WHEN TOP-K MISSES THE DECISION: TOOL-CALL DRIFT IN MULTI-TEACHER ON-POLICY DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Teacher top-K logits can preserve nearly all teacher probability mass while omitting corrections needed for a student's behavioral decisions. We study this gap in routed, two-teacher tool-use distillation: a tool teacher reinforces an entry coordinate that the response teacher usually omits even when the student overestimates it. In Qwen3.5-9B, response-teacher top-32 retains 99.99% of its mass yet contains the tool-entry token on only 0.4% of 500 response prompts. A frozen full-vocabulary audit shows that restoring this coordinate with its exact teacher logit recovers full-vocabulary descent on that coordinate. Forced-entry replay and matched training connect the omission to behavior: first-position restoration relocates calls to later tokens, whereas all-position restoration reduces full-generation over-calling from 14.2$\pm$2.1% to 3.7$\pm$0.5% while lowering required-call recall by 12.4 points. Teacher-probability-matched and output-logit-gradient-rerouted non-tool controls produce smaller shifts and do not reproduce exact restoration. We then compare support construction, loss shaping, and validation-tuned decoding bias. The comparison separates changes in the call threshold from modest gains in discrimination and exposes the access and capability costs of reducing calls. Llama-3.1-8B reproduces the directional support asymmetry under native JSON. Together, these results identify decision-critical support omission as a causal contributor in the primary Qwen setting and show why teacher-mass coverage alone cannot certify behavioral fidelity.

1 INTRODUCTION

On-policy distillation (OPD) supervises student-generated trajectories with teacher token distributions (Agarwal et al., 2024). Returning and processing full-vocabulary logits at every rollout position is expensive, so sparse-logit interfaces commonly retain only the teacher's top-K coordinates (Anshumann et al., 2025; Zhang et al., 2025). This compression is often audited by retained teacher probability mass: if top-K contains nearly all mass, the omitted tail appears unimportant.

Probability mass is teacher-centric, whereas distillation error is teacher–student relative. A token with negligible teacher probability can still require a large negative correction when the student overestimates it. If that token selects a generation mode rather than a local word, omitting it can remove supervision on an entire behavioral branch. We call this failure *decision-critical support omission*: compressed support preserves teacher mass but removes a coordinate needed to calibrate the student's decision boundary.

Routed multi-teacher tool use makes this failure observable. A tool-call teacher specializes in structured calls, a response teacher in direct answers, and metadata routes each example to its specialist (Ma et al., 2026). The tool teacher includes the tool-entry token and raises its logit. On response examples, the same coordinate has low teacher probability and is usually truncated, so the implemented objective cannot directly oppose a student's overestimate. Correct teacher routing therefore need not provide symmetric boundary supervision.

Prior work shows that teacher top-K truncation biases distillation, tail information can matter, and teacher–student support compatibility can guide position selection (Anshumann et al., 2025; Dasgupta et al., 2026; Wang et al., 2026b). We target behavior-level attribution by identifying an omitted within-position coordinate that is teacher-low-probability, student-overestimated, and behavior-switching, then connecting its omission to the learned generation boundary through matched interventions.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

This raises two questions: what student-relative mechanism produces the drift, and where in the learning and deployment stack can it be addressed? We study support-level coordinate construction, loss-level signal shaping, and decoding-level boundary adjustment.

We trace the first question through five links. A frozen full-vocabulary audit measures entry-coordinate support and descent; paired forced-entry replay tests how a local mode error changes the continuation; matched first- and all-position restoration identifies the role of temporal support coverage; a teacher-probability-matched non-tool coordinate tests arbitrary support expansion; and a synthetic output-logit-gradient-reroute control tests whether routing an equal-magnitude output-logit gradient to another coordinate reproduces the effect. The resulting chain shows that first-position restoration moves calls later, whereas all-position restoration changes complete generations and exposes a corresponding loss of required calls. We then compare support, loss, and decoding responses under shared behavioral metrics and access constraints.

Our contributions are:

- We identify decision-critical support omission: teacher-only top-K can preserve nearly all teacher mass while removing a student-relative coordinate needed to correct a behavioral branch.
- We causally implicate this omission in call-heavy drift in the primary Qwen setting through frozen coordinate gradients, forced-entry replay, and matched temporal-scope restoration, with the probability-matched placebo and gradient-reroute control narrowing alternative explanations.
- We compare support construction, loss shaping, and decoding adjustment, exposing their distinct operating points, correction scopes, and system requirements.

2 Setting and Behavioral Symptom

Routed on-policy distillation. Let x be a dialogue context and $y = (y_1, \ldots, y_T)$ a student-generated continuation. A metadata tag $m \in \{\text{tool}, \text{response}\}$ selects the specialized teacher for each example. The tag is extracted before text encoding, is not rendered into the student's input, and is removed before the student forward pass. Evaluation uses ordinary generation without the tag or a route oracle.

At position i, let p_t and p_s be the teacher and student next-token distributions and let $S_i^{(K)}$ be the teacher's top-K set. With $\mathcal{R}_S[p]$ denoting normalization of p on S, the implemented token loss is

$$d_i^{(K)} = \mathrm{JSD}\Big(\mathcal{R}_{S_i^{(K)}}[p_t] \,\|\, \mathcal{R}_{S_i^{(K)}}[p_s]\Big). \tag{1}$$

Teacher and student logits use temperature $\tau = 0.9$ before restriction, and the teacher API returns $K = 32$. For every omitted token $v \notin S_i^{(K)}$, $\partial d_i^{(K)}/\partial z_{i,v} = 0$: this support-conditioned objective is locally independent of the omitted student logit. With probability $\lambda = 0.8$, training uses a student rollout; otherwise it uses the dataset continuation. Dataset-source trajectories also receive a fixed SFT format anchor with weight $\alpha = 0.3$.

Decision-critical support. An omitted coordinate is decision-critical when the teacher assigns it little probability, the student assigns it enough probability to change a discrete behavior, and the full-vocabulary objective would correct it. In the Qwen protocol, `<tool_call>` enters a structured call. The tool teacher normally includes and reinforces this coordinate, whereas the response teacher usually omits it. Entry also has *trajectory leverage*: later tokens follow the selected tool or response mode.

Empirical setting. The primary experiments use a Qwen3.5-9B student and two specialized 9B teachers trained on APIGen-MT-derived data (Prabhakar et al., 2025). A conversation-disjoint balanced test contains 4,000 call/respond decision points. On this set, over-calling is the fraction of should-respond outputs parsed as calls; call and response recall are the correctly parsed fractions within their respective target classes, and decision accuracy averages the two recalls. With thinking disabled, E1 records whether the first generated assistant token enters tool mode, whereas complete-generation metrics parse the full output. BFCL (Patil et al., 2025), When2Call (Ross et al., 2025),

and a fixed-harness BFCL multi-turn diagnostic test single-turn transfer and interaction. A separate Llama-3.1-8B-Instruct study (Meta AI, 2024) changes model family, teachers, tokenizer, and tool serialization.

Model	Dec. Acc.	Over-call	Call Rec.	Resp. Rec.
Base	80.7	7.2	68.5	92.8
Mixed SFT	85.3	4.9	75.5	95.1
Vanilla GKD	88.6±0.3	14.2±2.1	91.5±1.7	85.8±2.1

Table 1: Behavioral symptom on APIGen-MT (percent). Vanilla GKD improves required-call recall but shifts the call/respond boundary toward excessive tool use. GKD values are mean±sample standard deviation over three seeds; Base and Mixed SFT are unmatched reference runs.

Table 1 motivates the mechanism study. Relative to Mixed SFT, vanilla GKD reaches much higher required-call recall, but its over-calling rises from 4.9% to 14.2%. The analyses below ask whether omitted response-side decision support is a causal contributor to this call-heavy shift; matched comparisons use paired GKD runs with common data, teachers, initialization, training budget, and evaluation inputs.

3 From Omitted Support to Behavior

3.1 Teacher mass does not preserve a student-relative correction

At the first supervised response position, we recompute full-vocabulary teacher and student logits for 500 unique validation prompts. Teacher-only top-32 retains 0.999900 mean raw response-teacher mass yet includes <tool_call> for only 0.4% of prompts. Coverage reaches 4.6%, 22.6%, and 52.2% at $K = 64, 128, 256$. The median strict competition rank is 236.5; the 90th and 99th percentiles are 614.1 and 1086.0, and ranks span 25–1243. Mean teacher $P(\texttt{<tool_call>})$ is only 6.14×10^{-8}. The complementary tool route exposes the same coordinate: on 500 unique validation prompts, the tool teacher ranks <tool_call> first and includes it in top-32 every time. Across 1,500 matched teacher–student pairs, its top-32 entry descent is positive in every pair (pooled $+0.001167$). Thus the tool route directly reinforces a coordinate whose response-side correction is almost always omitted; full routed results appear in Table 12.

Pairing the 500 prompts with three matched students yields 1,500 response-route audits. Table 2 reports entry-logit descent $-\partial d/\partial z_{\text{tool}}$, where a negative value lowers the tool-entry logit. Top-32 supplies essentially no direct correction. Inserting <tool_call> with its exact response-teacher logit recovers the full-vocabulary descent on this coordinate in every seed.

Student	K=32 cov.	K=32 descent	K=32 + entry	Full vocab.
Seed 42	0.4%	−0.000001	−0.037799	−0.037749
Seed 44	0.4%	−0.000003	−0.044787	−0.044739
Seed 60	0.4%	−0.000003	−0.039682	−0.039636
Pooled	0.4%	−0.000002	−0.040756	−0.040708

Table 2: Frozen response-side entry-coordinate audit. Adding the exact omitted coordinate to teacher top-32 recovers full-vocabulary descent on that coordinate, not the full gradient vector.

The correction is student-relative. Among 1,326 response-mode entries, mean student $P(\texttt{<tool_call>})$ is 0.093 and inserted descent is -0.0290; among 174 erroneous tool-mode entries, these values rise to 0.702 and -0.1303. Increasing teacher-only K recovers some coordinates but preserves the teacher-centric ordering that hides where the student is most wrong.

3.2 Wrong mode entry amplifies along the trajectory

Aggregate training statistics do not identify this directional mechanism. Across 31 paired diagnostic steps per method, confidence intervals for tool/response token exposure, per-token JSD, and a squared-

divergence proxy all contain parity. We therefore replay decisions rather than infer behavior from sequence averages.

For each Qwen seed, we replay 500 validation tool-tagged and 500 validation response-tagged prompts under their routed teachers. Tool-tagged rollouts that incorrectly answer directly have 55.45 times the per-token JSD of aligned tool rollouts (95% CI [41.74, 76.44]); response-to-tool mismatch is 2.03 times its aligned baseline [1.71, 2.39]. To control prompt difficulty, we additionally construct 200 paired validation prompts per teacher tag for seed 42: one branch forces tool entry and the other masks it to sample a response entry. These pairs yield ratios of 26.73 [14.87, 67.91] and 2.67 [2.29, 3.14], respectively. The pre-entry distribution is shared; divergence opens after the mode choice and persists over later tokens. Thus a local entry error has trajectory-level leverage, although replay alone does not identify why the error was learned.

