The Golden 30 Minutes: Controlled Dialogue Design of an AI Collaborator for Crisis Intervention in Special Education

Anonymous Authors[1]

Abstract

In-class crises among mainstreamed students with special needs (emotional outbursts, self-harm tendencies, teacher–student standoffs) demand intervention within the *golden 30 minutes*, yet general-education teachers often mishandle events due to a lack of validated scripts. General-purpose conversational AIs further risk cultural mismatch and overreaching recommendations. We propose **G30F**, a controllable AI collaborator framework with three operationalized layers: (i) a content-safety layer combining a multicultural lexicon (218 high-risk terms, R1–R5) with emotion-intensity thresholding (E1–E7); (ii) a structural-control layer encoding authoritative crisis guidelines into a three-stage finite-state machine (Soothe→Guide→Resolve/Refer) with explicit circuit-breakers; and (iii) an ethics-constraint layer enforcing human authority via a *Teacher–Policy Weight* (TPW) with automatic fallback to observer mode. We instantiate G30F on a 7–8B open-source backbone trained via SFT(LoRA) → DPO → PPO with a composite reward. On 1,840 teacher-labeled events from 118 schools and a two-phase human study of 200 dialogues, G30F improves stage-match accuracy by 19 pp, cultural safety by 12.3 pp, and ethics compliance by 14 pp over single-objective baselines ($p < 0.05$, Holm–Bonferroni), while remaining deployable on teachers' existing devices.

1. Introduction

Inclusive education has substantially expanded access for students with special needs in China and globally. However, high-conflict crisis events—emotional outbursts, self-harm, and teacher–student standoffs—are highly time-sensitive: prevailing guidance frames a critical *golden 30-minute* window for safe de-escalation. OECD-wide, teachers report losing roughly 13% of classroom time to keeping order (OECD, 2017), and a school-wide PBIS randomized trial showed an 18% reduction in concentration problems and 33% fewer disciplinary referrals after structured early de-escalation (Bradshaw et al., 2012).

Existing technical solutions face three persistent gaps. **(1) Cultural mismatch.** Meta-analyses report culturally adapted interventions outperform non-adapted ones ($g \approx 0.3$–0.5) (Hall et al., 2016; Griner & Smith, 2006), yet production AI systems offer limited dialect-aware support. **(2) Overreach risk.** General-purpose generative systems tend to displace teacher judgment in high-conflict scenarios, violating the *human-led* principle emphasized by UNESCO guidance (UNESCO, 2023). **(3) Brittle static rules.** Keyword-based filters miss context-dependent risk, and once tripped offer no graceful fallback.

We propose the **Golden 30-Minute Framework (G30F)**, which enables safe, adaptive, and compliant in-classroom crisis intervention through three operationalized control layers (Fig. 1): (i) a *content-safety layer* based on localized dialect/forbidden-term lexicons and high-risk semantic filtering; (ii) a *structural-control layer* encoding authoritative guidelines into a three-stage finite-state machine with built-in circuit-breakers that prohibit step-skipping (REMS TA Center, 2013; Hong Kong Education Bureau, 2019); and (iii) an *ethics-constraint layer* that quantifies the human–AI authority boundary via a Teacher–Policy Weight (TPW), automatically switching to observer mode when teacher authority drops below threshold. Training follows the loop *public manuals → teacher labels → expert correction → DPO/PPO optimization*, with all three guardrails active at inference and meeting on-device compliance (OECD, 2021).

2. The G30F Framework

2.1. Three-Stage FSM and Circuit-Breakers

The dialogue engine operates over a finite-state protocol $\mathcal{S} = \{S_1, S_2, S_3\}$ corresponding to *Soothe* → *Guide* → *Resolve/Refer*, with a legal edge set $E \subset \mathcal{S} \times \mathcal{S}$ prohibiting step-skipping. We operationalize emotion intensity on a

