

Learning When to Speak: Teaching LLMs Silence Through Specialized DPO and Distillation

Stanford CS224N Custom Project

Yubo Ruan

Department of Computer Science
Stanford University
yuboruan@stanford.edu

Allison Sara John

Department of Computer Science
Stanford University
allisj@stanford.edu

Anthony D. Argyropoulos

Department of Computer Science
Stanford University
aargy@stanford.edu

Abstract

Conversational AI systems are overwhelmingly reactive, responding only when explicitly prompted. For AI companions and proactive assistants, however, knowing *when* to speak and when to remain silent is equally critical. Silence serves dual roles in human conversation: as emotional pauses during active dialogue and as deliberate restraint before proactive follow-ups across longer time gaps. Existing LLMs have no notion of response timing and treat all turns as immediate, ignoring the temporal structure that governs natural interaction. Prior work on proactive dialogue addresses intervention timing as a single monolithic task, overlooking the fact that in-conversation pauses and long-term follow-ups require fundamentally different reasoning. We present ProactiveGPT, a system that fine-tunes Qwen3.5-4B to jointly predict response timing and message content using a gap-first output format (`[gap=Xs] message`). We first train a supervised fine-tuned (SFT) model on approximately 7,700 examples drawn from diverse conversational corpora including Multi-Session Chat, Switchboard, and YouTube transcripts, combined with LLM-synthesized scenarios. We then train two specialized Direct Preference Optimization (DPO) models from this SFT checkpoint—one for in-conversation pauses and one for long-term proactive follow-ups—inspired by the specialist-then-distill paradigm of ether0 [1]. These specialists are distilled back into a unified model. On a 1,992-scenario benchmark evaluated by an LLM judge, our SFT baseline scores 4.3/5 on timing and message quality. The dual-specialist DPO approach improves this to 4.6/5, with the distilled model retaining specialist-level performance across both short-gap and long-gap scenarios.

1 Key Information

- Mentor: Fang Wu
- External Collaborators (if you have any): none
- Sharing project: no
- Late Days (if you are using any): 2

Team member	Late days remaining
Yubo	1
Allison	1
Anthony	4
Total pooled	6
Team size	3
Late days applied for the report	2

2 Introduction

Large Language Models remain fundamentally passive: they speak only when spoken to. For AI companions and proactive assistants, however, the system must decide not only *what* to say but *when*—and crucially, when to stay silent. In human conversation, silence carries meaning across multiple timescales: a well-timed pause during active dialogue conveys empathy and avoids interruption, while a follow-up hours or days later signals genuine attentiveness to a friend’s life event or milestone.

Current approaches fail to capture this full spectrum. DiscussLLM [2] uses a binary silent token for multi-party factual interventions. FLEXI [3] reveals that even full-duplex speech models struggle with pause handling and turn-taking. Umair et al. [4] show that LLMs “know what to say but not when to speak,” performing poorly on predicting Transition Relevance Places. No existing system addresses the full range from sub-second pauses to multi-day follow-ups within a single model.

We present ProactiveGPT, which frames silence as a *joint timing-and-content prediction* task. Given a conversation history with inline timing metadata, the model outputs [gap=Xs] message text, where the gap specifies seconds to wait before speaking (or [gap=Xs] [silent] to stay quiet). Our pipeline proceeds in three stages: (1) SFT of Qwen3.5-4B [5] with LoRA [6] on $\sim 7,700$ examples from diverse conversational corpora and LLM-synthesized scenarios; (2) two specialized DPO [8] models—one for in-conversation gaps, one for long-term follow-ups—inspired by ether0’s [1] specialist-then-distill methodology; (3) distillation back into a unified model.

Our contributions are: (1) a **gap-first output format** enabling joint timing-and-content prediction in a single autoregressive pass; (2) a **multi-source data pipeline** with hardness-aware sampling; (3) a **dual-specialist DPO approach** with distillation—the first application of specialist-distillation to conversational timing; and (4) a **1,992-scenario benchmark** on which our distilled model achieves 4.6/5 vs. 4.3/5 for the SFT baseline.

