing
CONL (Sui and Hooi, 2026)	PG	critique-as-reward
RLME (Rentschler and Roberts, 2026)	PG	NL meta-judge
META-TTRL (Tan et al., 2026)	PG	test-time MM rubric
AERO (Gao et al., 2026)	PG	endogenous critique
SELF-JUDGE (Wu et al., 2026)	PG	judge-gated reward
GVU‡ (Pan et al., 2026)	PG	cross-branch evaluator

Table 4: Strict UPT methods in Family IV (§6). ‡ Family I/IV bridge case (discussed below).

5.3 Self-Generated Preference Pairs

A final sub-line synthesizes preference pairs and optimizes with DPO-style objectives: ranking responses by self-consistency (Prasad et al., 2025); multi-agent debate consensus (Samanta et al., 2025); *short-to-long* pairs (same instruction on short vs. long context) for long-context self-evolution (Chen et al., 2025a); and ranking chains of thought by an internal answer-confidence statistic (van Niekerk et al., 2025). CONFIDENT ST and RLSF select SFT or DPO targets by confidence, so the update-object rule assigns both to Family III; Appendix D contrasts evaluator-reward variants.

6 Internal Evaluator Bootstrapping

In Family IV, the model also self-elevates the *judge*: the internal update object is an evaluator (scorer, reward model, or meta-judge) produced and consumed by the same lineage. Where Family III bootstraps a *trainable target* and trains against it, Family IV bootstraps a *trainable verdict-emitter* and trains its outputs through this verdict. Because the verdict can take two natural forms, a pairwise preference or a scalar score, Table 4 splits the 10 methods: 4 DPO variants consume preference pairs from the judge, and 6 policy-gradient variants treat the judge's score as reward. Both sub-lines share one constraint: the evaluator must come from the same model lineage; otherwise Appendix D classifies the method as adjacent under (B4).

6.1 Self-Rewarding via Internal Judges

The line opens with SELF-REWARDING LM (Yuan et al., 2024): one LLM alternates between actor and judge roles; chosen–rejected pairs feed iterative DPO, and both capabilities improve in tandem. Subsequent methods stabilize the self-rewarding loop along three dimensions: CREAM enforces cross-iteration consistency (Wang et al., 2025c), META-REWARDING introduces meta-judgment (Wu et al., 2024), and TEMPORAL

SRLM anchors preferences across model generations (Wang et al., 2025b). CSR (Zhou et al., 2024) extends self-rewarding to vision-language models with a calibrated self-judge; Appendix D audits its external CLIP reward term.

6.2 Evaluator-Driven Policy Optimization

A second sub-line lets the evaluator drive a policy-gradient update directly: structured multi-agent debate with critique-helpfulness as Bradley–Terry-aggregated reward (Sui and Hooi, 2026); an internal evaluator's natural-language meta-judgments ("*correct? / logically consistent?*") as scalar rewards (Rentschler and Roberts, 2026); and a multimodal cohort-visible test-time recipe via a metacognitive introspector (Tan et al., 2026). Family IV routes reward through an explicit evaluator, whereas Family II derives it from multi-sample statistics, a distinction documented in Appendix D.

The same internal-evaluator pattern appears in broader self-evolution loops. AERO (Gao et al., 2026) uses self-generated tasks and counterfactual criticism for KTO-style updates, and SELF-JUDGE (Wu et al., 2026) uses a same-model-lineage frozen judge to modulate actor self-consistency. Two unified multimodal systems sit on the Family I/IV boundary: the update-object rule assigns SUDER (Hong et al., 2025) to Family I and GVU (Pan et al., 2026) to Family IV according to the primary signal each update consumes.

7 Cross-Family Synthesis

Four sources of leverage. UPT converts an internal proxy into an update: Family I exploits predictive statistics; Family II aggregates evidence across samples; Family III builds inspectable targets; and Family IV supplies semantic criteria for open-ended outputs. Increasing semantic flexibility also lengthens the feedback path, making consistency, diversity, and independent evaluation increasingly valuable.

Self-training theory links improvement from unlabeled data to class-consistent neighborhoods and expansion (Wei et al., 2021), while evidence on confirmation bias motivates consistency, diversity, and held-out checks (Arazo et al., 2020). Appendix B maps one representative result per family and signal-shaping task structures to these conditions (Tables 5 and 6).

From signal to update. The four families differ in where uncertainty is converted into a training decision. Family I optimizes a statistic at one model state. LANGADAPT CPT (Elhady et al., 2025) uses token likelihood on domain text, while entropy- and confidence-based variants reshape the predictive distribution; this short path depends on alignment between the statistic and downstream behavior rather than proxy sharpness alone. Family II aggregates rollouts. TTRL (Zuo et al., 2025) canonicalizes sampled answers, turns majority-class membership into a GRPO reward, and updates the policy against that relation. Agreement is informative when errors are diverse and answer equivalence is stable, but self-reinforcing when errors correlate. Family III materializes selected outputs as inspectable, reusable targets. In R-ZERO (Huang et al., 2025), a Challenger proposes problems near the Solver's ability boundary and the Solver trains on synthesized labels. The resulting data can be inspected and reused, while selection errors can persist after entering the target set. Family IV inserts a same-lineage evaluator before the update. SELF-REWARDING LM (Yuan et al., 2024) and CONL (Sui and Hooi, 2026) turn judgments or critiques into rewards for open-ended outputs, while actor–judge drift becomes part of the dynamics. Semantic reach grows with aggregation, target construction, and evaluation, together with the longer feedback path through which errors can recur.

Same observable, different gradient path. Majority vote and confidence recur across families, so classification follows the object consumed by the gradient. TTRL and ROIRL (Arzhantsev et al., 2025) consume majority agreement as reward, placing them in Family II; LRM SELF-TRAIN (Shafayat et al., 2025) uses it to select solutions and computes an SFT loss against the retained target set, placing it in Family III. CONFIDENT ST (Jang et al., 2025) and RLSF (van Niekerk et al., 2025) follow the same target path when confidence selects trajectories before SFT or DPO. Confidence consumed directly as loss or reward is a Family I statistic; a same-lineage evaluator's semantic score consumed as reward is a Family IV object. This gradient-path rule remains stable across labels such as self-training, self-rewarding, and confidence optimization; Appendix D records these assignments and the update objects that justify them across the taxonomy.

Reading reported evidence. Table 5 preserves one mechanism-matched result per family under its original setup. For Family I, LANGADAPT CPT reduces Basque perplexity from 23.64 to 3.35 and

raises aggregate downstream accuracy from 27.43 to 34.14, separating corpus adaptation from task transfer. For Family II, TTRL raises AIME 2024 from 12.9 to 40.2 and MATH-500 from 46.7 to 83.4, while its GPQA result ties majority reward to the target task structure. For Family III, QUIET-STAR (Zelikman et al., 2024) raises zero-shot GSM8K from 5.9 to 10.9 and CommonsenseQA from 36.3 to 47.2, supporting future-token selection of latent reasoning targets. For Family IV, CONL raises AIME 2024 from 60.0 to 76.5 and DeepMath from 70.5 to 87.1, demonstrating a strong same-policy semantic signal. Together, the rows connect the four update objects to corpus fit, answer equivalence, target selection, and evaluator quality; Appendix B retains the source-specific metrics, budgets, and adaptation protocols.

Choosing by available signal. When adaptation data are raw documents and distribution shift is the main problem, Family I is the direct baseline. When multiple samples are affordable and outcomes admit a defensible equivalence relation, Family II can exploit agreement. When generated targets can be inspected before an offline update, Family III offers the clearest data interface. When outputs are open-ended and equality is undefined, Family IV supplies learned semantic criteria through evaluator-mediated rollouts; externally grounded programmatic verifiers remain adjacent.