3.3 Restoring the coordinate changes the learned boundary

We train matched diagnostic interventions over seeds 42, 44, and 60. On response-tagged examples, exact restoration appends `<tool_call>` with its response-teacher logit whenever top-32 omits it; tool-tagged examples are unchanged. First-position restoration acts only at the first effective supervised response position, whereas all-position restoration applies the same rule at every supervised response position. At all-position scope, a placebo inserts the tail token closest in teacher probability to `<tool_call>`; a synthetic output-logit-gradient-reroute control (gradient reroute) keeps the exact-restoration forward value but reroutes its selected backward coefficient to that matched non-tool row.

Training	E1 resp. err.	E1 call rec.	Over-call	Call rec.
Vanilla	14.47±1.53	91.72±1.22	14.20±2.08	91.45±1.68
First-position	3.40±0.26	80.57±0.68	13.48±1.27	91.47±0.50
All-position	3.63±0.24	81.30±0.51	3.73±0.51	79.07±1.73
Non-tool placebo	13.37±1.10	91.30±1.00	13.25±1.96	90.37±2.31
Gradient reroute	9.25±3.43	85.83±2.92	11.65±2.88	86.93±3.31

Training	Calls/turn	Dialogue exact
Vanilla	1.515±0.122	7.04±0.31
First-position	1.237±0.154	5.83±1.01
All-position	0.975±0.045	3.38±0.50
Non-tool placebo	1.346±0.021	4.92±0.40
Gradient reroute	1.464±0.481	5.21±0.44

Table 3: Matched diagnostic controls over three seeds (mean±sample std). Upper metrics are percentages; the lower block is the BFCL diagnostic. Gradient rerouting matches the forward loss and selected output-logit gradient magnitude, not full parameter gradients, and does not reproduce exact restoration.

First-position restoration changes the local boundary but mostly postpones calls. Across 6,000 should-respond outputs, first-position calls fall from 13.42% to 3.07%, while delayed calls rise from 0.78% to 10.42%; complete over-calling therefore remains near vanilla. All-position restoration has a similar E1 operating point but reduces delayed calls to 0.65% and full-generation over-calling to 3.73%. Its paired over-calling reductions relative to first-position restoration are 10.10, 8.35, and 10.80 points across seeds. Temporal support coverage, rather than a larger first-token shift, separates the two interventions.

All-position restoration serves as a diagnostic intervention: it tests whether broader temporal support coverage explains the observed call relocation and exposes the resulting restraint–capability trade-off. APIGen call recall falls by 12.38 points and calls per turn from 1.515 to 0.975; dialogue exact success falls from 7.04% to 3.38%. The non-tool placebo controls position, support count, and teacher-probability scale. Its paired placebo-minus-vanilla changes are -0.95 ± 0.78 over-calling points and -1.10 ± 0.95 E1 response-error points, versus -10.47 ± 2.05 and -10.83 ± 1.66 for all-position exact restoration. This rules out arbitrary matched-scale support expansion, while leaving student probability and effective gradient strength unmatched.

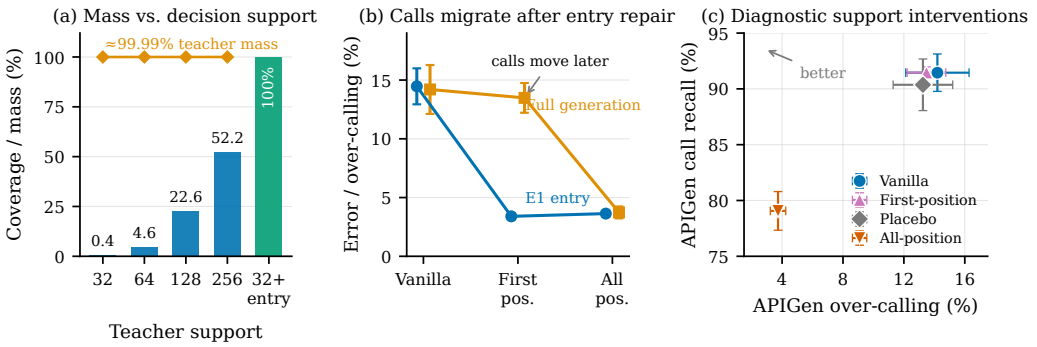

Figure 1: Evidence chain under teacher-top-32. (a) Near-complete response-teacher mass coexists with low decision-token coverage. (b) First-position restoration relocates calls, while all-position restoration closes the later-position path. (c) First-position restoration and the matched placebo remain near vanilla in full generation, whereas all-position restoration moves both over-calling and required-call recall. Error bars are sample standard deviations over seeds 42, 44, and 60.

The gradient-reroute control matches the frozen forward JSD and selected output-logit-gradient coefficient and norm, yet its E1 response error/APIGen over-calling remain 9.25/11.65%, versus 3.63/3.73% under exact restoration. Per-seed over-calling gaps are $+8.25$, $+4.90$, and $+10.60$ points. Routing the same selected output-logit-gradient magnitude to another coordinate does not reproduce exact restoration. Appendix B.7.1 reports the unmatched output row, full gradients, and collateral effects.

A separate direct penalty on tool-entry probability at all supervised response positions, tested at fixed weight 0.15, reduces over-calling from 14.20% to 12.93% while lowering dialogue exact from 7.04% to 4.96% across three matched seeds. This single setting provides a supplementary suppression control (Appendix B.7).

3.4 What the intervention identifies

Together, the frozen audit, replay, and temporal-scope comparison show that reinstating the omitted entry coordinate causally shifts call propensity in the primary Qwen setting. The non-tool controls narrow arbitrary-support and off-coordinate equal-magnitude-gradient explanations. Because restoration targets one coordinate, these results identify support omission as a contributor rather than a complete account of the drift or of full-vocabulary OPD.

4 Cross-Layer Interventions and Operating Points

The diagnosis suggests interventions at three levels of the learning and deployment stack. Table 4 makes their correction scope and access requirements explicit. Exact restoration and its matched controls remain diagnostic interventions rather than deployment candidates.

The support strategy uses a familiar teacher/student top-K union as the within-position support and queries teacher logits for student-selected coordinates. The loss strategies retain teacher top-K: Hard Clip caps token divergence, Global Reweight applies batch-relative weights (Hao et al., 2026a), and localized Soft Clamp rescales only above-threshold gradients while retaining nonzero updates. Unlike support construction, loss shaping does not restore omitted coordinates; it instead tests whether reducing the concentration of extreme retained-support signals can moderate the aggregate behavioral shift. The decoding strategy applies a validation-tuned non-positive bias to the first tool-entry logit. Detailed objectives and hyperparameters appear in the appendix.

Layer	Evaluated strategy	Correction scope	Retrain	Access requirement
Support	Student-aware support union	Distilled coordinates	Yes	Teacher probes for student-selected tokens
Loss	Soft Clamp (+ loss baselines)	Retained-signal strength	Yes	Original teacher top-K only
Decoding	Validation-tuned entry bias	Deployed boundary	No	Validation data; entry-logit control

Table 4: Intervention layers and system requirements. Among the evaluated intervention layers, only support-level construction places student-selected omitted coordinates into the training support; the other layers operate on retained signals or decoding.

Method	Dec. Acc.	Over-call	Call Rec.	Resp. Rec.
Vanilla GKD	88.6±0.3	14.2±2.1	91.5±1.7	85.8±2.1
Hard Clip	88.7±1.0	11.9±0.9	89.2±3.0	88.2±0.9
Global Reweight	88.6±1.3	9.7±0.8	86.9±3.3	90.3±0.8
Soft Clamp	88.7±0.7	9.0±0.2	86.5±1.4	91.0±0.2
Support union	89.8±0.8	7.4±0.6	87.0±2.0	92.6±0.6

Table 5: APIGen-MT decision results (percent). Loss shaping and support construction occupy different call/respond operating points; Base and Mixed SFT references appear in Table 1.

4.1 Primary Qwen operating points

All GKD variants share the teacher pair, routed data, initialization, one-epoch budget, $K = 32$, $\lambda = 0.8$, $\tau = 0.9$, and $\alpha = 0.3$ format anchor. Table 5 reports mean±sample standard deviation over three seeds.

In the frozen Qwen response-side audit, the union restores entry coverage to 100% over 1,500 prompt–student pairs. Entry descent is -0.040748, close to the full-vocabulary value of -0.040708 on that coordinate (Table 12). During training, the union averages 38.88 ± 0.08 coordinates: returned support grows by 21.5% over top-32, and observed steady-state step time rises by 16.0%. Thus the shared epoch budget includes an additional compute cost for this support strategy.

Vanilla has 14.2% over-calling and 91.5% call recall. Soft Clamp moves these metrics to 9.0% and 86.5%, while support union moves them to 7.4% and 87.0%. Diagnostic all-position restoration reaches 3.7% over-calling and 79.1% call recall, showing its more aggressive restraint–capability shift. Deployment preference among these interventions therefore depends on error costs and system access.

4.2 Cross-layer boundary effects and deployment trade-offs

The E1 first-token protocol separates ranking from threshold movement. AUC uses first-token call-entry probability with should-call as the positive class. First- and all-position restoration have AUCs 0.9712 and 0.9725, and Soft Clamp reaches 0.9710, all near vanilla's 0.9692; each lowers both response-entry error and call-entry recall. For each seed, entry bias is selected on validation to match Soft Clamp's response-entry error or call recall, then frozen on test. It reproduces most of Soft Clamp's mean entry suppression without retraining.

The support-level union reaches $8.72 \pm 0.74\%$ response-entry error and $90.55 \pm 1.00\%$ call-entry recall while increasing AUC to 0.9760 ± 0.0006, evidence for modest separation change as well as threshold movement. Figure 2 compares all three layers.

For the cost-sensitive comparison, let $\rho = C_{\text{over}}/C_{\text{miss}}$ be the cost of an unnecessary call relative to a missed required call. With class-conditional error rates e_{over} and e_{miss}, the class-balanced normalized risk is $R_\rho = (\rho e_{\text{over}} + e_{\text{miss}})/(\rho + 1)$. In E1, cost-aware decoding biases minimize validation risk separately for each seed and ρ, then remain fixed on test. On APIGen, vanilla has the lowest mean risk at $\rho \in \{0.25, 0.5\}$ and support union at $\rho \in \{1, 2, 4\}$; on E1, support union is lowest through $\rho = 2$, while validation-tuned decoding bias is lowest at $\rho = 4$ (Appendix C.2).