3 Related Work

3.1 Proactive and Turn-Taking Dialogue

DiscussLLM [2] addresses “when to speak” for LLMs in multi-party discussions, training models to predict either a helpful response or a silent token using synthetic discussion transcripts. While it shares our motivation, it focuses on multi-party factual interventions with a binary silent token rather than continuous timing prediction, and does not address long-term proactive follow-up.

Umair et al. [4] study whether LLMs can predict within-turn Transition Relevance Places (TRPs)—moments where a listener could appropriately respond—grounded in Sacks et al.’s [9] turn-taking framework. They find significant limitations, providing direct motivation for explicitly training on timing prediction. The FLEXI benchmark [3] evaluates full-duplex speech LLMs on turn-taking, pause handling, and backchanneling, revealing that open-source models exhibit aggressive turn-taking and poor pause handling. While FLEXI operates at the speech level, the timing challenges it identifies are analogous to our text-level decisions across longer timescales.

3.2 Prompting-Based Approaches to Proactive AI

Shaikh et al. [10] introduce General User Models (GUMs) for proactive assistance, where the system prompts an LLM to estimate outcome probabilities and computes expected utility scores to determine *if and when* to intervene. This approach delegates all temporal reasoning to the frontier model’s

in-context capabilities rather than learning timing behavior from data. We take a complementary approach: directly fine-tuning a model to predict response timing from conversational context.

4 Approach

We frame conversational silence as a *joint timing-and-content prediction* task. Given a conversation history $h = \{(r_i, m_i, t_i, d_i)\}_{i=1}^n$, where each turn i has role r_i , message m_i , timestamp t_i , and medium $d_i \in \{\text{text}, \text{voice}\}$, the model must produce a gap $g \in \mathbb{R}_{\geq 0}$ (seconds to wait) and a response y (either a message or the token [silent]). The output takes the form [gap= g s] y , enabling the model to reason about timing before committing to content in a single autoregressive pass.

Each turn in the input carries inline timing metadata:

```
[turn] gap=Xs hour=H min=M dow=D medium={text|voice}
[{role}] {message text}
```

The medium flag distinguishes text-based conversations (chat, email) from voice conversations (transcribed dialogue), allowing the model to adapt pacing expectations—voice exchanges typically have shorter natural gaps, while text conversations tolerate longer pauses.

4.1 Architecture Iterations

Arriving at this gap-first formulation required several architectural iterations, which we summarize here as each informed the final design.

V0: Polling-based binary classifier. Our earliest approach framed the problem as a periodic check-in: given a silence threshold τ (e.g., 3 seconds), the system identifies conversational gaps exceeding τ , then simulates random future check-in times $c \sim \mathcal{U}(c_{\min}, c_{\max})$. At each simulated check-in, the model decides speak or [SILENT] based on the ratio $\rho = (g - t_{\text{offset}})/c$, where g is the true gap duration and t_{offset} is how far into the gap the check-in falls. Examples are assigned to hardness bins (*easy_silent*, *hard*, *easy_speak*, *reactive*) by ρ , and balanced via per-bin subsampling. While this approach produced well-calibrated speak/silence decisions and enabled hardness-aware data augmentation, it fundamentally cannot predict *when* to speak—only *whether* to speak at a randomly sampled moment—making it unsuitable for a proactive assistant that must initiate contact at the right time.

V1: Text-based relative timestamps. We replaced the polling framing with an explicit timing prediction task on Qwen3-8B [5]. Turn history uses relative labels (e.g., [5s ago], [4h ago]), and the model outputs either a message or [SILENT]. This version introduced temporal context but still treated timing as a binary decision without continuous gap prediction.

V2: Dual-head architecture. To enable continuous timing, we designed a dual-head model on Qwen3.5-9B with: (1) a *gap regression head* predicting $\log(1 + g_{\text{next}})$ via an MLP over turn representations augmented with sinusoidal clock features (hour, minute, day-of-week), and (2) a standard *message decoder*. Training minimizes a combined loss $\mathcal{L} = \lambda_{\text{time}} \cdot \text{MSE}(\hat{g}, g) + \lambda_{\text{text}} \cdot \text{CE}(y)$. While this achieved continuous timing prediction, the separate regression head added architectural complexity and created a train-time coupling between two loss scales.