Safeguards follow the feedback path. Family I safeguards track calibration and downstream quality alongside the optimized statistic, distinguishing a sharper proxy from a better model. Family II tracks sample diversity, wrong-majority frequency, and agreement under alternative canonicalizers. Family III tracks target diversity and refreshes the generated dataset before selection errors persist across rounds. Family IV tracks actor–judge correlation, preference margins, cross-round consistency, and saturation. Across families, frozen baselines, rollback checkpoints, and held-out evaluation test whether gains persist. Each check acts where the internal proxy becomes a training signal, interrupting error accumulation before the next round reads a more biased proxy.

8 Timing of Adaptation

The update-object taxonomy asks what internal object supplies the signal. A timing view asks when target inputs become visible and how long the induced change persists. We call these two axes *Input Visibility* and *Update Persistence*. They are orthogonal to family membership: the same update object can be redeployed across regimes, so timing decides deployment cost while family decides supervision shape. Cross-cutting these two axes yields the regimes of Figure 2, which charts adjacent no-update inference-time optimization that uses similar internal signals but does not satisfy the explicit-update requirement of strict UPT.

Pre-sample regimes. The first four regimes differ in how much of the target distribution the update sees. *Offline corpus UPT* is the broadest, covering continued pretraining, instruction self-curation, reasoning self-training, and offline self-rewarding (e.g., Ke et al., 2023; Yuan et al., 2024; Huang et al., 2025); the target distribution is invisible. *Full-cohort transductive adaptation* updates over the entire target cohort at once (Zuo et al., 2025; Wei et al., 2025; Wang et al., 2025a; Xing et al., 2025; Zhao et al., 2026a; Wang et al., 2026b). *Few-sample target adaptation* sees only a small slice and generalizes to a held-out remainder (Singh et al., 2025; Gao et al., 2025b; Prabhudesai et al., 2025; Li et al., 2025); this is a strictly stronger generalization claim than full-cohort. *Streaming continual adaptation* is properly online: at sample t the model uses only prefix $1{:}t{-}1$, with updates accumulating forward (Hu et al., 2025a; Singh et al., 2025; Hedna et al., 2026; Liu et al., 2026).

Within-sample regimes. The remaining two interleave adaptation with prediction. *Test-time instance adaptation* fits a small update for the current instance and resets at its boundary (Hardt and Sun, 2024; Bansal et al., 2025; Hu et al., 2025b; Xu et al., 2025c, 2026; Kang et al., 2025). *Within-sequence adaptation* goes finer still: the update unfolds across chunks or token-states of one sequence and resets at its boundary (Feng et al., 2026; Sim, 2025); the update is to persistent local state, satisfying (B1). Protocols can instantiate multiple regimes. Figure 2 records each realized protocol; Family II $\times$ full-cohort transductive is the densest observed cell, exemplified by TTRL.

9 Challenges and Future Directions

Recursive error propagation is the central open problem: updates reinforce misordered outputs and shift the distribution from which the next proxy is computed. The four families expose different links in this chain and suggest distinct research priorities.

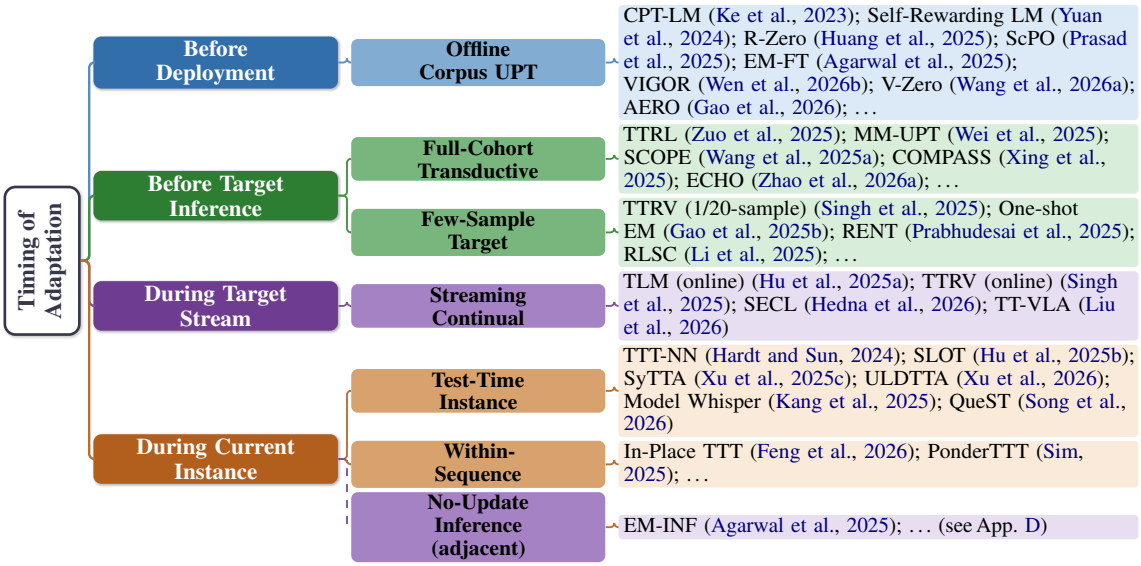

Figure 2: Timing of Adaptation, organized by Input Visibility × Update Persistence and orthogonal to Figure 1. Each leaf names representative methods; protocols spanning multiple regimes appear in more than one leaf. Adjacent no-update inference-time methods are shown for boundary clarity but are not counted as strict UPT.

Separate signal quality from task structure. Consensus can be informative because errors cancel, but only after an extractor defines which outputs agree. Boxed answers, finite choice sets, and code signatures provide such structure without determining correctness; execution and unit tests return a verdict and therefore cross (B3). Future evaluations should report the fields in Table 6 and test whether gains survive alternative canonicalizers and open-ended reformulations. This matters most for Families II–III; meanwhile, open-ended methods such as G-ZERO (Huang et al., 2026) probe the brittle boundary of equality-based consensus.

Interrupt confidence and majority amplification. Family I can turn low entropy into overconfidence; Family II can turn a popular error into a training reward. Among the 22 Family II methods, 13 use hard majority signals, so evaluations should include diversity and wrong-majority diagnostics. EVOL-RL (Zhou et al., 2025) adds semantic novelty, while DUAL CONSENSUS (Du et al., 2026b) and CSRS (Yu et al., 2026) soften or diversify consensus. T^{3}RL (Liao et al., 2026) shows that an independent execution channel can break false-popular modes, motivating independent internal views for strict UPT.

Measure target and judge drift across rounds. Family III freezes selected generations into training targets, so confirmation bias can survive even when the next sampling round looks more confident. Family IV adds evaluator drift: bias amplification, judge saturation, and weak chosen–rejected margins are addressed separately by CREAM, Meta-Rewarding, and Temporal SRLM (Wang et al., 2025c; Wu et al., 2024; Wang et al., 2025b). Across these interventions, a common longitudinal protocol would track target diversity, actor–judge correlation, held-out quality, and perturbation recovery.

Control initialization and timing. Intrinsic-feedback results depend on the starting checkpoint and update duration. Zhang et al. (2025e) show that these factors materially change the outcome of intrinsic-feedback training. Cross-family benchmarks should fix backbone, starting checkpoint, target split, rollout and update budgets, task-structure prior, and held-out evaluation to isolate the effects of signal family, initialization, and timing.

10 Conclusion

We survey 80 strict UPT methods around a single question: which model-derived object supplies the update signal? The resulting taxonomy has four families, while Input Visibility and Update Persistence provide an orthogonal map of when target inputs become visible and how long updates persist. Together, these views clarify UPT's central trade-off: more semantically expressive signals support open-ended outputs but create longer feedback paths through which proxy errors can be reinforced. Method selection and evaluation should therefore align the update object with task structure and deployment timing, and place safeguards where the proxy enters the update.