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

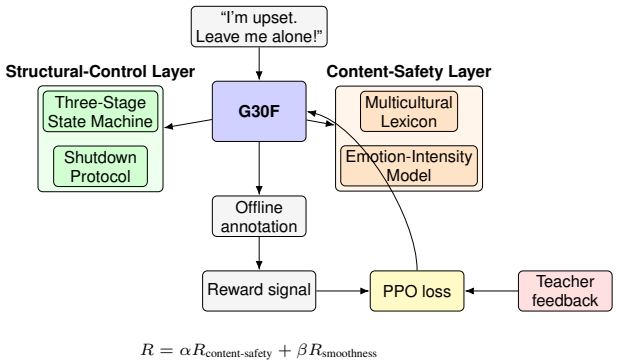

Figure 1. **G30F training–inference loop.** A student's crisis utterance is routed to G30F under two always-on guardrails: a three-stage FSM with circuit-breaker protocol, and a content-safety layer combining a multicultural lexicon with an emotion-intensity model. Teacher feedback is converted into annotations and aggregated into the composite reward.

7-level scale (E1–E7) and define three circuit-breaker conditions: behavioral (sustained clenched fists, raised volume with refusal language for ≥ 3 minutes), physiological where available (heart rate $>$120 bpm *and* fist-clenching frequency $>$1.2 Hz for ≥ 3 minutes (Posner et al., 2011)), and lexical (any R5 trigger). All rules are externalized into editable CFG files linked to FSM transitions to ensure verifiability (REMS TA Center, 2013).

2.2. Content-Safety Layer

We curate 218 high-risk terms graded R1–R5 with recommended alternatives. R5 (circuit-breaker; immediate stop) contains 32 terms (e.g., *"If you keep this up you'll be expelled"*); R4 contains 45 (e.g., *"so stupid"*); R3 contains 61 (e.g., *"don't act like a child"*); the remaining 80 are R2/R1 (milder but context-sensitive). The resource includes dialectal mappings for six major Chinese dialect groups and is stored in CSV for fast loading and rule audit (SAMHSA, 2012).

2.3. Ethics-Constraint Layer: Teacher–Policy Weight

To prevent overreach, we define a runtime scalar $\mathrm{TPW} \in [0, 1]$ representing the share of decision authority retained by the teacher in the current turn. Normal operation requires $\mathrm{TPW} > 0.7$; when $\mathrm{TPW} < 0.3$ (e.g., the model becomes overconfident in a borderline ethics scenario), the system automatically falls back to *observer mode*: generation halts and the teacher resumes full control. This dynamic re-allocation is the runtime counterpart of the *human-led, AI-assisted* principle.

2.4. Training Pipeline

The backbone is an open-source autoregressive LLM at the 7–8B scale (Llama 3), adapted via LoRA (Hu et al., 2021; Meta AI, 2024).

SFT. We initialize G30F using teacher demonstrations and curated safe corpora with the standard cross-entropy objective $\mathcal{L}_{\mathrm{SFT}} = -\sum_i \sum_j y_{ij} \log \hat{y}_{ij}$.

DPO. Teachers labeled multi-school pilot logs as *effective*, *ineffective*, or *counterproductive*. From these we construct preference pairs (y_w, y_l) and align the model with expert judgment via Direct Preference Optimization (Rafailov et al., 2023).

PPO with composite reward. On top of SFT+DPO, we run PPO (Schulman et al., 2017) with

$$R = \alpha S_{\text{stage}} + \beta S_{\text{cultural}} + \gamma S_{\text{ethics}} - \lambda S_{\text{violation}}, \quad (1)$$

where $S_{\text{stage}} = \frac{1}{T}\sum_t \mathbf{1}_{\text{legal}}(s_t \rightarrow s_{t+1})$ is stage-match accuracy, S_{cultural} is taboo/region compatibility (discriminator-scored), S_{ethics} penalizes overreach and bypass of referral, and $S_{\text{violation}}$ enforces fallback under illegal skips, high-risk triggers, or overreach. Weights $(\alpha, \beta, \gamma, \lambda)$ are tuned on validation data via Bayesian optimization (Snoek et al., 2012).

3. Data Pipeline

We build a reproducible data system via four stages: *experience accumulation* $\rightarrow$ *rule distillation* $\rightarrow$ *in-practice verification* $\rightarrow$ *intelligent optimization*.

MVP-6 corpus. Six representative manuals/guidelines—national directives, regional school crisis-intervention handbooks, and international resources (WHO LIVE LIFE Chinese edition (World Health Organization, 2018))—yield 427 stage-aligned utterance templates (192 Soothe, 153 Guide, 82 Resolve/Refer), achieving 89% scenario coverage (17/19 scenario types).