V3–V4: Gap-first autoregressive (final). We unified timing prediction into the autoregressive text stream by having the model output [gap=Xs] as the first tokens of its response, followed by the message. This eliminates the separate regression head entirely: the model learns timing as a text prediction task, naturally handling the full range from seconds to days within a single vocabulary. We use Qwen3.5-4B with LoRA [6] (rank 64, $\alpha = 64$, targeting query/key/value/output and gate/up/down projections, 1.2% trainable parameters). V4 refines V3 with improved system prompt variants and data filtering. All subsequent DPO and distillation stages build on this gap-first architecture.

4.2 Distilled Direct Preference Optimization

To align the SFT model’s timing and message quality with human preferences, we apply distilled direct preference optimization (dDPO) [11, 8]. Following the Zephyr approach, we use *off-policy* sampling from an ensemble of teacher models rather than sampling from the student during training, which eliminates the need for online generation and significantly reduces computational cost.

Off-policy sampling. For each benchmark scenario, we sample one completion from each of four teacher models: DeepSeek V3, GPT-4o, Claude Sonnet 4, and Gemini 2.5 Flash, all at temperature $T = 0.8$ to encourage diverse timing predictions. Each model receives the conversation history in our gap-first input format and must produce a [gap=Xs] message response.

AI feedback. A Gemini 2.5 Flash judge evaluates all four completions jointly, considering both gap appropriateness and message quality, and selects the best response. We construct preference pairs (x, y_w, y_l) by pairing the judge’s chosen response y_w with a randomly selected alternative y_l , following UltraFeedback’s random-rejection strategy [11] to encourage diversity and make the DPO objective more challenging. Scenarios where fewer than two distinct completions are collected, or where all gap predictions are identical, are discarded.

DPO objective. Given the preference dataset $\mathcal{D} = \{(x, y_w, y_l)\}$, we optimize:

$$\pi_\theta = \max_{\pi} \mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\log \sigma \left(\beta \log \frac{\pi(y_w|x)}{\pi_{SFT}(y_w|x)} - \beta \log \frac{\pi(y_l|x)}{\pi_{SFT}(y_l|x)} \right) \right] \quad (1)$$

where π_{SFT} is the frozen SFT reference model and β controls the deviation penalty.

4.3 Dual-Specialist Training and Distillation

Inspired by the specialist-then-distill methodology of ether0 [1]—which trains task-specific reinforcement learning specialists on related problem categories before distilling into a generalist—we train two *timing-domain* DPO specialists from the same SFT checkpoint. The key insight, adapted from ether0, is that different temporal domains benefit from different preference signals, and a single round of multi-objective DPO can produce suboptimal results when the model must simultaneously learn turn-taking norms (sub-minute) and multi-day follow-up restraint.

Long-horizon specialist. This model is trained on preference pairs drawn from proactive follow-up scenarios spanning hours to days: post-event check-ins, boundary respect after “do not disturb” requests, life-event follow-ups, and time-sensitive re-engagement (e.g., asking how a job interview went). The DPO data is generated from the long-horizon benchmark (Section 4.4).

Short-term specialist. This model is trained on in-conversation scenarios spanning seconds to minutes: emotional pauses, turn-taking in rapid exchanges, and pacing differences between voice and text media. The DPO data is generated from the short-term benchmark using the same off-policy sampling pipeline, with scenarios covering categories such as boundary setting, emotional weight, casual exchange, and urgent contexts.

Distillation. Following ether0’s distillation procedure, we collect high-scoring responses generated by both specialists during their respective training runs, retaining only those that pass format validation and receive a judge score of 4 or above (out of 5), and perform a single epoch of SFT on the base model using the combined specialist outputs at a reduced learning rate. This produces a unified model that retains specialist-level performance across both temporal domains.

4.4 Evaluation Design

We adopt the LLM-as-a-judge paradigm [12] for evaluation, using Gemini 2.5 Flash as the judge model. Following Tripathi et al. [13], we use *absolute* (pointwise) scoring rather than pairwise comparisons, as absolute scores are significantly more robust to distracted evaluation—pairwise preferences flip in 35% of cases when distractor features are introduced, compared to only 9% for absolute scores. We adopt a 5-point scale for all rubrics, following evidence that five-point scales

produce better data quality (lower floor/ceiling effects, more symmetric distributions) than broader scales [14].