Limitations

The inventory is frozen in May 2026 and therefore excludes later work. Representative results retain each source paper's backbone, budget, metric, and adaptation protocol, supporting mechanism-level synthesis rather than pooled effect-size estimation. Hybrid methods follow the primary update object consumed by the gradient during adaptation. Finally, (B1)–(B4) constrain the update rule rather than the deployment stack. Because no external oracle scores the updated model, reward hacking and silent per-problem degradation can accumulate unobserved, so independent held-out evaluation, monitoring, and red-teaming remain necessary and are not excluded by the strict boundary; we organize such safeguards by feedback path (§7) but do not evaluate their effectiveness, which requires controlled study.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Grant Nos. 92370204 and 62506318); in part by the National Key R&D Program of China (Grant No. 2023YFF0725001); in part by the Guangdong Provincial Key Laboratory of Frontier Basic Science for All-domain Intelligence; in part by the Guangdong Basic and Applied Basic Research Foundation (Grant No. 2023B1515120057); in part by the Key-Area Special Project of Guangdong Provincial Ordinary Universities (Grant No. 2024ZDZX1007); in part by the Guangdong Provincial Department of Education Project (Grant No. 2024KQNCX028); in part by the Scientific Research Projects for the Higher-educational Institutions, Education Bureau of Guangzhou Municipality (Grant No. 2024312096); and in part by the Guangzhou-HKUST(GZ) Joint Funding Program, Education Bureau of Guangzhou Municipality (Grant No. 2025A03J3957).

References

Istabrak Abbes, Gopeshh Subbaraj, Matthew Riemer, Nizar Islah, Tsuguchika Tabaru, Hiroaki Kingetsu, Sarath Chandar, and Irina Rish. 2026. Revisiting replay and gradient alignment for continual pretraining of large language models. In *Proceedings of The 4th Conference on Lifelong Learning Agents*, volume 330 of *Proceedings of Machine Learning Research*, pages 465–486. PMLR.

Shivam Agarwal, Zimin Zhang, Lifan Yuan, Jiawei Han, and Hao Peng. 2025. The unreasonable effectiveness of entropy minimization in LLM reasoning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Eric Arazo, Diego Ortego, Paul Albert, Noel E. O'Connor, and Kevin McGuinness. 2020. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *2020 International Joint Conference on Neural Networks*, pages 1–8.

Aleksei Arzhantsev, Otmane Sakhi, and Flavian Vasile. 2025. RoiRL: Efficient, self-supervised reasoning with offline iterative reinforcement learning. In *NeurIPS 2025 Workshop on Efficient Reasoning*.

Rachit Bansal, Aston Zhang, Rishabh Tiwari, Lovish Madaan, Sai Surya Duvvuri, Devvrit Khatri, David Brandfonbrener, David Alvarez-Melis, Prajjwal Bhargava, Mihir Sanjay Kale, and Samy Jelassi. 2025. Let's (not) just put things in context: Test-time training for long-context llms. *Preprint*, arXiv:2512.13898.

Guanzheng Chen, Xin Li, Michael Shieh, and Lidong Bing. 2025a. LongPO: Long context self-evolution of large language models through short-to-long preference optimization. In *The Thirteenth International Conference on Learning Representations*.

Qiguang Chen, Libo Qin, Jinhao Liu, Dengyun Peng, Jiannan Guan, Peng Wang, Mengkang Hu, Yuhang Zhou, Te Gao, and Wanxiang Che. 2025b. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. *Preprint*, arXiv:2503.09567.

Lu Dai, Yijie Xu, Jinhui Ye, Hao Liu, and Hui Xiong. 2025. Seper: Measure retrieval utility through the lens of semantic perplexity reduction. In *International Conference on Learning Representations (ICLR)*.

Shijian Deng, Kai Wang, Tianyu Yang, Harsh Singh, and Yapeng Tian. 2025. Self-improvement in multimodal large language models: A survey. *Preprint*, arXiv:2510.02665.

Bodong Du, Xuanqi Huang, and Xiaomeng Li. 2026a. Distribution-aware reward estimation for test-time reinforcement learning. *Preprint*, arXiv:2601.21804.

Kaixuan Du, Meng Cao, Hang Zhang, Yukun Wang, Xiangzhou Huang, and Ni Li. 2026b. Dual consensus: Escaping from spurious majority in unsupervised rlvr via two-stage vote mechanism. *Preprint*, arXiv:2603.16223.

Ahmed Elhady, Eneko Agirre, and Mikel Artetxe. 2025. Emergent abilities of large language models under continued pre-training for language adaptation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32174–32186, Vienna, Austria. Association for Computational Linguistics.

Guhao Feng, Shengjie Luo, Kai Hua, Ge Zhang, Wenhao Huang, Di He, and Tianle Cai. 2026. In-place test-time training. In *The Fourteenth International Conference on Learning Representations*.

Yao Fu, Rameswar Panda, Xinyao Niu, Xiang Yue, Hannaneh Hajishirzi, Yoon Kim, and Hao Peng. 2024. Data engineering for scaling language models to 128k context. *Preprint*, arXiv:2402.10171.

Chaochen Gao, Xing Wu, Zijia Lin, Debing Zhang, and Songlin Hu. 2025a. Longmagpie: A self-synthesis method for generating large-scale long-context instructions. *Preprint*, arXiv:2505.17134.

Zhitao Gao, Jie Ma, Xuhong Li, Pengyu Li, Ning Qu, Yaqiang Wu, Hui Liu, and Jun Liu. 2026. Aero: Autonomous evolutionary reasoning optimization via endogenous dual-loop feedback. *Preprint*, arXiv:2602.03084.

Zitian Gao, Lynx Chen, Haoming Luo, Joey Zhou, and Bryan Dai. 2025b. One-shot entropy minimization. *Preprint*, arXiv:2505.20282. Work in progress.

Yiduo Guo, Jie Fu, Huishuai Zhang, and Dongyan Zhao. 2025. Efficient domain continual pretraining by mitigating the stability gap. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32850–32870, Vienna, Austria. Association for Computational Linguistics.

Moritz Hardt and Yu Sun. 2024. Test-time training on nearest neighbors for large language models. In *The Twelfth International Conference on Learning Representations*.

Haoyang He, Zihua Rong, Liangjie Zhao, Yunjia Zhao, Lan Yang, and Honggang Zhang. 2026. Ttsr: Test-time self-reflection for continual reasoning improvement. *Preprint*, arXiv:2603.03297.

Mohamed Rissal Hedna, Jan Strich, Martin Semmann, and Chris Biemann. 2026. Self-calibrating language models via test-time discriminative distillation. *Preprint*, arXiv:2604.09624.

Jixiang Hong, Yiran Zhang, Guanzhong Wang, Yi Liu, Ji-Rong Wen, and Rui Yan. 2025. Suder: Self-improving unified large multimodal models for understanding and generation with dual self-rewards. *Preprint*, arXiv:2506.07963.

Jinwu Hu, Zitian Zhang, Guohao Chen, Xutao Wen, Chao Shuai, Wei Luo, Bin Xiao, Yuanqing Li, and Mingkui Tan. 2025a. Test-time learning for large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 24823–24849. PMLR.

Yang Hu, Xingyu Zhang, Xueji Fang, Zhiyang Chen, Xiao Wang, Huatian Zhang, and Guojun Qi. 2025b. Slot: Sample-specific language model optimization at test-time. *Preprint*, arXiv:2505.12392.

Chengsong Huang, Haolin Liu, Tong Zheng, Runpeng Dai, Langlin Huang, Jinyuan Li, Zongxia Li, Zhepei Wei, Yu Meng, and Jiaxin Huang. 2026. G-zero: Self-play for open-ended generation from zero data. *Preprint*, arXiv:2605.09959.