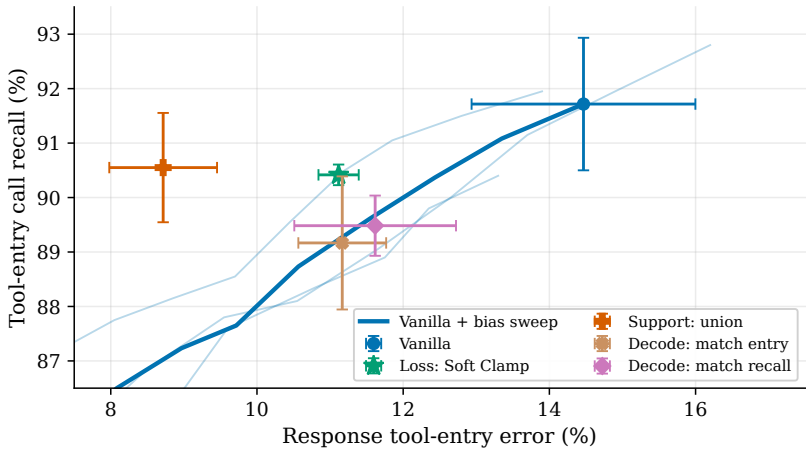

Figure 2: E1 tool-entry operating points. Light curves are per-seed vanilla inference-bias sweeps and the dark curve is their mean; markers show fixed trained methods and validation-matched decoding points. Upper-left is better. Support union is consistent with modest separation change, whereas loss shaping and tuned decoding points mainly move the threshold.

4.3 Interaction trade-offs

Our fixed BFCL diagnostic contains 800 multi-turn tasks and 3,136 user turns, returns simulated tool observations, and permits five call steps per turn. Calls/turn counts emitted calls; Loop@3 flags three consecutive call steps. Dialogue exact matches execution and state on required turns and forbids calls on empty-ground-truth turns; it excludes final-answer language quality.

Table 6 pairs call and loop reductions with their capability costs. The stronger call reduction under support union coincides with a larger loss of required-call coverage and dialogue exact. Single-turn transfer metrics vary less (Appendix C.3).

Table 6: Interaction trade-offs under the main vLLM BFCL harness. Values are mean±sample std over three seeds, in percent except calls/turn. Required call denotes call coverage on required turns; dialogue exact checks execution and state, excluding final-answer language quality.

Method	Calls/turn	Loop@3	Required call	Dialogue exact
Vanilla GKD	1.515±0.122	15.1±3.0	92.90±0.33	7.04±0.31
Soft Clamp	1.289±0.023	10.3±0.4	88.50±2.01	6.17±0.38
Support union	1.118±0.057	8.0±0.9	81.22±2.38	4.12±0.33

Under a separate strict local-decoder protocol, entry-matched decoding bias moves calls/turn from 1.734 to 1.492 and Loop@3 from 20.5% to 15.7% without retraining. These absolute values are protocol-specific; the decoding point is compared only with the same-decoder vanilla baseline (Appendix C.2). Across layers, fewer unnecessary calls do not imply better endpoints, and the appropriate operating point depends on error costs and system access.

4.4 Cross-family directional replication

Llama-3.1-8B changes the model family, tokenizer, teachers, and tool serialization. Matched vanilla GKD has 28.80% over-calling and 97.27% call recall; Soft Clamp moves these to 22.35% and 95.05%, and support union to 11.08% and 91.23%. Both trade recall for restraint. Separately trained Mixed SFT is an unmatched reference.

The native JSON-entry audit reproduces the directional asymmetry: response top-32 retains 99.977% mass but gives zero entry coverage through $K = 256$, whereas the tool teacher ranks the coordinate first on all 500 prompts and reinforces it. Across 1,500 response-teacher–student pairs, top-32

union restores 97.93% entry coverage and descent -0.04923, near full-vocabulary -0.04918. Qwen supplies the matched causal chain; Llama shows the direction is not XML-specific.

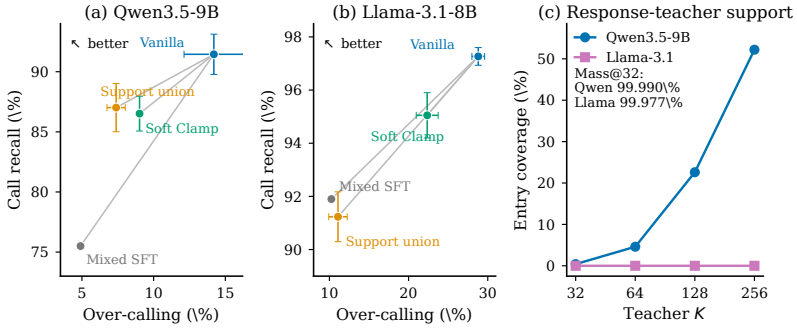

Figure 3: Cross-family behavior and response-teacher support. Panels (a)–(b) show APIGen operating points; upper-left is better for the displayed call/respond trade-off. Panel (c) shows decision-token coverage as teacher K grows, despite 99.990% Qwen and 99.977% Llama mass at top-32. Error bars are sample standard deviations over three GKD seeds.

5 Related Work

Sparse and tail-aware distillation. Sparse Logit Sampling estimates truncated gradients with sampled tail logits (Anshumann et al., 2025); Tail-Aware Distillation models aggregate tail contribution (Dasgupta et al., 2026), while top-K-censored API work characterizes unresolved distributions (Nie et al., 2026). We instead identify an omitted coordinate by its student-relative behavioral role.

Selective OPD and multiple teachers. Selective methods prioritize positions by uncertainty or teacher–student disagreement (Xu et al., 2026; Wang et al., 2026b); Teachability-Aware OPD also uses support union for position compatibility, so the set union is not new. Multi-teacher work studies routing or conflicting supervision (Song & Zheng, 2026; Ma et al., 2026; Wang et al., 2026a; Sumit et al., 2026; Chen et al., 2026). We hold routing and positions fixed and identify a behavior-correcting within-position coordinate.

Behavioral contribution. Tool overuse and repeated calls also occur in reinforcement learning and agentic search (Qian et al., 2025; Zeng et al., 2026; Hao et al., 2026b). We behaviorally validate a routed-OPD failure where one route exposes a mode-entry coordinate while the other truncates its direct correction.

6 Limitations

The matched chain is established for Qwen teacher-top-32; Llama replicates only the routed asymmetry. We restore one coordinate and do not train full-vocabulary OPD, so the experiments implicate support omission as a contributor rather than a complete account of the drift.

The natural placebo matches position, support count, and teacher probability, but not student probability or gradient strength. The gradient-reroute control matches the forward loss and selected output-logit gradient magnitude, but not the output row or full parameter gradients, and degrades BFCL/When2Call relative to vanilla. Support union needs teacher probes, loss shaping cannot restore omissions, and entry bias needs decoder control. The local BFCL study is not an official final-answer leaderboard evaluation.

7 CONCLUSION

Teacher-only top-K can preserve nearly all teacher mass while omitting a behavior-correcting coordinate. In routed tool-use OPD, frozen audits, replay, and temporal restoration causally implicate this omission in Qwen's call-heavy boundary; matched non-tool controls narrow arbitrary-support and off-coordinate equal-magnitude-gradient explanations. The resulting support-, loss-, and decoding-level interventions expose distinct trade-offs in correction scope, system access, and required-call retention. Llama reproduces the routed asymmetry under native JSON. Teacher-mass coverage alone therefore cannot certify behavioral fidelity.

AI USE STATEMENT

Generative AI tools assisted with literature discovery, code and analysis support, and language and structural editing. All cited work, experimental outputs, numerical claims, and AI-assisted text were reviewed by the authors; code and analysis changes were checked through tests and reproducible runs. The authors take responsibility for the final content, claims, and artifacts.

ETHICS STATEMENT

This study analyzes calibration failures in tool-using language models. Both unnecessary calls and missed required calls can create downstream cost or reliability risks. We therefore report both sides of the operating-point trade-off and do not recommend indiscriminate call suppression. The work uses existing model and benchmark data and does not involve a new human-subject study.

REPRODUCIBILITY STATEMENT

Appendices A–H document the data split, training configuration, objectives, metrics, per-seed comparisons, frozen-support audits, and multi-turn harness. The anonymous supplementary package will include implementation and aggregate artifacts needed to reproduce the reported analyses.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. On-Policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2306.13649.

Anshumann, Mohd Abbas Zaidi, Akhil Kedia, Jinwoo Ahn, Taehwak Kwon, Kangwook Lee, Haejun Lee, and Joohyung Lee. Sparse logit sampling: Accelerating knowledge distillation in LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025. URL https://arxiv.org/abs/2503.16870.

Tianlei Chen, Jiao Ou, Ziyuan Liu, Ruiming Tang, Jian Liang, and Han Li. Counteraction-aware multi-teacher On-Policy distillation for general capability recovery with domain preservation. *arXiv preprint arXiv:2605.27115*, 2026. URL https://arxiv.org/abs/2605.27115.

Sayantan Dasgupta, Trevor Cohn, and Timothy Baldwin. Don't ignore the tail: Decoupling top-K probabilities for efficient language model distillation. In *International Conference on Machine Learning (ICML)*, 2026. URL https://arxiv.org/abs/2602.20816.

Dongze Hao, Zhiwei Jin, Chen Chen, and Haonan Lu. Stabilizing On-Policy distillation for MLLM reasoning with global normalization. *arXiv preprint arXiv:2606.09091*, 2026a. URL https://arxiv.org/abs/2606.09091.

Yupu Hao, Zhuoran Jin, Huanxuan Liao, Kang Liu, and Jun Zhao. Why multi-step tool-use reinforcement learning collapses and how supervisory signals fix it. *arXiv preprint arXiv:2606.26027*, 2026b. URL https://arxiv.org/abs/2606.26027.

Wenhan Ma, Jianyu Wei, Liang Zhao, Hailin Zhang, Bangjun Xiao, Lei Li, Qibin Yang, Bofei Gao, Yudong Wang, Rang Li, Jinhao Dong, Zhifang Sui, and Fuli Luo. MOPD: Multi-teacher On-Policy

distillation for capability integration in LLM post-training. *arXiv preprint arXiv:2606.30406*, 2026. URL https://arxiv.org/abs/2606.30406.

Meta AI. Llama 3.1-8B-Instruct model card. Hugging Face model card, 2024. URL https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct.

Wenhua Nie, ZiCheng Zhu, Jianan Wu, Binhan Luo, Haoran Zheng, and Jyh-Shing Roger Jang. Identified-set geometry of distributional model extraction under top-K censored API access. *arXiv preprint arXiv:2605.10407*, 2026. URL https://arxiv.org/abs/2605.10407.

Shishir G. Patil, Huanzhi Mao, Charlie Cheng-Jie Ji, Fanjia Yan, Vishnu Suresh, Ion Stoica, and Joseph E. Gonzalez. The berkeley function calling leaderboard (BFCL): From tool use to agentic evaluation of large language models. In *International Conference on Machine Learning (ICML)*, 2025. URL https://gorilla.cs.berkeley.edu/leaderboard.html.