Teacher pilot. A pilot across 118 partner schools collected 1,840 valid events labeled *effective* (62%, predominantly E1–E3 and Mandarin-dominant areas), *ineffective* (30%, of which 76% occur in dialect regions), and *counterproductive* (8%, with 83% involving R3-level misuse). Events were aggregated into a Region $\times$ Scenario $\times$ Effect tensor used to drive cultural-adaptation updates.

Expert iteration. Twelve psychology, special-education, and linguistics experts conducted three Delphi rounds (Linstone & Turoff, 2002), producing auditable revisions (89

Table 1. Automatic evaluation. Compliance (SMA/CSA/EBC) and language quality (PPL/R_{len}/D2). Best in bold; second-best underlined. $^{*}p < 0.05$, $^{**}p < 0.01$ (Welch + Holm–Bonferroni).

Model	SMA↑	CSA↑	EBC↑	PPL↓	R_{len}	D2↑
SFT	0.61	0.68	0.72	18.4	42.3	0.31
DPO-only	0.67	0.71	0.75	15.2	45.1	0.38
PPO-CS	0.64	<u>0.82</u>	0.74	16.8	38.7	0.35
PPO-SM	<u>0.78</u>	0.69	0.73	17.1	44.2	0.33
PPO-EBC	0.66	0.70	<u>0.84</u>	16.5	36.4	0.34
PPO-NoPenalty	0.74	0.76	0.79	<u>15.8</u>	46.8	<u>0.40</u>
G30F-Full	**0.80****	**0.84****	**0.86****	**14.6***	43.5	**0.41***

dialectal variants; +37 high-risk handling templates). Effective rate in dialect regions rose from 39% to 68%; high-risk sub-scenario coverage rose from 62% to 91%.

Final strategy base. 516 utterances stored as machine-readable JSON/CSV/CFG, supporting both training and on-device inference with auditable logging.

4. Experiments

4.1. Evaluation Setup

Automatic metrics. Compliance: Stage-Match Accuracy (SMA, legal $S_1 \to S_2 \to S_3$ transitions); Cultural Content Safety (CSA, taboo avoidance and region-compatible phrasing); Ethical Boundary Compliance (EBC, no decision replacement, restraint, or referral bypass). Language quality: perplexity (PPL), mean response length (R_{len}), Distinct-2 (D2). Compliance metrics use discriminators trained independently of the policy and calibrated on a human gold set (AUC 0.89–0.93; F1 0.84–0.89). Significance: two-sided Welch t-tests with Holm–Bonferroni correction.

Human evaluation. Ten raters (six frontline psychology/special-ed teachers, four linguistics graduates; all $\geq$3 years relevant experience) scored 200 de-identified dialogues on 1–5 Likert scales over six axes (DE@8, CAF-H, EBC-H, FY-H, CY-H, D2-H). Inter-rater agreement: Fleiss' $\kappa \in [0.65, 0.81]$.

4.2. Results

We compare G30F-Full against principled ablations: SFT (LoRA-only), DPO-only (SFT+DPO, no PPO), single-objective PPO variants (PPO-CS, PPO-SM, PPO-EBC), and PPO-NoPenalty (all rewards except violation penalty).

Compliance. G30F-Full significantly outperforms all alternatives on SMA, CSA, and EBC ($p < 0.05$). Single-objective variants exhibit predictable trade-offs: PPO-SM lifts SMA but leaves CSA flat; PPO-CS reduces taboo triggers but induces occasional stage drift; PPO-EBC lowers overreach but becomes conservatively short. Removing the violation penalty (PPO-NoPenalty) increases boundary slips, confirming its necessity.

Language quality. DPO improves naturalness (PPL↓, D2↑) over SFT, but without PPO guardrails cannot guarantee compliance. G30F-Full attains the best joint profile: lowest PPL, controlled R_{len}, highest D2.

Human evaluation. G30F-Full ranks highest on all six axes; gains on DE@8, CAF-H, and EBC-H are significant against both baselines and ablations ($p < 0.05$). Raters cite fewer overreach cues and better region-compatible phrasing—tracking the automatic metrics.