Chengsong Huang, Wenhao Yu, Xiaoyang Wang, Hongming Zhang, Zongxia Li, Ruosen Li, Jiaxin Huang, Haitao Mi, and Dong Yu. 2025. R-zero: Self-evolving reasoning llm from zero data. *Preprint*, arXiv:2508.05004.

Jiaxin Huang, Shixiang Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. 2023. Large language models can self-improve. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1051–1068, Singapore. Association for Computational Linguistics.

Adam Ibrahim, Benjamin Thérien, Kshitij Gupta, Mats L. Richter, Quentin Anthony, Timothée Lesort, Eugene Belilovsky, and Irina Rish. 2024. Simple and scalable strategies to continually pre-train large language models. *Preprint*, arXiv:2403.08763.

Hyosoon Jang, Yunhui Jang, Sungjae Lee, Jungseul Ok, and Sungsoo Ahn. 2025. Self-training large language models with confident reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 14925–14939, Suzhou, China. Association for Computational Linguistics.

Xinyue Kang, Diwei Shi, and Li Chen. 2025. Model whisper: Steering vectors unlock large language models' potential in test-time. *Preprint*, arXiv:2512.04748.

Zixuan Ke, Yijia Shao, Haowei Lin, Tatsuya Konishi, Gyuhak Kim, and Bing Liu. 2023. Continual pre-training of language models. *Preprint*, arXiv:2302.03241.

Komal Kumar, Tajamul Ashraf, Omkar Thawakar, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, Phillip H.S. Torr, Fahad Shahbaz Khan, and Salman Khan. 2025. Llm post-training: A deep dive into reasoning large language models. *Preprint*, arXiv:2502.21321.

Pengteng Li, Pinhao Song, Wuyang Li, Huizai Yao, Weiyu Guo, Yijie Xu, Dugang Liu, and Hui Xiong. 2026. See&trek: Training-free spatial prompting for multimodal large language model. *Advances in Neural Information Processing Systems*, 38:62098–62131.

Pengyi Li, Matvey Skripkin, Alexander Zubrey, Andrey Kuznetsov, and Ivan Oseledets. 2025. Confidence is all you need: Few-shot rl fine-tuning of language models. *Preprint*, arXiv:2506.06395.

Siheng Li, Cheng Yang, Zesen Cheng, Lemao Liu, Mo Yu, Yujiu Yang, and Wai Lam. 2024. Large language models can self-improve in long-context reasoning. *Preprint*, arXiv:2411.08147.

Xian Li, Ping Yu, Chunting Zhou, Timo Schick, Omer Levy, Luke Zettlemoyer, Jason Weston, and Mike Lewis. 2023. Self-alignment with instruction backtranslation. *Preprint*, arXiv:2308.06259.

Jian Liang, Ran He, and Tieniu Tan. 2023. A comprehensive survey on test-time adaptation under distribution shifts. *Preprint*, arXiv:2303.15361.

Xun Liang, Shichao Song, Zifan Zheng, Hanyu Wang, Qingchen Yu, Xunkai Li, Rong-Hua Li, Yi Wang, Zhonghao Wang, Feiyu Xiong, and Zhiyu Li. 2024. Internal consistency and self-feedback in large language models: A survey. *Preprint*, arXiv:2407.14507.

Ruotong Liao, Nikolai Röhrich, Xiaohan Wang, Yuhui Zhang, Yasaman Samadzadeh, Volker Tresp, and Serena Yeung-Levy. 2026. Tool verification for test-time reinforcement learning. *Preprint*, arXiv:2603.02203.

Changyu Liu, Yiyang Liu, Taowen Wang, Qiao Zhuang, James Chenhao Liang, Wenhao Yang, Renjing Xu, Qifan Wang, Dongfang Liu, and Cheng Han. 2026. On-the-fly vla adaptation via test-time reinforcement learning. *Preprint*, arXiv:2601.06748.

Jia Liu, ChangYi He, YingQiao Lin, MingMin Yang, FeiYang Shen, and ShaoGuo Liu. 2025. Ettrl: Balancing exploration and exploitation in llm test-time reinforcement learning via entropy mechanism. *Preprint*, arXiv:2508.11356.

Jiaheng Liu, Zhiqi Bai, Yuanxing Zhang, Chenchen Zhang, Yu Zhang, Ge Zhang, Jiakai Wang, Haoran Que, Yukang Chen, Wenbo Su, Tiezheng Ge, Jie Fu, Wenhu Chen, and Bo Zheng. 2024. E2-LLM: Efficient and extreme length extension of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4243–4253, Bangkok, Thailand. Association for Computational Linguistics.

Mohammad Mahdi Moradi and Sudhir Mudur. 2026. Disctt: Consensus-guided self-curriculum for efficient test-time adaptation in reasoning. *Preprint*, arXiv:2603.05357.

Tergel Munkhbat, Namgyu Ho, Seo Hyun Kim, Yongjin Yang, Yujin Kim, and Se-Young Yun. 2025. Self-training elicits concise reasoning in large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25127–25152, Vienna, Austria. Association for Computational Linguistics.

Jiadong Pan, Liang Li, Yuxin Peng, Yu-Ming Tang, Shuohuan Wang, Yu Sun, Hua Wu, Qingming Huang, and Haifeng Wang. 2026. Learning to generate via understanding: Understanding-driven intrinsic rewarding for unified multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22174–22184.

Mihir Prabhudesai, Lili Chen, Alex Ippoliti, Katerina Fragkiadaki, Hao Liu, and Deepak Pathak. 2025. Maximizing confidence alone improves reasoning. *Preprint*, arXiv:2505.22660.

Archiki Prasad, Weizhe Yuan, Richard Yuanzhe Pang, Jing Xu, Maryam Fazel-Zarandi, Mohit Bansal, Sainbayar Sukhbaatar, Jason E Weston, and Jane Yu. 2025. Self-consistency preference optimization. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 49737–49751. PMLR.

Micah Rentschler and Jesse Roberts. 2026. Reinforcement learning from meta-evaluation: Aligning language models without ground-truth labels. *Preprint*, arXiv:2601.21268.

Ankur Samanta, Akshayaa Magesh, Runzhe Wu, Ayush Jain, Youliang Yu, Daniel Jiang, Boris Vidolov, Paul Sajda, Yonathan Efroni, and Kaveh Hassani. 2025. Self-improvement of language models by post-training on multi-agent debate. *Preprint*, arXiv:2509.15172.

Sheikh Shafayat, Fahim Tajwar, Ruslan Salakhutdinov, Jeff Schneider, and Andrea Zanette. 2025. Can large reasoning models self-train? *Preprint*, arXiv:2505.21444.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y.K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

Zhanming Shen, Hao Chen, Yulei Tang, Shaolin Zhu, Wentao Ye, Xiaomeng Hu, Haobo Wang, Gang Chen, and Junbo Zhao. 2025. CYCLE-INSTRUCT: Fully seed-free instruction tuning via dual self-training and cycle consistency. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 5123–5137, Suzhou, China. Association for Computational Linguistics.

Gihyeon Sim. 2025. When to ponder: Adaptive compute allocation for code generation via test-time training. *Preprint*, arXiv:2601.00894.

Akshit Singh, Shyam Marjit, Wei Lin, Paul Gavrikov, Serena Yeung-Levy, Hilde Kuehne, Rogerio Feris, Sivan Doveh, James Glass, and M. Jehanzeb Mirza. 2025. Ttrv: Test-time reinforcement learning for vision language models. *Preprint*, arXiv:2510.06783.

Chaehee Song, Minseok Seo, Yeeun Seong, Doyi Kim, and Changick Kim. 2026. Query-conditioned test-time self-training for large language models. *Preprint*, arXiv:2605.13369.