Akshara Prabhakar, Zuxin Liu, Ming Zhu, Jianguo Zhang, Tulika Awalgaonkar, Shiyu Wang, Zhiwei Liu, Haolin Chen, Thai Hoang, Juan Carlos Niebles, Shelby Heinecke, Weiran Yao, Huan Wang, Silvio Savarese, and Caiming Xiong. APIGen-MT: Agentic pipeline for multi-turn data generation via simulated agent-human interplay. *arXiv preprint arXiv:2504.03601*, 2025. URL https://arxiv.org/abs/2504.03601.

Cheng Qian, Emre Can Acikgoz, Hongru Wang, Xiusi Chen, Avirup Sil, Dilek Hakkani-Tür, Gokhan Tur, and Heng Ji. SMART: Self-aware agent for tool overuse mitigation. In *Findings of the Association for Computational Linguistics: ACL 2025*, 2025. URL https://aclanthology.org/2025.findings-acl.239/.

Hayley Ross, Ameya Sunil Mahabaleshwarkar, and Yoshi Suhara. When2Call: When (not) to call tools. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2025. URL https://github.com/NVIDIA/When2Call.

Mingyang Song and Mao Zheng. A survey of On-Policy distillation for large language models. *arXiv preprint arXiv:2604.00626*, 2026. URL https://arxiv.org/abs/2604.00626.

Dipto Sumit, Ankan Kumar Roy, Sadia Khair Rodela, Atia Haque Asha, Mourchona Afrin, Niloy Farhan, and Farig Yousuf Sadeque. Reliability gated multi-teacher distillation for low resource abstractive summarization. *arXiv preprint arXiv:2604.03192*, 2026. URL https://arxiv.org/abs/2604.03192.

Jianze Wang, Ying Liu, Jinlong Chen, Xuchun Hu, Qilong Zhang, Yu Cao, Jun Wang, Hua Yang, Yong Xie, and Qianglong Chen. MAD-OPD: Breaking the ceiling in On-Policy distillation via multi-agent debate. *arXiv preprint arXiv:2605.01347*, 2026a. URL https://arxiv.org/abs/2605.01347.

Yuanyi Wang, Su Lu, Yanggan Gu, Pengkai Wang, Yifan Yang, Zhaoyi Yan, Congkai Xie, Jianmin Wu, and Hongxia Yang. Not all disagreement is learnable: Token teachability in On-Policy distillation. *arXiv preprint arXiv:2605.26844*, 2026b. URL https://arxiv.org/abs/2605.26844.

Yuanda Xu, Hejian Sang, Zhengze Zhou, Ran He, Zhipeng Wang, and Alborz Geramifard. TIP: Token importance in On-Policy distillation. *arXiv preprint arXiv:2604.14084*, 2026. URL https://arxiv.org/abs/2604.14084.

Yirong Zeng, Shen You, Yufei Liu, Qunyao Du, Xiao Ding, Yutai Hou, Yuxian Wang, Wu Ning, Haonan Song, Dandan Tu, Bibo Cai, and Ting Liu. The tool-overuse illusion: Why does LLM prefer external tools over internal knowledge? *arXiv preprint arXiv:2604.19749*, 2026. URL https://arxiv.org/abs/2604.19749.

Xinlu Zhang, Na Yan, Yang Su, Yansha Deng, and Toktam Mahmoodi. Communication-aware knowledge distillation for federated LLM fine-tuning over wireless networks. In *IEEE Global Communications Conference (GLOBECOM)*, 2025. doi: 10.1109/GLOBECOM59602.2025.11432269. URL https://arxiv.org/abs/2509.01750.

A Additional Experimental Details

A.1 Training data and supervision

The main GKD experiments use two filtered training splits derived from APIGen-MT: 15,419 tool-call examples and 15,245 response examples. We first split 5,000 APIGen-MT conversations at the conversation level with split seed 42: 3,500 conversations for training, 500 for validation, and 1,000 for test. The conversation IDs are disjoint across splits. All GKD variants use the same post-filter files. The validation split is approximately balanced, with 2,188 tool-call examples and 2,172 response examples. The APIGen-MT-derived decision test is extracted only from the held-out test conversations, with four balanced decision buckets of 1,000 examples each, for 4,000 total test decision points. Each GKD training example is routed to one teacher by its teacher tag. The tool-call teacher supervises examples whose final target is a structured tool call, and the response teacher supervises examples whose final target is a natural-language answer.

Only the final assistant turn is supervised; earlier turns provide context. Because our ms-swift/Megatron backend consumes this target as text, we pre-render tool calls in the model template's XML-style form using tags such as `<tool_call>`, `<function=...>`, and `<parameter=...>`. This backend-specific step prevents unrendered objects from corrupting the learned schema.

A.2 Tool-call target rendering

For structured tool-call examples, the final supervised target is rendered as text before training in our backend. A representative target is:

```
<tool_call>
<function=get_flight_cost>
<parameter=travel_to>
LAX
</parameter>
</function>
</tool_call>
```

This appendix example is illustrative; the experiments use the rendered targets from the APIGen-MT-derived training split. Other frameworks may serialize the same structured call differently.

A.3 Main training configuration

Table 7 lists the configuration shared by the four loss-comparison variants and the support-construction interventions. All runs use the same data, teachers, training length, and batch settings within each seed. The loss-comparison variants differ only in their modifier; exact restoration and its matched controls append one matched coordinate, while the support union probes student-selected coordinates as described in Sections 3.3 and 4.

All GKD variants share the supervised anchor for structured-output stability. Student rollouts are sampled with probability $\lambda = 0.8$; the SFT anchor applies only to dataset-source trajectories, as defined in Section 2.

All primary Qwen GKD results use the final one-epoch checkpoint 319; intermediate 50-step saves are used only for trajectory analysis, never model selection. Llama likewise reports final one-epoch checkpoint 306. Neither validation nor test metrics select checkpoints.

Teacher and SFT reference training. The 9B tool-call teacher, response teacher, and Mixed SFT reference are full fine-tunes from the Qwen3.5-9B checkpoint. They use the same ms-swift SFT backend, bfloat16 precision, DeepSpeed ZeRO-3, Qwen3.5 template, maximum length 10,000, learning rate 5×10^{-6}, cosine schedule with 0.1 warmup fraction, micro batch size 4, gradient accumulation 2, and two epochs. The tool-call teacher is trained only on the tool-call SFT split, the response teacher only on the response SFT split, and Mixed SFT on the mixed SFT split. The 4B Mixed SFT reference uses the same recipe with Qwen3.5-4B. We use the fixed final epoch-2 checkpoints throughout the evaluation.

Item	Value
Initialization checkpoint	Qwen/Qwen3.5-9B
Training backend	ms-swift Megatron GKD backend
Precision	bfloat16
Epochs	1
Micro batch size	4
Global batch size	96
Maximum input length	14,000 tokens
Maximum completion length	512 tokens
Learning rate	5×10^{-6}
Warmup fraction	0.1
Tensor parallel size	8
GKD rollout mixture λ	0.8
GKD β	0.5
Distillation / rollout temperature	0.9
Rollout top-p / top-k	1.0 / disabled
Sequence KD	false
Supervised format anchor	`sft_alpha`=0.3
Teacher logit top-K	32
Loss scale	`last_response+ignore_empty_think`
Loss-comparison checkpoint saves	Every 50 steps plus final step 319
Reported primary Qwen checkpoint	Final one-epoch step 319

Table 7: Shared GKD training configuration.

Loss-modifier hyperparameters. We perform no joint validation search. Before multi-seed test evaluation, we fix Soft Clamp at its design default $k = 3$, Hard Clip at $c = 0.5$, and Global Reweight at $(\alpha, z_{\max}, w_{\min}, w_{\max}) = (0.3, 3.0, 0.25, 2.0)$. These are fixed design settings; the latter two baselines use single contrasts, not grids. Only the decoding biases in Appendix C.2 use validation selection.

B Metric Definitions

B.1 Evaluation metrics

For the APIGen-MT-derived decision set, *decision accuracy* is the fraction of examples for which the model makes the correct tool-use mode-entry decision. We remove thinking blocks and apply a family-specific parser before normalizing detected calls to one canonical representation. The Qwen parser detects its XML `<tool_call>` block; the Llama parser detects a native JSON object with `name` and `parameters` fields. *Over-calling* is the fraction of should-respond examples for which the corresponding parser detects a call. *Call recall* is detection recall on should-call examples, and *respond recall* is recall on should-respond examples. These decision metrics do not require schema-valid or executable arguments; schema validity is instead tested by the separate BFCL call-quality evaluation. The APIGen-MT-derived decision test used in this paper is balanced, so decision accuracy is the average of call recall and respond recall, while over-calling equals one minus respond recall. We retain all columns to make the call/response operating-point trade-off explicit. In this paper, boundary calibration means operating-point calibration between tool-call and direct-response decisions; it is not probability calibration measured by ECE, Brier score, or reliability diagrams.

E1 boundary AUC treats *should call* as positive and uses the continuous first-token probability summed over IDs that can begin `<tool_call>`. We compute ROC AUC on each seed's balanced 4,000-example test set and report its mean and sample standard deviation; a constant entry bias moves the threshold but preserves score ordering.

Decoding settings. Main single-turn APIGen-MT, BFCL, and When2Call inference uses a vLLM-backed OpenAI-compatible server with thinking disabled, greedy decoding (`temperature`=0, `top_p`=1), and dataset-specific generation lengths: 512 new tokens for APIGen-MT, 256 for BFCL, 16 for When2Call multiple choice, and 512 for When2Call judge-style outputs. The BFCL multi-turn diagnostic also uses greedy decoding, a maximum of five tool-call steps per user turn, and a fixed

simulated tool-observation harness. The strict E1 decoding evaluation in Appendix C.2 instead uses a local Transformers decoder so that a first-token-only bias can be applied; that bias is applied at the first generated assistant token of every assistant generation step.

For BFCL, we use local BFCL v4-style processed data from seven single-turn subsets: simple-python, multiple, parallel, live-simple, live-multiple, irrelevance, and live-irrelevance. The processed files contain 400, 200, 200, 258, 1053, 240, and 884 examples, respectively. Our project-local scorer checks exact AST-style matches for tool-call subsets, allowing any acceptable BFCL ground-truth value, and counts irrelevance examples as correct when no tool call is parsed. These are local BFCL-v4-style scores rather than official leaderboard submissions.

For When2Call, we use a fixed 1,000-example subset sampled with seed 42 from the 3,652-example public processed MCQ snapshot. The evaluated subset contains 329 tool-call, 289 request-for-information, and 382 cannot-answer targets. The four choices map A to direct response, B to tool call, C to request for information, and D to cannot answer; the scorer extracts the final standalone A–D answer letter. The raw 300-example judge-style file is preprocessed with a 512-token generation length for completeness, but the reported tables use only the MCQ protocol.