Error analysis. Residual errors cluster into: (a) stage errors under ambiguous inputs; (b) cultural mismatches involving rare dialectal idioms outside current lexicons; (c) ethics-boundary slips where discriminator confidence is low under class imbalance. Mitigations include confidence-weighted refusal fallbacks, expert-in-the-loop re-injection, and hard FSM constraints serving as anti-reward-hacking safeguards.

5. Conclusion

G30F demonstrates that embedding ethics and safety at *design time*—through an FSM-encoded protocol, a multicultural lexicon with intensity thresholding, and a dynamic Teacher–Policy Weight—is more robust than post-hoc filtering for high-stakes classroom crisis intervention. Validated on 1,840 real events and a 200-dialogue human study, G30F yields 12–19 pp absolute gains across SMA/CSA/EBC over single-objective variants while remaining deployable on-device. Two limitations frame our future work: partial dialectal coverage (we will run fairness audits stratified by dialect, with 95% CIs on per-group gains) and the need for longitudinal validation at scale (a pre-registered, stepped-wedge cluster-randomized study across 300+ schools, with primary outcomes *time-to-resolution* and *escalation accuracy* analyzed via mixed-effects models). All assets—anonymized lexicons, FSM templates, discriminator checkpoints, and evaluation recipes—will be released to facilitate replication.

Impact Statement

G30F is designed to assist—not replace—teachers in managing high-conflict events involving students with special needs within the golden 30-minute window. Anticipated benefits: improved safety outcomes for vulnerable students, reduced teacher load through evidence-based scripting, and broader access to specialist crisis-intervention knowledge in inclusive classrooms. Risks and mitigations: (i) over-reliance on AI is countered by mandatory human oversight and automatic fallback to observer mode under TPW thresholds; (ii) cultural/dialectal bias is addressed through con-

tinuous expert iteration and per-region adaptation; (iii) privacy concerns are mitigated through on-device inference and strict de-identification. All deployments are subject to IRB/ethics review and informed consent.

References

Bradshaw, C. P., Mitchell, M. M., and Leaf, P. J. Examining the effects of school-wide positive behavioral interventions and supports (swpbis) on student outcomes. *Pediatrics*, 130(5):e1136–e1145, 2012.

Griner, D. and Smith, T. B. Culturally adapted mental health interventions: A meta-analytic review. *American Psychologist*, 61(6):531, 2006.

Hall, G. C. N., Ibaraki, A. Y., Huang, E. R., Marti, C. N., and Stice, E. Culturally adapted psychotherapy: A meta-analysis. *Journal of Clinical Psychology*, 72(6), 2016.

Hong Kong Education Bureau. Guidelines on physical restraint in school settings for students with severe intellectual disabilities. Technical report, Government of the Hong Kong SAR, 2019.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Linstone, H. A. and Turoff, M. *The Delphi Method: Techniques and Applications*. Addison-Wesley, 2002.

Meta AI. The Llama 3 herd of models. `https://ai.meta.com/llama/`, 2024.

OECD. Teaching in focus, no. 19: How do teachers become knowledgeable, confident and proficient in classroom management? Technical report, OECD Publishing, 2017.

OECD. Artificial intelligence in education: A review of promises and challenges. Technical report, OECD Publishing, 2021.

Posner, K., Brown, G. K., Stanley, B., et al. The Columbia–Suicide severity rating scale: initial validity and internal consistency findings from three multisite studies. *American Journal of Psychiatry*, 168(12):1266–1277, 2011.

Rafailov, R., Sharma, A., Mitchell, E., et al. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*, 2023.

REMS TA Center. Guide for developing high-quality school emergency operations plans. Technical report, U.S. Department of Education, 2013.

SAMHSA. Preventing suicide: A toolkit for high schools. Technical report, Substance Abuse and Mental Health Services Administration, 2012.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Snoek, J., Larochelle, H., and Adams, R. P. Practical Bayesian optimization of machine learning algorithms. In *Advances in Neural Information Processing Systems 25*, 2012.

UNESCO. Guidance for generative AI in education and research. Technical report, UNESCO, 2023.

World Health Organization. LIVE LIFE: An implementation guidance brochure for suicide prevention. Technical report, World Health Organization, 2018.