Gaurav Srivastava, Zhenyu Bi, Meng Lu, and Xuan Wang. 2025. DEBATE, TRAIN, EVOLVE: Self-Evolution of language model reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32764–32810, Suzhou, China. Association for Computational Linguistics.

Yuan Sui and Bryan Hooi. 2026. Conversation for non-verifiable learning: Self-evolving LLMs through meta-evaluation. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*.

Lit Sin Tan, Junzhe Chen, Xiaolong Fu, Lichen Ma, Junshi Huang, Jianzhong Shi, Yan Li, and Lijie Wen. 2026. Meta-ttrl: A metacognitive framework for self-improving test-time reinforcement learning in unified multimodal models. *Preprint*, arXiv:2603.15724.

Zhengwei Tao, Ting-En Lin, Xiancai Chen, Hangyu Li, Yuchuan Wu, Yongbin Li, Zhi Jin, Fei Huang, Dacheng Tao, and Jingren Zhou. 2024. A survey on self-evolution of large language models. *Preprint*, arXiv:2404.14387.

Omkar Thawakar, Shravan Venkatraman, Ritesh Thawkar, Abdelrahman Shaker, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, and Fahad Khan. 2025. Evolmm: Self-evolving large multimodal models with continuous rewards. *Preprint*, arXiv:2511.16672.

Guiyao Tie, Zeli Zhao, Dingjie Song, Fuyang Wei, Rong Zhou, Yurou Dai, Wen Yin, Zhejian Yang, Jiangyue Yan, Yao Su, Zhenhan Dai, Yifeng Xie, Yihan Cao, Lichao Sun, Pan Zhou, Lifang He, Hechang Chen, Yu Zhang, Qingsong Wen, and 7 others. 2025. Large language models post-training: Surveying techniques from alignment to reasoning. *Preprint*, arXiv:2503.06072.

Carel van Niekerk, Renato Vukovic, Benjamin Matthias Ruppik, Hsien chin Lin, and Milica Gašić. 2025. Post-training large language models via reinforcement learning from self-feedback. *Preprint*, arXiv:2507.21931.

Han Wang, Yi Yang, Jingyuan Hu, Minfeng Zhu, and Wei Chen. 2026a. V-zero: Self-improving multimodal reasoning with zero annotation. *Preprint*, arXiv:2601.10094.

Ru Wang, Wei Huang, Qi Cao, Yusuke Iwasawa, Yutaka Matsuo, and Jiaxian Guo. 2026b. SELF-HARMONY: LEARNING TO HARMONIZE SELF-SUPERVISION AND SELF-PLAY IN TEST-TIME REINFORCEMENT LEARNING. In *The Fourteenth International Conference on Learning Representations*.

Weiqin Wang, Yile Wang, Kehao Chen, and Hui Huang. 2025a. Beyond majority voting: Towards fine-grained and more reliable reward signal for test-time reinforcement learning. *Preprint*, arXiv:2512.15146.

Yidong Wang, Xin Wang, Cunxiang Wang, Junfeng Fang, Qiufeng Wang, Jianing Chu, Xuran Meng, Shuxun Yang, Libo Qin, Yue Zhang, Wei Ye, and Shikun Zhang. 2025b. Temporal self-rewarding language models: Decoupling chosen-rejected via past-future. *Preprint*, arXiv:2508.06026.

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

Zhaoyang Wang, Weilei He, Zhiyuan Liang, Xuchao Zhang, Chetan Bansal, Ying Wei, Weitong Zhang, and Huaxiu Yao. 2025c. CREAM: Consistency regularized self-rewarding language models. In *The Thirteenth International Conference on Learning Representations*.

Colin Wei, Kendrick Shen, Yining Chen, and Tengyu Ma. 2021. Theoretical analysis of self-training with deep networks on unlabeled data. In *International Conference on Learning Representations*.

Lai Wei, Yuting Li, Chen Wang, Yue Wang, Linghe Kong, Weiran Huang, and Lichao Sun. 2025. First SFT, second RL, third UPT: Continual improving multi-modal LLM reasoning via unsupervised post-training. In *Advances in Neural Information Processing Systems*, volume 38.

Wen Wen, tianwu zhi, Kanglong FAN, Yang Li, Xinge Peng, Yabin ZHANG, Yiting Liao, Junlin Li, and Li zhang. 2026a. Self-evolving vision-language models for image quality assessment via voting and ranking. In *The Fourteenth International Conference on Learning Representations*.

Xuexiang Wen, Hang Yu, Linchao Zhu, and Gaoang Wang. 2026b. Verifier-free rl for llms via intrinsic gradient-norm reward. *Preprint*, arXiv:2605.09920.

Jianghao Wu, Yasmeen George, Jin Ye, Yicheng Wu, Daniel F. Schmidt, and Jianfei Cai. 2025. Spine: Token-selective test-time reinforcement learning with entropy-band regularization. *Preprint*, arXiv:2511.17938.

Tianhao Wu, Weizhe Yuan, Olga Golovneva, Jing Xu, Yuandong Tian, Jiantao Jiao, Jason Weston, and Sainbayar Sukhbaatar. 2024. Meta-rewarding language models: Self-improving alignment with llm-as-a-meta-judge. *Preprint*, arXiv:2407.19594.

Xiaobao Wu. 2025. Sailing by the stars: A survey on reward models and learning strategies for learning from rewards. *Preprint*, arXiv:2505.02686.

Zhengxian Wu, Kai Shi, Chuanrui Zhang, Zirui Liao, Jun Yang, Ni Yang, Qiuying Peng, Luyuan Zhang, Hangrui Xu, Tianhuang Su, Zhenyu Yang, Haonan Lu, and Haoqian Wang. 2026. When models judge themselves: Unsupervised self-evolution for multimodal reasoning. *Preprint*, arXiv:2603.21289.

Jiahao Xie, Alessio Tonioni, Nathalie Rauschmayr, Federico Tombari, and Bernt Schiele. 2026. Ssl-r1: Self-supervised visual reinforcement post-training for multimodal large language models. *Preprint*, arXiv:2604.20705.

Jingyu Xing, Chenwei Tang, Xinyu Liu, Deng Xiong, Shudong Huang, Wei Ju, Jiancheng Lv, and Ziyue Qiao. 2025. Rewarding the journey, not just the destination: A composite path and answer self-scoring reward mechanism for test-time reinforcement learning. *Preprint*, arXiv:2510.17923.

Wenhan Xiong, Jingyu Liu, Igor Molybog, Hejia Zhang, Prajjwal Bhargava, Rui Hou, Louis Martin, Rashi Rungta, Karthik Abinav Sankararaman, Barlas Oguz, Madian Khabsa, Han Fang, Yashar Mehdad, Sharan Narang, Kshitiz Malik, Angela Fan, Shruti Bhosale, Sergey Edunov, Mike Lewis, and 2 others. 2024. Effective long-context scaling of foundation models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4643–4663.

Fangzhi Xu, Hang Yan, Chang Ma, Haiteng Zhao, Qiushi Sun, Kanzhi Cheng, Junxian He, Jun Liu, and Zhiyong Wu. 2025a. Genius: A generalizable and purely unsupervised self-training framework for advanced reasoning. *Preprint*, arXiv:2504.08672.

Fengli Xu, Qianyue Hao, Zefang Zong, Jingwei Wang, Yunke Zhang, Jingyi Wang, Xiaochong Lan, Jiahui Gong, Tianjian Ouyang, Fanjin Meng, Chenyang Shao, Yuwei Yan, Qinglong Yang, Yiwen Song, Sijian Ren, Xinyuan Hu, Yu Li, Jie Feng, Chen Gao, and Yong Li. 2025b. Towards large reasoning models: A survey of reinforced reasoning with large language models. *Preprint*, arXiv:2501.09686.