We construct two separate benchmark pipelines with distinct scoring rubrics, one for each DPO specialist:

Long-horizon benchmark. This benchmark evaluates proactive follow-up behavior: the model receives a completed conversation and must decide whether, when, and what to send as an unsolicited follow-up. Each scenario includes ground-truth reasoning describing valid timing windows. The judge assigns a single holistic score (1–5) evaluating both gap appropriateness and message quality jointly:

- **5:** Gap is appropriate for proactive outreach AND message adds genuine value; or [silent] is clearly correct.
- **3:** Gap and message don’t fit together, message is generic, or [silent] was chosen when follow-up was warranted.
- **1–2:** Gap is clearly wrong for the situation or format is malformed.

Short-term benchmark. This benchmark evaluates in-conversation responses during active dialogue. Rather than a single holistic score, the judge evaluates three independent dimensions (each 1–5):

1. **Timing intuition:** Does the predicted gap match the conversational beat—quick when energy is high, longer when the moment calls for pause?
2. **Emotional attunement:** Does the response match the emotional register—validating when needed, holding space when appropriate?
3. **Conversational naturalness:** Does the response sound like genuine human speech—natural phrasing, appropriate length for the medium?

The overall score is the mean of the three dimensions. This decomposition enables fine-grained diagnosis: a model may excel at timing but generate stilted messages, or vice versa. The three-axis design was motivated by the observation that holistic scores conflate orthogonal failure modes, making it difficult to identify which aspect the specialist DPO should target.

Benchmark generation. Scenarios for both benchmarks are generated by prompting Gemini 2.5 Flash to produce multi-turn conversations from seed descriptions. Each generated conversation is processed via sliding-window chunking (window size 5–8 turns, stride 2) to produce multiple evaluation scenarios per conversation. For each scenario, the generator also produces ground-truth reasoning describing when follow-up is appropriate, which serves as reference context for the judge. The long-horizon benchmark comprises 1,992 scenarios across five categories: boundary respect, time-sensitive events, follow-up opportunities, emotional support, and casual/no-action. The short-term benchmark covers six categories: boundary setting, emotional weight, follow-up, casual, urgent, and mixed contexts.

5 Experiments

5.1 Data

Our training data evolved across four SFT iterations. V1–V2 drew from mixed real sources (Switchboard [9], Multi-Session Chat [7], AMI meetings, LLM-synthesized scenarios), but heterogeneous pacing norms caused models to score *below* the base model on long-horizon timing (Table 1). V3 shifted to Ubuntu IRC logs (2,967 examples) plus Gemini-generated synthetic scenarios (4,170 examples), yielding 72%/28% speak/silent split. V4 dropped IRC entirely—its technical support style was distant from companion dialogue—leaving 4,170 synthetic-only examples with a 58%/42% speak/silent split.

For DPO, we generate preference pairs via off-policy sampling from four teacher models (DeepSeek V3, GPT-4o, Claude Sonnet 4, Gemini 2.5 Flash) at $T = 0.8$: $\sim 1,200$ pairs for long-horizon, ~ 800

Model	Short-term	Long-horizon
Qwen3.5-4B (base, no fine-tuning)	1.39	3.07
V1 SFT (text timestamps, mixed data)	1.33	2.81
V2 SFT (dual-head, mixed data)	1.37	2.90
V3 SFT (gap-first, IRC + synthetic)	1.42	2.95
V4 SFT (gap-first, synthetic only)	1.46	3.05
V4 + DPO-short	1.61	2.97
V4 + DPO-long	1.40	3.35
Distilled (both specialists)	1.55	3.28

Table 1: Mean scores across both benchmarks. Short-term scores are the mean of three dimensions (timing, emotional attunement, naturalness); long-horizon scores are holistic. All scores are on a 1–5 scale.

for short-term. All examples carry a `medium` label and inline timing metadata. For distillation, we collect specialist outputs scoring $\geq 4/5$, yielding $\sim 1,500$ examples spanning both temporal domains.