B.2 Training process metrics

The sample-level diagnostic records contain both raw and adjusted statistics. Raw statistics are computed before any loss modifier. Adjusted statistics are computed after Soft Clamp, Hard Clip, or Global Reweighting. In vanilla GKD, raw and adjusted values are identical. We compute every reported tag-level quantity from records that preserve the pre-packing sample boundary; packed-row aggregate tag fields are not used.

The main aggregate diagnostic is the raw per-token top-K JSD ratio:

$$R_{\mathrm{JSD}} = \frac{\sum_{i \in \mathrm{tool}} d_i / N_{\mathrm{tool}}}{\sum_{i \in \mathrm{resp}} d_i / N_{\mathrm{resp}}}, \tag{2}$$

where d_i is the token-level top-K divergence and N is the number of supervised loss tokens. In teacher-API runs, teacher log probabilities are returned for the top 32 teacher tokens; student logits are gathered on those token IDs, and both distributions are renormalized on that truncated support before computing JSD. We also use a squared-divergence proxy ratio:

$$R_{\mathrm{grad}} = \frac{\sum_{i \in \mathrm{tool}} d_i^2 / N_{\mathrm{tool}}}{\sum_{i \in \mathrm{resp}} d_i^2 / N_{\mathrm{resp}}}. \tag{3}$$

This proxy is not a parameter-gradient norm; it is a scale-sensitive diagnostic for whether a small number of large token divergences dominate the divergence mass. These ratios remove the first-order effect of response length and batch composition. Values below one mean that response examples have larger average full-sequence divergence or squared-divergence proxy.

Decision-boundary metrics inspect the first supervised token. We log the student's probability of `<tool_call>`, its log-odds margin over non-tool tokens and common response starters, and whether it is top-1. These metrics track entry pressure; full-generation evaluation captures later behavior.

B.3 Aggregate sanity diagnostics

Table 8 gives packed-safe sampled exposure and divergence diagnostics from representative runs. T/R denotes tool-call divided by response. Each method contributes 31 paired diagnostic steps, with one tool-call and one response sample retained every ten steps. Intervals use paired bootstrap resampling over steps.

Method	Sampled token T/R	Raw per-token JSD T/R	Raw squared-div. T/R
Vanilla GKD	0.813 [0.650, 1.013]	1.031 [0.734, 1.434]	1.339 [0.810, 2.168]
Hard Clip	0.811 [0.637, 1.035]	1.014 [0.692, 1.487]	1.301 [0.742, 2.215]
Global Reweight	0.807 [0.644, 1.008]	0.950 [0.679, 1.296]	1.187 [0.741, 1.848]
Soft Clamp	0.934 [0.738, 1.178]	0.983 [0.675, 1.404]	1.214 [0.706, 1.953]

Table 8: Packed-safe sampled sanity checks from representative diagnostic runs. Brackets are paired 95% bootstrap intervals. Every interval includes one, so these sparse full-sequence diagnostics do not establish a stable direction of teacher dominance.

B.4 Token-level concentration diagnostics

Table 9 gives the multi-seed numerical values behind the concentration discussion in Section 3. The concentration columns report the average share of per-sample JSD carried by the largest 1%, 5%, and 10% of supervised tokens. The shrinkage column reports the mean adjusted squared-divergence proxy divided by the raw squared-divergence proxy on diagnostic samples. The threshold-event column reports the fraction of diagnostic tokens that explicitly exceed a hard or soft clamp threshold; it is not defined for vanilla GKD or Global Reweight, because Global Reweight modifies token weights broadly rather than producing explicit clamp events.

Method	Top 1% JSD	Top 5% JSD	Top 10% JSD	Proxy shrinkage	Threshold events
Vanilla GKD	0.412±0.004	0.768±0.029	0.910±0.023	1.000±0.000	N/A
Hard Clip	0.389±0.003	0.763±0.026	0.907±0.022	0.972±0.002	0.003±0.001
Global Reweight	0.327±0.024	0.672±0.015	0.859±0.018	0.462±0.040	N/A
Soft Clamp	**0.154±0.032**	**0.507±0.021**	**0.794±0.031**	0.283±0.047	0.074±0.005

Table 9: Multi-seed diagnostic token-level signal concentration and compression. Values are mean±std over seeds 42, 44, and 60. Lower concentration indicates that fewer extreme tokens dominate the token-level JSD mass.

B.5 Mode-mismatch replay and decision-support audit

We use the APIGen-MT validation split, with 500 tool-tagged and 500 response-tagged prompts. Each of the three vanilla checkpoint-319 students samples with thinking disabled, temperature 0.9, top-$p = 1$, and at most 512 new tokens. We replay the resulting 3,000 trajectories under their routed teachers with the training objective: beta 0.5 JSD, teacher-top-32 support, and support renormalization. Confidence intervals use 10,000 prompt bootstrap replicates.

Student seed	Tool mismatch	Tool JSD ratio	Response mismatch	Response JSD ratio
42	5.8%	50.63 [32.22, 89.83]	11.6%	2.40 [1.82, 3.09]
44	4.6%	57.15 [38.41, 94.72]	12.8%	1.76 [1.23, 2.39]
60	5.6%	64.39 [36.94, 130.31]	10.4%	1.96 [1.49, 2.57]
Pooled	5.3%	55.45 [41.74, 76.44]	11.6%	2.03 [1.71, 2.39]

Table 10: Natural mode-mismatch replay. A tool mismatch is a response-mode rollout under the tool teacher; a response mismatch is a tool-mode rollout under the response teacher. JSD ratios divide mismatched by aligned per-token JSD. Brackets are 95% bootstrap intervals.

Natural mismatch may select harder prompts. We therefore sample 200 prompts per teacher tag for seed 42 and create paired continuations: one forces `<tool_call>` at the first generated position, while the other masks that token and samples a response entry. The student generates all later tokens normally. Under the tool teacher, mismatched JSD is 0.029140 per token versus 0.001090 when aligned, a ratio of 26.73 [14.87, 67.91]; 99.5% of prompts have higher mismatched JSD. Under the response teacher, the corresponding values are 0.013040 and 0.004886, a ratio of 2.67 [2.29, 3.14]; 87.5% of prompts are higher under mismatch.

Teacher → student mode	First 1	First 4	Tokens 5–10	After 10
Tool → tool	0.00266	0.00079	0.00193	0.00064
Tool → response	0.00279	0.06407	0.07706	0.03220
Response → response	0.01407	0.00817	0.00672	0.00568
Response → tool	0.01395	0.10543	0.01758	0.01055

Table 11: Per-token JSD by trajectory window in the forced-entry pairs. The first-position distributions are shared before the intervention. The mismatch appears after mode entry and extends beyond the first few tokens.

For the Qwen routed support audit, we recompute the first supervised position with local full-vocabulary teacher and matched-student logits, one prompt at a time. This avoids padding-dependent numerical changes and preserves 500 fixed validation prompts per route across seeds. Teacher coverage, mass, and rank use the 500 unique teacher distributions per route; student-dependent gradient and union quantities use all 1,500 matched prompt–student pairs. Mean response-teacher full-vocabulary $P(\texttt{<tool_call>})$ is 6.14×10^{-8}. Its strict competition rank has median 236.5, with 90th, 95th, and 99th percentiles of 614.1, 781.9, and 1086.0; ranks span 25–1243. Although top-32 retains 0.999900 of teacher probability mass, it contains <tool_call> for only 0.4% of prompts.

Support	Entry coverage	Raw teacher mass	Entry-logit descent
Response $K = 32$	0.40%	0.999900	-0.000002
Response T∪S top-32	100.00%	0.999911	-0.040748
Response $K = 32$ + entry	100.00%	–	-0.040756
Response full vocabulary	100.00%	1.000000	-0.040708
Tool $K = 32$	100.00%	1.000000	$+0.001167$
Tool full vocabulary	100.00%	1.000000	$+0.020174$

Table 12: Frozen Qwen routed support audit. Entry descent is $-\partial d/\partial z_{\text{tool}}$; negative values lower the tool-entry logit and positive values raise it. Response-teacher top-32 omits the coordinate despite near-complete raw mass, whereas teacher–student union restores it with 38.46 coordinates on average and closely matches full-vocabulary descent. The tool teacher ranks the coordinate first on all 500 tool prompts and directly reinforces it through top-32.

On the tool route, top-32 preserves the reinforcing direction but substantially attenuates its full-vocabulary magnitude despite near-unit retained teacher mass.

The seed-level and pooled entry-coordinate descents are reported in Table 2. Adding the omitted decision coordinate to top-32 recovers full-vocabulary descent on that coordinate in every seed.

The intervention adapts to the student's current boundary error. Among 1,326 response-mode entries, mean student $P(\texttt{<tool_call>})$ is 0.093 and injected descent is -0.0290. Among 174 erroneous tool-mode entries, these values rise to 0.702 and -0.1303. The injected-minus-full descent difference is -0.000048 [-0.000054, -0.000043], small relative to the full mean magnitude. By contrast, $K = 128$ minus full is 0.031549 [0.029098, 0.034049]. This audit isolates a local logit-space property of the objective; it does not measure how shared parameters integrate that signal during training.

B.6 Matched support-scope training

We train first-position and all-position support students for vanilla seeds 42, 44, and 60. All runs share initialization, teacher pair, routed data, anchored GKD settings, 319-step budget, and evaluation inputs. On response-tagged examples, the intervention appends <tool_call> with its exact teacher logit whenever top-32 omits it. The first-position variant changes only the first effective supervised position; the all-position variant applies the same rule at every supervised response position. Tool-tagged examples are unchanged.

First-position restoration. The first-position intervention supplies a direct test of the entry gradient. Table 13 reports its matched change from vanilla.

Seed	ΔE1 resp. entry	ΔE1 call entry	ΔAPIGen over-call	ΔCalls/turn	ΔLoop@3	ΔDialogue exact
42	-10.80	-11.85	-0.40	-0.101	-1.98	-0.63
44	-12.65	-11.45	-4.15	-0.561	-10.27	-2.25
60	-9.75	-10.15	$+2.40$	-0.172	-2.49	-0.75
Mean	-11.07	-11.15	-0.72	-0.278	-4.91	-1.21

Table 13: First-position support minus vanilla GKD. Entries are percentage points except calls/turn. E1 measures first-token entry; APIGen measures whether the complete generation contains <tool_call>; exact success uses the joint task-level definition in Appendix H.

The first-token E1 shift is large and consistent but not class-selective. Mean response entry error falls from 14.47% to 3.40%, while mean call entry falls from 91.72% to 80.57%. First-token accuracy changes by only -0.04 points and AUC changes from 0.9692 to 0.9712. The intervention moves the operating point instead of improving threshold-free separation.