Longhuan Xu, Cunjian Chen, and Feng Yin. 2026. Unsupervised layer-wise dynamic test time adaptation for llms. *Preprint*, arXiv:2602.09719.

Yijie Xu, Huizai Yao, Zhiyu Guo, Pengteng Li, Aiwei Liu, Xuming Hu, Weiyu Guo, and Hui Xiong. 2025c. You only need 4 extra tokens: Synergistic test-time adaptation for llms. *Preprint*, arXiv:2510.10223.

Dong Yan, Jian Liang, Yanbo Wang, Shuo Lu, Ran He, and Tieniu Tan. 2026. What if consensus lies? selective-complementary reinforcement learning at test time. *Preprint*, arXiv:2603.19880.

Chengyi Yang, Zhishang Xiang, Yunbo Tang, Zongpei Teng, Chengsong Huang, Fei Long, Yuhan Liu, and Jinsong Su. 2026a. Ttcs: Test-time curriculum synthesis for self-evolving. *Preprint*, arXiv:2601.22628.

Haoyan Yang, Mario Xerri, Solha Park, Huajian Zhang, Yiyang Feng, Sai Akhil Kogilathota, and Jiawei Zhou. 2026b. Self-improvement of large language models: A technical overview and future outlook. *Preprint*, arXiv:2603.25681.

Yunyao Yu, Zhengxian Wu, Zhuohong Chen, Hangrui Xu, Zirui Liao, Xiangwen Deng, Zhifang Liu, Senyuan Shi, and Haoqian Wang. 2026. Stabilizing unsupervised self-evolution of mllms via continuous softened retracing resampling. *Preprint*, arXiv:2604.03647.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason E Weston. 2024. Self-rewarding language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 57905–57923. PMLR.

Wenzhen Yuan, Shengji Tang, Weihao Lin, Jiacheng Ruan, Ganqu Cui, Bo Zhang, Tao Chen, Ting Liu, Yuzhuo Fu, Peng Ye, and Lei Bai. 2025. Wisdom of the crowd: Reinforcement learning from coevolutionary collective feedback. *Preprint*, arXiv:2508.12338.

Eric Zelikman, Georges Harik, Yijia Shao, Varuna Jayasiri, Nick Haber, and Noah D. Goodman. 2024. Quiet-STaR: Language models can teach themselves to think before speaking. In *First Conference on Language Modeling*.

Zheni Zeng, Yuxuan Chen, Shi Yu, Ruobing Wang, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, and Maosong Sun. 2025. KBAlign: Efficient self adaptation on specific textual knowledge bases. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 13519–13532, Suzhou, China. Association for Computational Linguistics.

Jin Zhang, Flood Sung, Zhilin Yang, Yang Gao, and Chongjie Zhang. 2025a. Learning to plan before answering: Self-teaching llms to learn abstract plans for problem solving. *Preprint*, arXiv:2505.00031.

Kongcheng Zhang, Qi Yao, Shunyu Liu, Yingjie Wang, Baisheng Lai, Jieping Ye, Mingli Song, and Dacheng Tao. 2025b. Consistent paths lead to truth: Self-rewarding reinforcement learning for llm reasoning. *Preprint*, arXiv:2506.08745.

Nonghai Zhang, Weitao Ma, Zhanyu Ma, Jun Xu, Jiuchong Gao, Jinghua Hao, Renqing He, and Jingwen Xu. 2026a. Silence the judge: Reinforcement learning with self-verifier via latent geometric clustering. *Preprint*, arXiv:2601.08427.

Qingyang Zhang, Haitao Wu, Changqing Zhang, Peilin Zhao, and Yatao Bian. 2025c. Right question is already half the answer: Fully unsupervised llm reasoning incentivization. *Preprint*, arXiv:2504.05812. Ongoing work. First released on April 8, 2025.

Xiaoying Zhang, Baolin Peng, Ye Tian, Jingyan Zhou, Yipeng Zhang, Haitao Mi, and Helen M. Meng. 2025d. Self-tuning: Instructing LLMs to effectively acquire new knowledge through self-teaching. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5688–5724, Vienna, Austria. Association for Computational Linguistics.

Yanzhi Zhang, Zhaoxi Zhang, Haoxiang Guan, Yilin Cheng, Yitong Duan, Chen Wang, Yue Wang, Shuxin Zheng, and Jiyan He. 2025e. No free lunch: Rethinking internal feedback for llm reasoning. *Preprint*, arXiv:2506.17219.

Zizhuo Zhang, Jianing Zhu, Xinmu Ge, Zihua Zhao, Zhanke Zhou, Xuan Li, Xiao Feng, Jiangchao Yao, and Bo Han. 2026b. Co-rewarding: Stable self-supervised RL for eliciting reasoning in large language models. In *The Fourteenth International Conference on Learning Representations*.

Andrew Zhao, Yiran Wu, Yang Yue, Tong Wu, Quentin Xu, Yang Yue, Matthieu Lin, Shenzhi Wang, Qingyun Wu, Zilong Zheng, and Gao Huang. 2025. Absolute zero: Reinforced self-play reasoning with zero data. *Preprint*, arXiv:2505.03335.

Chu Zhao, Enneng Yang, Yuting Liu, Jianzhe Zhao, and Guibing Guo. 2026a. Echo: Entropy-confidence hybrid optimization for test-time reinforcement learning. *Preprint*, arXiv:2602.02150.

Xuandong Zhao, Zhewei Kang, Aosong Feng, Sergey Levine, and Dawn Song. 2026b. Learning to reason without external rewards. In *The Fourteenth International Conference on Learning Representations*.

Yiyang Zhou, Zhiyuan Fan, Dongjie Cheng, Sihan Yang, Zhaorun Chen, Chenhang Cui, Xiyao Wang, Yun Li, Linjun Zhang, and Huaxiu Yao. 2024. Calibrated self-rewarding vision language models. *Preprint*, arXiv:2405.14622.

Yujun Zhou, Zhenwen Liang, Haolin Liu, Wenhao Yu, Kishan Panaganti, Linfeng Song, Dian Yu, Xiangliang Zhang, Haitao Mi, and Dong Yu. 2025. Evolving language models without labels: Majority drives selection, novelty promotes variation. *Preprint*, arXiv:2509.15194.

Yuxin Zuo, Kaiyan Zhang, Li Sheng, Shang Qu, Ganqu Cui, Xuekai Zhu, Haozhan Li, Yuchen Zhang, Xinwei Long, Ermo Hua, Biqing Qi, Youbang Sun, Zhiyuan Ma, Lifan Yuan, Ning Ding, and Bowen Zhou. 2025. TTRL: Test-time reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 38.

Appendix Contents

A Survey Protocol and Screening Details

A.1 Search Sources and Query Strings

We searched ACL Anthology, arXiv, Semantic Scholar, and Google Scholar for work dated January 2023–May 2026. The following frozen templates make the search reconstructable; each quoted mechanism phrase was run separately to avoid engine-specific Boolean limits.

- **ACL Anthology search:** `"<mechanism>" ("large language model" OR LLM OR multimodal)`.
- **arXiv API:** `(ti:"<mechanism>" OR abs:"<mechanism>") AND (all:"large language model" OR all:"multimodal large language model")`, with submitted-date bounds 2023-01-01 and 2026-05-31.
- **Semantic Scholar API:** `query="<mechanism> large language model", year=2023-2026, fieldsOfStudy=Computer Science`.
- **Google Scholar:** `"<mechanism>" ("large language model" OR LLM OR MLLM) -survey`, with a custom 2023–2026 year range.