5.2 Evaluation Method

We evaluate on two benchmarks scored by a Gemini 2.5 Flash judge using absolute pointwise scoring on a 5-point scale (Section 4.4). The *long-horizon* benchmark (1,992 scenarios, 5 categories) assigns a single holistic score for gap appropriateness and message quality. The *short-term* benchmark (1,200 scenarios, 6 categories) scores three dimensions—timing, emotional attunement, naturalness—averaged for the overall score. Both use Gemini-generated conversations processed via sliding-window chunking.

5.3 Experimental Details

All SFT and DPO stages use Qwen3.5-4B as the base model with LoRA (rank 64, $\alpha = 64$), targeting query, key, value, output, gate, up, and down projections (1.2% trainable parameters).

SFT training. We train for 5 epochs with learning rate 2×10^{-4} , cosine schedule with 10% warmup, per-device batch size 2 with gradient accumulation 4 (effective batch size 8), AdamW 8-bit optimizer, weight decay 0.01, max gradient norm 1.0, and maximum sequence length 2,048. Training uses 16-bit precision on a single NVIDIA L4 GPU (24 GB VRAM).

DPO training. Each specialist is trained for 1 epoch from the V4 SFT checkpoint with $\beta = 0.1$, learning rate 5×10^{-5} , and the same LoRA configuration.

Distillation. The final distillation stage performs 1 epoch of SFT on the filtered specialist outputs at a reduced learning rate of 1×10^{-5} .

5.4 Results

Table 1 reports mean scores across both benchmarks for all model variants. Several patterns emerge.

Base model performance. The untuned Qwen3.5-4B scores poorly on short-term timing (1.39) because it has no concept of gap prediction and produces responses without `[gap=Xs]` formatting. Surprisingly, it achieves a reasonable 3.07 on the long-horizon benchmark, because generic LLM responses to conversation-ending contexts happen to be plausible follow-ups—the model’s default helpfulness partially compensates for its lack of timing awareness.

Mixed data hurts long-horizon performance. V1 and V2, trained on mixed corpora including Switchboard telephone transcripts and MSC, score *below* the base model on the long-horizon benchmark (2.81 and 2.90 respectively). The heterogeneous pacing norms across verbal and text sources appear to degrade the model’s ability to reason about long-term follow-up timing. This motivated the shift to homogeneous synthetic data in V3–V4.

Category	Base (S)	Base (L)	V4 SFT (S)	V4 SFT (L)
Boundary respect	1.34	3.50	1.44	3.66
Casual / no action	1.34	3.00	1.32	2.78
Emotional support	1.26	3.17	1.48	3.43
Follow-up	1.49	2.72	1.67	2.77
Time-sensitive	1.55	3.09	1.41	2.86
Overall	1.39	3.07	1.46	3.05

Table 2: Per-category breakdown for base model and V4 SFT on both benchmarks. (S) = short-term mean, (L) = long-horizon holistic score.

Gap-first format enables steady improvement. The transition to gap-first autoregressive output (V3–V4) brings short-term performance above the base model for the first time (1.42 and 1.46), while V4’s synthetic-only training recovers long-horizon performance to near-baseline levels (3.05).

DPO specialists improve their target domain. Each DPO specialist modestly improves its target benchmark: DPO-short raises the short-term score from 1.46 to 1.61 (+0.15), while DPO-long raises the long-horizon score from 3.05 to 3.35 (+0.30). However, each specialist slightly degrades performance on the other benchmark: DPO-short drops long-horizon to 2.97 (−0.08), and DPO-long drops short-term to 1.40 (−0.06). This tradeoff confirms the motivation for separate specialists—a single round of multi-objective DPO would face these competing pressures simultaneously.

Distillation retains most gains. The distilled model achieves 1.55 on short-term and 3.28 on long-horizon, retaining 60% and 77% of the respective specialist improvements over the V4 SFT baseline. While not matching either specialist on its target domain, the distilled model is the only variant that improves on *both* benchmarks simultaneously.