Call-position counts explain the weaker full-generation result. Across 6,000 should-respond outputs, first-position calls fall from 805 (13.42%) to 184 (3.07%), while delayed calls rise from 47 (0.78%) to 625 (10.42%). The no-call count changes only from 5,148 (85.80%) to 5,191 (86.52%). On 6,000 should-call outputs, first-position calls fall from 90.52% to 77.68%, delayed calls rise from 0.93% to 13.78%, and the no-call rate remains 8.5%. Thus support injection often postpones entry rather than changing the final mode.

The BFCL multi-turn reductions are consistent at the task level. The paired 95% bootstrap intervals for calls/turn are $[-0.157, -0.044]$, $[-0.632, -0.491]$, and $[-0.212, -0.130]$ for seeds 42, 44, and 60. The corresponding Loop@3 intervals are $[-3.40, -0.55]$, $[-11.96, -8.64]$, and $[-3.55, -1.43]$ percentage points. Broader capability metrics expose the cost: BFCL single-turn tool-call quality falls from 79.41±1.54% to 75.72±2.74%, and BFCL single-turn overall falls by 1.69 points on average. When2Call improves by 2.63 points, consistent with a more conservative call policy. Overall dialogue exact success changes by -0.63, -2.25, and -0.75 points; its paired intervals are $[-1.88, 0.63]$, $[-3.75, -0.75]$, and $[-2.13, 0.63]$. Thus the loop reduction is consistent, whereas an endpoint-completion gain is absent and the completion decrease is resolved only for seed 44.

All-position restoration. First-position restoration leaves later response positions outside the support intervention. All-position restoration tests whether this temporal scope explains the call migration. Table 14 gives the complete-generation and multi-turn aggregates.

Metric	Vanilla	First-position	All-position
E1 response entry error	14.47±1.53	3.40±0.26	3.63±0.24
E1 tool-call entry recall	91.72±1.22	80.57±0.68	81.30±0.51
E1 boundary AUC	0.9692±0.0023	0.9712±0.0014	0.9725±0.0010
APIGen decision accuracy	88.63±0.28	88.99±0.79	87.67±0.94
APIGen over-calling	14.20±2.08	13.48±1.27	3.73±0.51
APIGen call recall	91.45±1.68	91.47±0.50	79.07±1.73
BFCL single-turn overall	78.98±1.72	77.29±1.66	78.10±1.56
BFCL tool-call quality	79.41±1.54	75.72±2.74	73.41±2.66
BFCL irrelevance refusal	78.17±2.91	80.25±3.35	86.92±3.09
When2Call MCQ	65.57±1.07	68.20±0.72	67.73±1.27
Calls/turn	1.515±0.122	1.237±0.154	0.975±0.045
Loop@3	15.08±3.01	10.17±2.43	6.82±0.22
Non-tool final	87.31±7.24	90.99±0.31	93.82±0.08
Overall dialogue exact success	7.04±0.31	5.83±1.01	3.38±0.50

Table 14: Three-seed support-scope comparison. Values are percentages except calls/turn and report mean±sample std. BFCL single-turn overall combines tool-call and irrelevance subsets, so the all-position row's higher refusal offsets its lower tool-call quality.

Call positions expose the mechanism directly. Table 15 removes the initial thinking wrapper and separates calls that begin the answer from calls that appear after natural-language tokens.

Target	Training	First-position call	Delayed call	No call
Should respond	Vanilla	13.42	0.78	85.80
	First-position	3.07	10.42	86.52
	All-position	3.08	0.65	96.27
Should call	Vanilla	90.52	0.93	8.55
	First-position	77.68	13.78	8.53
	All-position	77.23	1.83	20.93

Table 15: Tool-call position over 6,000 complete generations per target class, pooled across three seeds (percent). First-position restoration relocates calls; all-position restoration removes most delayed calls in both classes.

Relative to first-position support, all-position support reduces APIGen over-calling by 10.10, 8.35, and 10.80 points across seeds 42, 44, and 60. The same seeds lose 12.75, 10.70, and 13.75 points of APIGen call recall. Loop@3 falls by 5.84, 1.02, and 3.19 points. The intervention therefore converts the previous call relocation into a consistent full-generation effect, but does so by moving the shared operating point beyond the first-position treatment.

E1 confirms that temporal coverage, rather than a larger initial boundary shift, separates the two interventions. All-position response entry error is $3.63\pm0.24\%$, tool-call entry recall is $81.30\pm0.51\%$, and AUC is 0.9725 ± 0.0010—all close to the first-position values of $3.40\pm0.26\%$, $80.57\pm0.68\%$, and 0.9712 ± 0.0014. The large complete-generation gap therefore opens after the first token.

The task-level decomposition gives the same conclusion. Empty-ground-truth no-call accuracy progresses from $37.70\pm0.51\%$ for vanilla to $47.41\pm6.52\%$ for first-position and $61.49\pm4.54\%$ for all-position support. Observed required-turn call coverage moves in the opposite direction: $92.90\pm0.33\%$, $83.15\pm5.61\%$, and $73.24\pm4.62\%$. All-position overall dialogue exact success is $3.38\pm0.50\%$, versus $7.04\pm0.31\%$ for vanilla and $5.83\pm1.01\%$ for first-position support. Training with the restored coordinate therefore causally changes the learned boundary, but shared parameters do not confine the shift to response examples.

B.7 Non-tool controls and student-aware support

The probability-matched placebo uses the same all-position response scope as exact restoration, but appends a non-tool tail token whose teacher probability is closest to that of <tool_call>. The support-level strategy appends all student-top-32 tokens missing from teacher top-32 and queries their teacher logits. Across 319 steps and three seeds, teacher–student overlap is 25.12 ± 0.08 tokens, so the union support averages only 38.88 ± 0.08 rather than 64.

Its within-position support is $U_i^{(K)} = \mathrm{TopK}(p_t)_i \cup \mathrm{TopK}(p_s)_i$, and the same restricted JSD is evaluated on $U_i^{(K)}$. Returned support grows by 21.5% relative to teacher top-32, and observed steady-state step time rises by 16.0%. Table 12 reports the frozen routed closure; it does not claim full mediation of the training effect.

The placebo controls support count, scope, and teacher-probability scale, but not student probability or effective gradient strength. Its loop rate and overall dialogue exact success also move, while its local entry and full-generation shifts remain much smaller than exact restoration. A separate response-side <tool_call> logit-penalty control ($\lambda = 0.15$) produces only a small operating-point shift, reducing APIGen over-calling from $14.20\pm2.08\%$ to $12.93\pm1.14\%$ while lowering overall dialogue exact success from $7.04\pm0.31\%$ to $4.96\pm0.07\%$ across three matched seeds. As the student-aware support-level strategy, the support union raises mean APIGen decision accuracy, BFCL single-turn overall, and When2Call MCQ by 1.19, 2.55, and 2.70 matched points. Its required-turn call coverage nevertheless falls by 11.69 points, explaining the lower overall dialogue exact success.

Metric	Vanilla	Prob. placebo	Exact	Grad. reroute	Support union
E1 response entry error	14.47±1.53	13.37±1.10	3.63±0.24	9.25±3.43	8.72±0.74
E1 tool-call entry recall	91.72±1.22	91.30±1.00	81.30±0.51	85.83±2.92	90.55±1.00
E1 boundary AUC	0.9692±0.0023	0.9695±0.0036	0.9725±0.0010	0.9609±0.0082	0.9760±0.0006
APIGen over-calling	14.20±2.08	13.25±1.96	3.73±0.51	11.65±2.88	7.38±0.64
APIGen call recall	91.45±1.68	90.37±2.31	79.07±1.73	86.93±3.31	87.02±2.01
BFCL single-turn overall	78.98±1.72	78.95±0.44	78.10±1.56	76.76±1.04	81.53±0.93
When2Call MCQ	65.57±1.07	68.53±0.92	67.73±1.27	26.70±22.32	68.27±1.11
Calls/turn	1.515±0.122	1.346±0.021	0.975±0.045	1.464±0.481	1.118±0.057
Loop@3	15.08±3.01	12.44±1.41	6.82±0.22	15.78±9.97	8.05±0.87
Required-turn call	92.90±0.33	86.64±4.06	73.24±4.62	83.43±7.84	81.22±2.38
Overall dialogue exact success	7.04±0.31	4.92±0.40	3.38±0.50	5.21±0.44	4.12±0.33

Table 16: Three-seed non-tool controls and support-construction comparison (percent except AUC and calls/turn; mean±sample std). Gradient rerouting matches the forward loss and selected output-logit-gradient magnitude but does not reproduce exact-restoration behavior; student-aware union is the support-level strategy.

Support union − Vanilla	Seed 42	Seed 44	Seed 60	Mean [95% CI]
APIGen decision accuracy	+1.87	+0.98	+0.72	+1.19 [−0.31, +2.69]
APIGen over-calling	−5.95	−9.85	−4.65	−6.82 [−13.54, −0.09]
E1 boundary AUC	+0.46	+0.78	+0.79	+0.68 [+0.20, +1.15]
BFCL single-turn overall	+2.75	+4.08	+0.81	+2.55 [−1.54, +6.63]
When2Call MCQ	+3.50	+2.50	+2.10	+2.70 [+0.91, +4.49]
Required-turn call	−12.41	−13.29	−9.36	−11.69 [−16.81, −6.57]
Overall dialogue exact success	−3.38	−3.13	−2.25	−2.92 [−4.38, −1.45]

Table 17: Matched support-union deltas in percentage points. The interval is a two-sided 95% t interval over the three training-seed deltas; with three seeds it is descriptive rather than a high-powered significance test.

B.7.1 Output-logit-gradient-reroute control

The synthetic output-logit-gradient-reroute control reuses the natural placebo's all-position eligibility and legal non-tool candidate. Let z_t and z_c be the student logits of the tool-entry token t and selected candidate c. The appended student value is the straight-through proxy

$$\widetilde{z}_{t\to c} = \mathrm{sg}(z_t) + z_c - \mathrm{sg}(z_c),$$
$$\widetilde{z}_{t\to c} = z_t \text{ (forward)}, \qquad \frac{\partial \widetilde{z}_{t\to c}}{\partial z_t} = 0, \qquad \frac{\partial \widetilde{z}_{t\to c}}{\partial z_c} = 1. \tag{4}$$

Here sg stops gradients. The appended teacher value remains the response-teacher log-probability of t, so the forward loss equals exact restoration while its selected output-logit gradient is routed to c.

Of 1,500 frozen prompt–student pairs, 1,494 are eligible and all pass in every seed. The remaining six are the same two prompts repeated across three student seeds; response-teacher top-32 already contains t, so no restoration or rerouting is applicable. On the eligible pairs, maximum forward-JSD and coefficient errors are zero, and maximum L1/L2 logit-gradient norm relative errors are 2.30×10^{-7} and 1.19×10^{-7}.

Table 16 reports the behavioral comparison. The gradient-reroute-minus-exact APIGen over-calling differences are $+8.25$, $+4.90$, and $+10.60$ points across seeds. This supports the identity of the output coordinate as one contributor, but not full-parameter-gradient isolation: the hidden-state term changes from gW_{tool} to $gW_{\mathrm{candidate}}$. The candidate/tool row-norm ratio has median 1.1525 and row cosine median -0.0172; embeddings are untied. The low, high-variance When2Call result and lower BFCL single-turn score are therefore substantive collateral effects.