The mechanism list was: *unsupervised post-training*, *self-improvement*, *self-rewarding*, *self-training*, *test-time training*, *test-time adaptation*, *test-time reinforcement learning*, *internal reward*, *self-consistency*, *majority vote*, *intrinsic reward*, and *evaluator-driven RL*. Forward and backward snowballing expanded the four strands in §2.

A.2 Inclusion and Exclusion Criteria

Inclusion: a candidate is included as strict UPT iff it satisfies all four boundary checks (B1)–(B4) of §2, has a verifiable algorithmic description in a paper, preprint, or extended technical report, and operates on foundation-scale text or multimodal models. Methods on smaller classifiers, image-only encoders predating LLM/MLLM self-improvement, or pure decoding-time tricks without an explicit update are not included as strict UPT. Tool-grounded, verifier-grounded, seed-supervised, stronger-teacher, and external-evaluator methods are kept as adjacent methods. **Exclusion**: surveys, benchmarks-only papers, hardware/system papers, and unrelated domain-specific applications without a post-training contribution.

A.3 Screening Flow

Each candidate paper was reviewed with a structured note recording: update target (parameters, adapters, memories, persistent local state, none); signal source (model samples, internal aggregates, internal evaluator, external verifier, external label); whether any seed, tool, or stronger-teacher signal entered the update loop; the internal update object against which the gradient is actually computed; and the timing regime (offline corpus, full-cohort transductive, few-sample target, streaming continual, test-time instance, within-sequence, or no-update inference-time). Boundary cases receive a source-section recheck and a recorded assignment rationale in Appendix D.

After deduplication, the inventory contains **94 method records from 91 papers**: **80 strict rows from 78 papers**, **8 adjacent rows**, and **6 prose-only boundary or antecedent records**. Multiple algorithmic variants in one paper are separate method rows.

Family / method	Backbone and evaluation	Internal update signal	Originally reported result	Interpretation and task structure
I / LangAdapt CPT (Elhady et al., 2025)	Llama 2 7B; Basque LM PPL, downstream accuracy, Copain	Next-token NLL on unlabeled Basque and English text	PPL $23.64 \rightarrow 3.35$; downstream $27.43 \rightarrow 34.14$; Copain $44.67 \rightarrow 43.43$	Large perplexity and downstream gains coexist with task-specific variation, separating language adaptation from multiple-choice evaluation.
II / TTRL (Zuo et al., 2025)	Qwen2.5-Math-7B; pass@1 on AIME 2024, MATH-500, GPQA-Diamond	Same-model answer majority as GRPO reward	AIME $12.9 \rightarrow 40.2$; MATH-500 $46.7 \rightarrow 83.4$; GPQA $29.1 \rightarrow 27.7$	Transductive adaptation yields large benchmark-specific gains; the GPQA result shows that the update does not automatically transfer across evaluation distributions.
III / Quiet-STaR (Zelikman et al., 2024)	Mistral 7B; zero-shot GSM8K and CommonsenseQA	Thoughts selected by improvement in future-token prediction	GSM8K $5.9 \rightarrow 10.9$; CommonsenseQA $36.3 \rightarrow 47.2$	Zero-shot gains arise from future-token-trained thoughts, with thought and lookahead lengths defining the operating point.
IV / CoNL (Sui and Hooi, 2026)	Qwen3-8B; pass@1 on AIME 2024 and DeepMath	Same-policy multi-agent critiques and rankings as diagnostic reward	AIME $60.0 \rightarrow 76.5$; DeepMath $70.5 \rightarrow 87.1$	Same-policy multi-agent evaluation yields strong reasoning gains without an external judge; rollout cost and open-ended transfer define the next evaluation axes.

Table 5: Representative results, one method per family. The table preserves each source paper's backbone, data, budget, and checkpoint rules to expose mechanism-specific evidence and task conditions.

]

Task-structure prior	What it contributes	Representative exposure	Boundary interpretation
Answer extractor / canonicalizer	An equivalence relation over free-form strings; easier vote counting	Math-answer consensus in TTRL, ETTRL, and related Family II methods	Supplies an equivalence relation without correctness supervision; sensitivity should be evaluated across alternative normalizers.
Finite answer alphabet	A small support for clustering or voting	Multiple-choice reasoning and classification evaluations	Supplies a compact comparison space; open-ended reformulations test whether the gain extends beyond that structure.
Code signature vs. execution	A signature constrains output form; execution returns a correctness-bearing verdict	Code-oriented self-training and adjacent executor-based methods	Signature alone is a prior; tests, interpreters, and environment rewards fail B3.
Full-cohort transduction	Visibility of the target input distribution before final prediction	TTRL, MM-UPT, and other cohort-visible test-time updates	Provides target-distribution visibility and should be reported separately from held-out generalization.
Open-ended output space	No canonical equality test; comparison requires generated targets or an evaluator	G-Zero, CoNL, dialogue, summarization, and creative generation	Exposes the regime where generated targets and internal evaluators provide the comparison signal.

Table 6: Task structure is audited independently from explicit supervision. A prior can make an internal signal more informative without specifying the correct answer; an executed verifier crosses the strict boundary.

B Representative Evidence and Task-Structure Audit

Tables 5 and 6 provide the empirical and task-structure detail referenced by the main-text synthesis. The first preserves each source paper's reported setup; the second records structural priors separately from correctness-bearing supervision.

	Primary timing regime						Update target		
Family	**Off.**	**Coh.**	**Few**	**Str.**	**Inst.**	**Seq.**	**Param.**	**Local**	**Total**
I	17	0	1	1	6	1	22	4	26
II	15	7	0	0	0	0	22	0	22
III	18	3	0	0	1	0	21	1	22
IV	9	1	0	0	0	0	10	0	10
Total	**59**	**11**	**1**	**1**	**7**	**1**	**75**	**5**	**80**

Table 7: Cross-section of the strict inventory by family, primary timing regime, and update target. Family I includes the SUDER bridge row. Off.: offline corpus; Coh.: full-cohort transductive; Few: few-sample target; Str.: streaming continual; Inst.: test-time instance; Seq.: within-sequence; Local: sample-local state. Each method record contributes once.

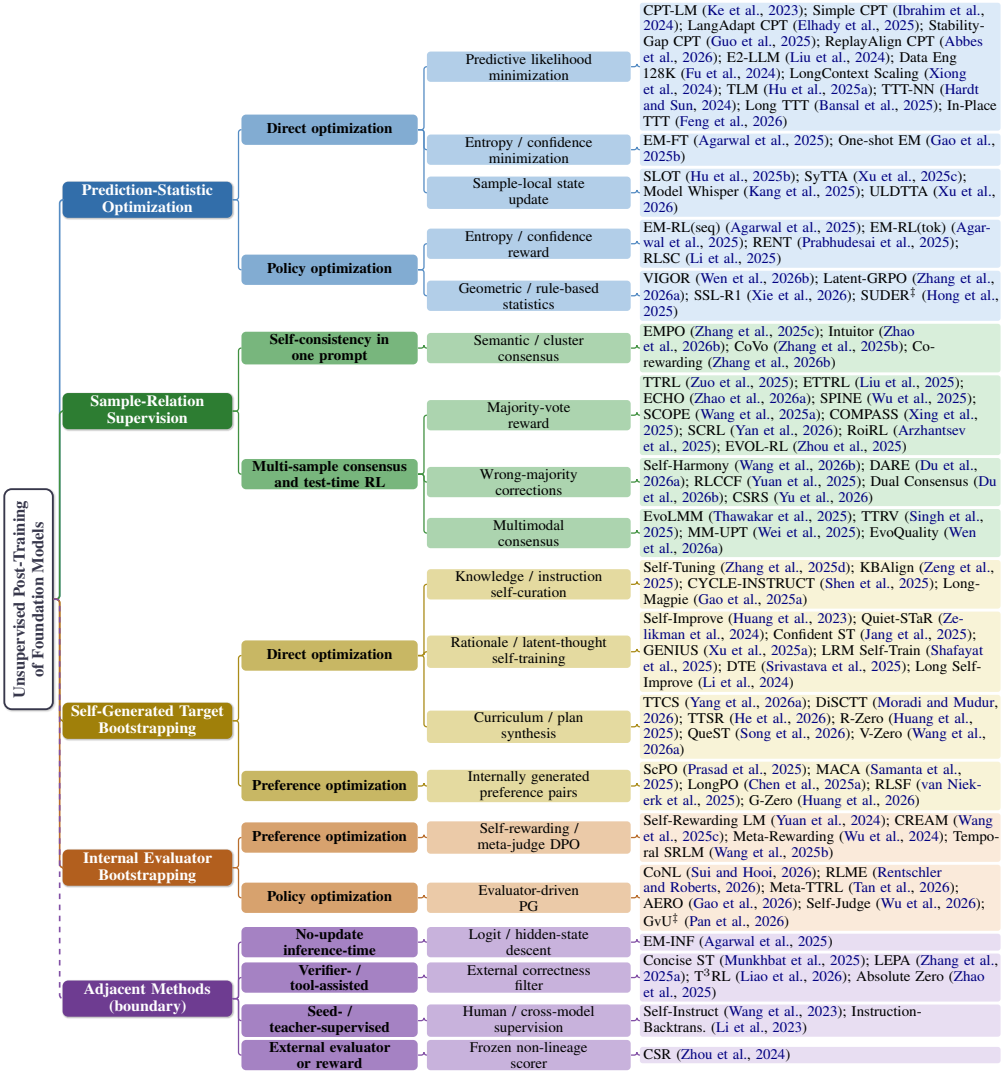

Figure 3: Full update-object taxonomy. Each of the 80 strict UPT methods appears once under its primary family and sub-class; ‡ marks the Family I/IV bridge cases (SUDER and GvU). The dashed adjacent branch places eight neighboring methods by the boundary they cross.

C Full Method Inventory

Figure 3 maps all 80 strict methods by family, subclass, and update object. Tables 1– 4 compare method attributes; the tree provides the complete hierarchy in a single view.

C.1 Family, Timing, and Update Target

The companion inventory provides machine-readable method records and per-paper rationales. Table 7 aggregates its 80 strict rows along the timing and update-target axes used in the main text.

Offline updates dominate all four families (59/80). Full-cohort transduction is concentrated in Family II (7/11), whereas all five sample-local-state methods occur in Families I and III. The cross-tabulation therefore connects the update-object taxonomy to the deployment regimes in §8.

D Boundary Decisions

Adjacent methods are organized by the boundary check they fail. The cases below cover the recurring ambiguities; stronger-teacher distillation fails (B2)–(B3) directly.

D.1 Representative Adjacent Cases

No-update inference-time optimization. EM-INF (Agarwal et al., 2025) performs inference-time entropy descent over logits or hidden states without modifying parameters, adapters, memories, or persistent local state, failing (B1). Training-free multimodal prompting provides the same boundary test: SEE&TREK (Li et al., 2026) changes spatial prompt construction without updating model parameters or persistent state. Likewise, SEPER (Dai et al., 2025) uses semantic-perplexity reduction to measure retrieval utility; an internal model statistic does not satisfy (B1) unless it drives an explicit update.

Verifier- or tool-assisted self-training. T^{3}RL (Liao et al., 2026) pairs majority-vote selection with code-interpreter verification; ABSOLUTE ZERO (Zhao et al., 2025) closes a propose-solve loop with a code executor as the truth oracle. Both fail (B3). They are important neighboring evidence for false-popular collapse fixes (§9).

Human- or seed-supervised bootstrapping. SELF-INSTRUCT (Wang et al., 2023) and instruction-backtranslation pipelines (Li et al., 2023) bootstrap from human-written seeds, failing (B3) at the seed stage. They are treated as precursors of self-generated target bootstrapping rather than strict UPT.

External reward or evaluator methods. The full CSR (Zhou et al., 2024) system includes a CLIP-derived visual-relevance term in its reward. Because the evaluator is not derived from the same model lineage, CSR fails (B4). It is routed to adjacent unless an internal-only variant is analyzed separately.

D.2 Family II vs. Family III

LRM SELF-TRAIN (Shafayat et al., 2025) uses majority vote to filter candidate solutions before SFT on the survivors. Because the gradient is computed against the kept solutions (self-generated targets), not against the consensus statistic itself, the update-object rule assigns it to Family III. The same rule places TTRL, ROIRL, and methods whose gradient is computed directly against $r = \mathbf{1}[y = \mathrm{maj}]$ in Family II.

D.3 Family III vs. Family IV

CONFIDENT ST (Jang et al., 2025) and RLSF (van Niekerk et al., 2025) use a self-confidence or self-rated score on candidate trajectories. When the score acts as a selection mask before SFT or DPO, the gradient is computed against the kept generations (Family III). When the score itself appears as the scalar reward in PG, the gradient is computed through the evaluator (Family IV). The update-object rule assigns each method according to its actual gradient path; the companion inventory records the mapping.

D.4 Strict UPT vs. Adjacent Methods

ECHO and SPINE use multi-sample consensus as a selection mask while a separate intrinsic term shapes advantages or selects gradient-receiving tokens. Their consensus reward places them in Family II, with the intrinsic term acting as a within-family modulation. EM-INF and the full CSR are routed to adjacent for the reasons in Appendix D.

Survey	Venue	Real update	No external signal	Update-object taxonomy	MLLM coverage
Liang et al. (2023) (*TTA*)	arXiv 2023	●	◎	○	○
Tao et al. (2024) (*Self-Evolution*)	arXiv 2024	◎	◎	○	◎
Liang et al. (2024) (*Internal Consistency*)	arXiv 2024	◎	◎	○	○
Xu et al. (2025b) (*Reinforced Reasoning*)	arXiv 2025	◎	○	○	○
Kumar et al. (2025) (*LLM Post-Training*)	arXiv 2025	◎	○	○	◎
Chen et al. (2025b) (*Long CoT*)	arXiv 2025	◎	○	○	●
Tie et al. (2025) (*Alignment to Reasoning*)	arXiv 2025	◎	○	○	◎
Wu (2025) (*Learning from Rewards*)	arXiv 2025	◎	◎	◎	◎
Deng et al. (2025) (*MLLM Self-Improvement*)	arXiv 2025	◎	◎	○	●
Yang et al. (2026b) (*Self-Improvement*)	arXiv 2026	◎	◎	○	◎
This survey (*Strict UPT*)	–	●	●	●	●

Table 8: Scope dimensions of this survey and the closest existing surveys, grouped by publication year. ● explicit focus; ◎ mixed coverage; ○ outside the primary scope. *Real update*: every surveyed method must induce an explicit parameter or state update. *No external signal*: the update loop uses no ground-truth answers, verifier feedback, human labels, or stronger-teacher labels. *Update-object taxonomy*: methods are organized by the internal object the gradient consumes. *MLLM coverage*: vision-language or other multimodal foundation models are included.

E Comparison with Existing Surveys

Table 8 places this survey against the closest surveys along four scope dimensions. Prior work surveys self-feedback and self-improvement (Tao et al., 2024; Liang et al., 2024; Deng et al., 2025; Yang et al., 2026b), broad LLM post-training and reward learning (Kumar et al., 2025; Tie et al., 2025; Wu, 2025), test-time adaptation (Liang et al., 2023), and reinforced reasoning (Xu et al., 2025b; Chen et al., 2025b). These scopes overlap with parts of strict UPT, but none makes all four dimensions a joint inclusion rule. Our distinctive unit of analysis is the internal object consumed by an explicit update under the no-external-signal boundary, across text and multimodal models.