C Additional Benchmark Results

C.1 Additional APIGen-MT results

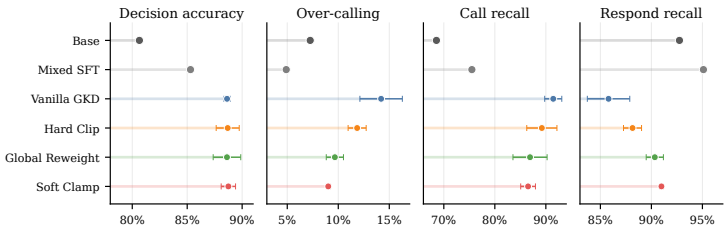

Figure 4: APIGen-MT loss-shaping trade-off. Error bars mark standard deviation over three seeds. Soft Clamp reduces over-calling relative to vanilla GKD while preserving decision accuracy, at the cost of lower call recall. Support construction is reported separately in Table 16.

Case	User state	Vanilla GKD	Soft Clamp
Airline baggage	User asks to add one checked bag and believes it is covered by free allowance.	Calls a baggage-update tool directly.	Explains the free-baggage allowance and asks for confirmation before updating.
Retail address	User asks to change Suite 716 to Suite 1000 at the same street address.	Calls an address-modification tool directly.	Summarizes the address change and asks the user to confirm.
Flight refund	User asks whether a completed trip can receive refund or credit due to membership status.	Starts another tool-call trajectory after stating the policy.	Directly explains that completed flights cannot be refunded or cancelled.

Table 18: Qualitative APIGen-MT should-respond examples. Vanilla GKD enters a tool-call trajectory when the correct behavior is to answer, explain policy, or request confirmation. Soft Clamp stays on the response side of the behavior boundary.

C.2 Decoding-level entry-bias evaluation

We evaluate a non-positive scalar bias $b \in [-5, 0]$ on vanilla GKD's first-token `<tool_call>` logit as a decoding-level strategy and test whether it can reproduce Soft Clamp's boundary effect. With thinking disabled, *response tool-entry error* is the fraction of should-respond examples whose top-1 first token is `<tool_call>`; full-generation APIGen over-calling remains a separate metric.

For each seed, we sweep b from -5 to 0 in steps of 0.05 on validation and select operating points matching either Soft Clamp's response entry error or call recall. Ties use the smallest $|b|$; the bias is then frozen on test. E1 uses a common local Transformers scorer, whereas full-generation APIGen uses vLLM, so the decoding comparison remains within its own protocol.

Setting	Resp. entry err.	Tool call rec.	First-token acc.
Vanilla GKD, no bias	14.5±1.5	91.7±1.2	88.6±0.4
Soft Clamp, no inference bias	11.1±0.3	90.4±0.2	89.7±0.1
Support union, no inference bias	8.7±0.7	90.6±1.0	90.9±0.2
Biased vanilla, matched entry error	11.2±0.6	89.2±1.2	89.0±0.6
Biased vanilla, matched call recall	11.6±1.1	89.5±0.6	88.9±0.5

Table 19: Cross-layer first-token comparison on APIGen-MT test. Values are percentages over three paired seeds, with sample standard deviations. Biases are tuned on validation and frozen on test. The support union changes training support and is not bias-matched.

Seed	Val-matched response-entry bias	Val-matched call-recall bias
42	-0.40	-0.50
44	-0.65	-0.40
60	-0.25	-0.25

Table 20: Validation-selected non-positive <tool_call> entry biases used for the decoding-level evaluation. Biases are selected on APIGen-MT validation and frozen for APIGen-MT test and strict BFCL multi-turn evaluation.

Validation-tuned bias nearly matches Soft Clamp's response entry error (11.2% versus 11.1%) but has lower call recall (89.2% versus 90.4%) and accuracy (89.0% versus 89.7%). Its unchanged AUC supports threshold movement, whereas support union's 0.9760 AUC is consistent with modest separation change.

We next make the deployment preference explicit. Let $\rho = C_{\text{over}}/C_{\text{miss}}$ be the cost of an unnecessary call relative to a missed required call. For response-side false-positive rate e_{over} and call-side false-negative rate e_{miss}, we report the class-balanced normalized risk

$$R_\rho = \frac{\rho e_{\text{over}} + e_{\text{miss}}}{\rho + 1}. \tag{5}$$

At each ρ, the inference-bias row selects b independently for each seed by minimizing validation risk, then freezes it on test. Figure 5 keeps the two protocols separate: panel (a) uses APIGen complete generations for the fixed trained models, whereas panel (b) uses E1 first-token decisions.

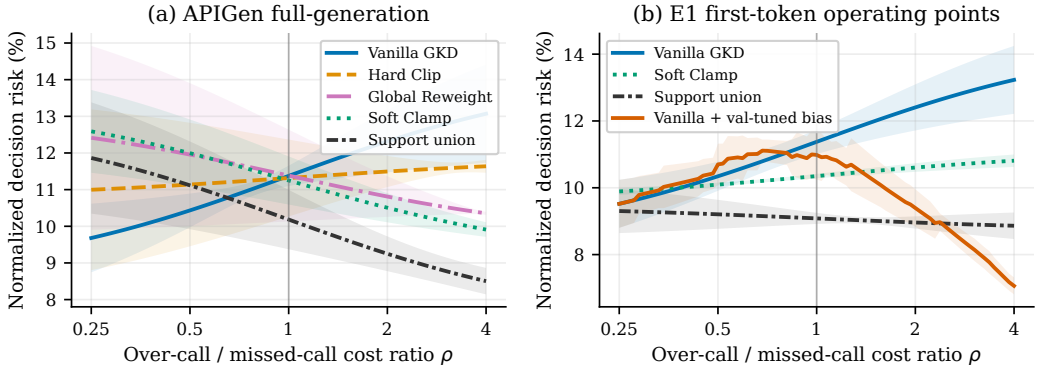

Figure 5: Cost-sensitive operating points. Lower is better. Panel (a) compares fixed trained GKD variants on APIGen full-generation decisions. Panel (b) compares fixed vanilla GKD, Soft Clamp, and the support union with a per-seed inference bias selected on E1 validation for each cost ratio and frozen on test. Lines and bands are three-seed means and sample standard deviations. The panels use distinct protocols and are not pooled.

Vanilla has the lowest mean APIGen risk at $\rho \in \{0.25, 0.5\}$ and support union at $\rho \in \{1, 2, 4\}$. On E1, union is lowest through $\rho = 2$; cost-aware bias is lowest at $\rho = 4$ (7.06%). These are descriptive three-seed comparisons.

Vanilla GKD has 14.5±1.5% first-token entry error and 14.3±2.2% full-generation over-calling, close to the main-table 14.2±2.1%. For Soft Clamp, the values are 11.1±0.3% and 9.0±0.2%. The larger gap comes from cases that start with <tool_call> but do not retain the marker in the thinking-stripped output. Because E1 and full generation use different backends, this disagreement is protocol-specific.

We also run a strict BFCL multi-turn decoding comparison with the same local Transformers decoder, prompt rendering, parser, and bias hook for every row. The main BFCL diagnostic uses vLLM, which lacks a first-token-only hook; backend, batching, and template differences can change absolute values. We therefore compare rows only within Table 21.

Setting	Calls/turn	Loop@3	Loop@5	Repeat	Non-tool fin.	Invalid
Vanilla	1.734±0.057	20.5±1.5	15.1±1.6	23.7±4.3	78.8±7.0	2.3±2.2
Soft Clamp	1.494±0.053	15.6±0.8	10.1±1.1	16.6±1.5	88.2±1.3	0.7±0.1
Biased, entry-matched	1.492±0.089	15.7±1.0	10.3±1.2	17.6±0.6	85.8±1.5	1.6±1.2
Biased, recall-matched	1.525±0.062	16.3±1.3	11.0±1.3	18.8±4.0	84.7±4.8	1.8±1.5

Table 21: Strict BFCL multi-turn decoding comparison with the same local first-token-bias decoder for all rows. Values are three-seed mean±sample std. The per-seed biases selected on APIGen-MT validation are reused here without BFCL tuning. Inference-time scalar bias reproduces nearly all of the mean reduction in tool-call frequency and loop rate under this strict decoder, but Soft Clamp retains higher non-tool-final rate and lower invalid-call rate. Non-tool final is a termination metric, not answer correctness.

C.3 BFCL AND WHEN2CALL TABLES

Method	Overall	Tool Call Quality	Irrelevance Refusal
Base	82.2	83.5	79.8
Mixed SFT	82.9	82.0	84.4
Vanilla GKD	79.0±1.7	79.4±1.5	78.2±2.9
Hard Clip	79.9±0.6	78.1±1.2	83.2±1.2
Global Reweight	79.8±2.7	77.8±5.7	83.5±3.4
Soft Clamp	80.8±0.1	79.7±1.3	82.7±2.4
Support union	81.5±0.9	78.9±0.6	86.5±1.5

Table 22: BFCL results. All values are percentages. GKD rows report mean±std over three seeds; columns rank the strategies differently.

Method	MCQ Acc	tool_call	request_for_info	cannot_answer
Base	72.8	88.1	61.6	68.1
Mixed SFT	71.6	88.8	63.3	63.1
Vanilla GKD	65.6±1.1	88.0±1.4	58.9±5.8	51.2±4.1
Hard Clip	64.3±1.3	87.9±1.8	57.3±4.7	49.2±4.5
Global Reweight	67.1±2.2	87.3±2.5	61.5±6.9	53.8±3.4
Soft Clamp	65.9±2.5	87.7±1.2	60.3±2.2	51.3±6.7
Support union	68.3±1.1	87.6±2.0	59.8±6.2	58.0±4.6

Table 23: When2Call results. All values are percentages. GKD rows report mean±std over three seeds; category means rank the strategies differently.

D SECONDARY QWEN3.5-4B STUDENT-SCALE CHECK

This secondary experiment replaces the Qwen3.5-9B student with Qwen3.5-4B while retaining the fixed 9B teachers. It predates the support-level interventions and therefore tests only whether the loss-shaping pattern persists at a smaller student scale; it is not evidence for support-aware correction.

Method	Dec. Acc	Over-call	Call Rec.	Resp. Rec.
Mixed SFT	88.5	7.5	84.5	92.5
Vanilla GKD	88.4±0.8	12.8±0.7	89.6±2.1	87.2±0.7
Global Reweight	87.8±0.8	9.5±0.8	85.1±1.9	90.5±0.8
Soft Clamp	87.6±0.6	9.2±1.2	84.4±1.8	90.8±1.2

Table 24: Qwen3.5-4B APIGen-MT decision results (percent). GKD rows report mean±sample std over seeds 42, 44, and 60; Mixed SFT is a single reference.

Global Reweight and Soft Clamp reduce over-calling from 12.8±0.7% to 9.5±0.8% and 9.2±1.2%. Their Loop@3 rates are 8.9±1.4% and 9.2±2.6%, versus 15.7±3.2% for vanilla; neither raises

overall dialogue exact success. These results retain a scale check for loss shaping without conflating it with the paper's support-level mechanism claim.

E LLAMA-3.1 CROSS-FAMILY REPLICATION

The Llama replication uses Meta-Llama-3.1-8B-Instruct for the student and both specialized teachers. Tool-call SFT renders native assistant JSON of the form `{"name": ..., "parameters": ...}`, while tool observations use the model's native tool-response role. Mixed, response-only, and tool-only SFT checkpoints train for two epochs. Mixed SFT is a separate supervised reference; Vanilla GKD, Soft Clamp, and the support union run each initialize from the original Instruct checkpoint and share one epoch, seeds 42/44/60, teacher top-32, $\lambda = 0.8$, $\alpha = 0.3$, $\beta = 0.5$, $\tau = 0.9$, and a 512-token rollout limit. The evaluation renderer and normalizer preserve this native protocol before mapping calls to the common metric representation.

Model	Dec. Acc	Over-call	Call Rec.	Resp. Rec.	When2Call	BFCL
Base	57.13	59.65	73.90	40.35	46.10	47.94
Mixed SFT	90.83	10.25	91.90	89.75	37.80	42.29
Response SFT	50.00	0.00	0.00	100.00	14.80	34.74
Tool-call SFT	50.00	100.00	100.00	0.00	33.80	45.29
Vanilla GKD	84.24±0.27	28.80±0.80	97.27±0.33	71.20±0.80	12.93±3.86	52.27±8.24
Soft Clamp	86.35±0.71	22.35±1.39	95.05±0.85	77.65±1.39	12.77±3.93	48.75±8.26
Support union	90.08±0.13	11.08±1.16	91.23±0.94	88.92±1.16	17.27±4.16	53.59±8.02

Table 25: Llama-3.1-8B single-turn results (percent). GKD rows report mean±sample std over three seeds; Base and SFT rows are single references. The one-sided SFT teachers verify the intended routing extremes. BFCL varies substantially across GKD seeds, and all GKD variants remain below mixed SFT on When2Call.

The frozen response-side support audit samples 500 unique response prompts and evaluates the same prompts against vanilla students from all three seeds. A complementary tool-side audit repeats this protocol on 500 unique tool prompts with the tool teacher. The configured entry prefix `{"name"` begins with `{"` (ID 5018), which is the audited behavior coordinate. Teacher-only support metrics use the 500 unique prompts per route. Support-union and gradient metrics pool 1,500 matched teacher–student pairs per route.

Support	Entry coverage	Raw teacher mass	Entry-logit descent
Response $K = 32$	0.00%	0.999770	0.000000
Response $K = 64$	0.00%	0.999899	0.000000
Response $K = 128$	0.00%	0.999947	0.000000
Response $K = 256$	0.00%	0.999968	0.000000
Response T∪S top-32	97.93%	0.999786	−0.049229
Response full vocabulary	100.00%	1.000000	−0.049180
Tool $K = 32$	100.00%	1.000000	+0.005174
Tool full vocabulary	100.00%	1.000000	+0.009028

Table 26: Frozen Llama routed support audit. Entry descent is $-\partial d/\partial z_{\text{entry}}$; negative values lower the JSON-entry logit and positive values raise it. Response-teacher top-32 retains 99.977% raw mass while omitting the decision coordinate on every response prompt; student-aware union recovers nearly the full-vocabulary correction with 40.49 coordinates on average. Tool-teacher top-32 contains the coordinate on every tool prompt and directly reinforces it.

The response teacher's strict competition entry rank spans 10,381–127,884, with median 105,952.5 and 90th percentile 122,199.9. Union entry coverage is 99.2%, 97.6%, and 97.0% for vanilla student seeds 42, 44, and 60. These endpoint students assign enough probability to the omitted coordinate for it to enter their top-32, which is precisely the error teacher-only support cannot directly correct. By contrast, the tool teacher ranks the entry token first on all 500 tool prompts, with mean full-vocabulary

probability 0.999999; its top-32 entry descent is positive for all three matched students. The trained support-union models' APIGen results in Section 4.4 provide the corresponding behavioral test; the broader matched temporal-scope causal chain remains the Qwen primary study.

F Loss Modifier Details

F.1 Vanilla GKD

Vanilla GKD uses the original token divergence d_i without additional calibration. It is the reference point for measuring whether a loss modifier reduces behavior imbalance.

F.2 Hard Clip

Hard Clip applies

$$d'_i = \min(d_i, c), \tag{6}$$

with $c = 0.5$ in the main experiment. This baseline tests whether simply truncating extreme token losses can reduce over-calling. It is a practical contrast rather than a mechanism-matched control: it uses a fixed threshold, removes marginal gradients above the threshold, and affects a different number of tokens from Soft Clamp.

F.3 Global Reweight

Global Reweight is a batch-relative baseline. Let

$$z_i = \text{clip}\left(\frac{d_i - \text{mean}(d)}{\text{std}(d)}, -z_{\max}, z_{\max}\right). \tag{7}$$

The token weight is

$$w_i = \frac{\exp(-\alpha z_i)}{\text{mean}_j \exp(-\alpha z_j)}, \tag{8}$$

followed by clipping to $[w_{\min}, w_{\max}]$. The adjusted divergence is $d'_i = d_i \,\text{stopgrad}(w_i)$. We use $\alpha = 0.3$, $z_{\max} = 3.0$, $w_{\min} = 0.25$, and $w_{\max} = 2.0$. This method is useful as a contrast because it reweights all tokens by batch-relative divergence, while Soft Clamp leaves non-extreme tokens unchanged.

F.4 Soft Clamp

Soft Clamp sets a detached dynamic threshold $C = \text{stopgrad}(k \,\text{mean}_i(d_i))$ for each batch. We use $k = 3.0$ in the main experiment. Tokens with $d_i \le C$ are unchanged. Tokens with $d_i > C$ use

$$d'_i = d_i \frac{C}{\text{stopgrad}(d_i)}. \tag{9}$$

The forward value is capped at C, but the token still receives a nonzero gradient scaled by C/d_i. Detaching the threshold prevents gradients from flowing through the batch mean and matches the implementation.

G Reproducibility Checklist

Within each family, matched GKD variants share initialization, teachers, split, training budget, batch and sequence settings, rollout sampling, temperature, anchor weight, teacher top-K, and evaluation data. Training uses three seeds. Loss runs differ only by modifier; support runs change only distilled support and teacher probes. We store outputs, scores, logs, diagnostic samples, and plotting inputs. The artifact includes the modified backend, sanitized training templates, evaluation and plotting scripts, metric definitions, and aggregate results.

H BFCL MULTI-TURN LOOP DIAGNOSTIC

The BFCL multi-turn diagnostic evaluates whether a model with a stronger tool-call prior becomes less usable in an interactive tool environment. It uses 800 BFCL v4 multi-turn tasks and 3,136 user turns sampled from the local processed BFCL multi-turn data. When the model emits a tool call, the harness returns a fixed simulated tool observation generated by the local BFCL-style harness and continues the dialogue until the model gives a non-tool final answer or reaches the maximum step limit. Invalid calls are counted when the model emits a tool-call marker that the harness cannot parse into a supported function call.

Tables 27 and 28 report the full interaction and correctness results before we define their components.

Method	Calls/turn	Loop@3	Loop@5	Repeat call	Non-tool final	Dialogue exact
Mixed SFT	0.974	5.1	0.7	2.5	96.5	5.1
Vanilla GKD	1.515±0.122	15.1±3.0	8.8±3.5	17.7±6.4	87.3±7.2	7.0±0.3
Hard Clip	1.348	11.5	6.3	14.4	91.5	6.0
Global Reweight	1.297±0.088	10.8±1.4	5.5±1.3	13.1±2.7	93.6±1.6	5.3±0.9
Soft Clamp	1.289±0.023	10.3±0.4	5.0±0.3	11.3±0.6	93.9±0.2	6.2±0.4
Support union	1.118±0.057	8.0±0.9	3.3±0.5	8.9±1.0	92.3±0.6	4.1±0.3

Table 27: Full BFCL multi-turn interaction results. Values except calls/turn are percentages. "Non-tool final" records termination, not correctness. Three-seed methods report mean±sample std; Hard Clip and Mixed SFT are single runs.

Method	Required call	Supported req. exec/state	Empty-GT no-call	Turn prot.	Overall dlg. exact	Supported dlg. exact
Mixed SFT	79.77	29.46	54.85	32.56	5.12	6.83
Vanilla GKD	92.90±0.33	34.38±0.62	37.70±0.51	35.57±0.54	7.04±0.31	9.39±0.42
Hard Clip	88.88	32.69	44.90	34.55	6.00	8.00
Global Reweight	87.78±4.62	30.58±2.53	43.37±6.80	32.65±2.06	5.33±0.92	7.11±1.23
Soft Clamp	88.50±2.01	32.15±0.80	42.96±6.13	33.93±1.17	6.17±0.38	8.22±0.51
First-position support	83.15±5.61	29.67±3.18	47.41±6.52	32.49±2.50	5.83±1.01	7.78±1.35
All-position support	73.24±4.62	24.76±1.80	61.49±4.54	28.60±1.50	3.38±0.50	4.50±0.67

Table 28: Turn-checkpoint and dialogue-level correctness in the 9B BFCL multi-turn diagnostic (percent). Required call and Empty-GT no-call use all 3,136 observed user turns. "Supported" metrics exclude the missing-function category not fully implemented by the local harness. "Turn prot." denotes turn-level protocol success; "dlg. exact" denotes dialogue exact success. Multi-seed rows report mean±sample std; Mixed SFT and Hard Clip are single runs.

Calls/turn counts emitted calls per user turn; Loop@3 and Loop@5 mark runs of at least three or five consecutive call steps; *repeat* requires the same function and arguments. The five-step harness makes Loop@5 identical to a max-step hit. *Non-tool final* records termination, not answer correctness.

For dialogue exact success, the stored checker executes model and reference calls from the same initial state and requires matching observations and environment state on required turns, plus no executable call on empty-ground- truth turns. Turn-level execution/state uses the same check at each observed checkpoint. This custom-harness metric is not an official BFCL leaderboard score and does not evaluate final-answer language quality.

The harness did not implement BFCL's missing-function supplementation prompt; its 200 placeholders are excluded from observed-turn denominators, and *supported* summaries exclude that category. Offline replay reproduces all 13,600 stored checker decisions. On supported categories, vanilla dialogue exact rises from 7.04% to 9.39%, with 35.57% turn-level protocol success; fewer loops still do not improve endpoint correctness.