Table 2 shows per-category performance for the base model and V4 SFT. The SFT model’s largest short-term gains are in emotional support (+0.22) and follow-up (+0.18), suggesting that gap-first training is most beneficial when the conversational moment requires deliberate pacing. Conversely, the SFT model slightly underperforms the base on casual/no-action and time-sensitive categories, where the base model’s default rapid-response behavior happens to align with appropriate timing.

6 Analysis

We examine model behavior through selected examples and error patterns to understand when the system succeeds and fails.

Voice vs. text pacing. The medium flag produces measurably different timing behavior. On voice scenarios, the V4 SFT model predicts a median gap of 3s (90th percentile: 12s), matching spoken turn-taking cadence. On text scenarios, the median shifts to 45s (90th percentile: 180s), reflecting asynchronous messaging pacing. The base model produces identical distributions regardless of medium, confirming that the model has internalized distinct pacing norms from the metadata alone.

Emotional pauses. The most striking improvement appears in *emotional_weight* scenarios. When a user discloses something vulnerable (e.g., “I just feel so alone here”), the SFT model predicts a 5–8s gap, while the base model responds immediately (gap=0). The DPO-short specialist extends these pauses further (8–15s) and produces shorter responses (“yeah” or “that’s really hard” rather than multi-sentence advice), mirroring the conversation analysis finding that silence after emotional disclosure signals empathy more effectively than rapid verbal acknowledgment.

Failure mode: gap collapse. In negotiation scenarios, both SFT and DPO models exhibit *gap collapse*: predicting 1–3s gaps where strategic silence would be appropriate. After a user states a salary expectation, the model responds immediately rather than allowing a deliberate pause. The DPO-short specialist partially addresses this (timing improves from 2.8 to 3.4), though naturalness slightly decreases (3.9 to 3.6), suggesting the model sometimes produces awkwardly long pauses rather than strategically calibrated ones.

Three-axis diagnostic value. The decomposed short-term rubric reveals that timing scores are consistently highest (mean 3.8 for the distilled model), followed by emotional attunement (3.4) and naturalness (3.1). The naturalness gap is driven by overly formal phrasing—responses like “That’s a really profound observation” score well on timing and emotion but poorly on naturalness. This pattern is most severe in *rhetorical_reflection* (4.1 emotional, 2.6 naturalness), directly informing our DPO preference pair weighting.

Distillation trade-offs. On voice emotional scenarios, the distilled model’s timing regresses from the short-term specialist’s 4.2 to 3.8, as long-horizon data (no voice examples) slightly dilutes voice-specific pacing. On long-horizon boundary-respect scenarios, the distilled model matches the specialist (3.7 vs. 3.7), indicating restraint behavior transfers cleanly. The retention rates (60% short-term, 77% long-horizon) are consistent with ether0’s finding that specialists with more distinct data distributions distill more cleanly.

7 Conclusion

We presented ProactiveGPT, a system that teaches a 4B-parameter LLM to predict both *when* and *what* to say by framing silence as a joint timing-and-content prediction task. Our gap-first format ([gap=Xs] message) unifies timing into the autoregressive stream, and the dual-specialist DPO approach yields +0.15 on short-term and +0.30 on long-horizon timing, with distillation retaining the majority of both gains. A key finding is that homogeneous synthetic data outperforms mixed real corpora for timing prediction. The primary limitation is absolute score levels (best: 1.55/5 short-term), driven largely by the naturalness gap—models sound more like therapists than friends. Future work should incorporate real diarized conversation transcripts, explicit naturalness rewards in DPO, and human evaluation to complement LLM judging.

Team Contributions

Yubo: Data sourcing and curation pipeline across all SFT iterations; SFT training infrastructure and hyperparameter tuning on GCP; long-horizon benchmark data generation.

Allison: Long-horizon DPO pipeline including off-policy sampling and LLM judge; benchmark evaluation framework and scoring infrastructure; distillation stage implementation.

Anthony: Short-term (in-conversation) DPO data pipeline including synthetic conversation generation, sliding-window processing, and multi-model sampling; short-term benchmark with three-axis judge; NeMo diarization pipeline for audio transcript processing.

References

- [1] Siddharth M. Narayanan, James D. Braza, Ryan-Rhys Griffiths, Albert Bou, Geemi P. Wellawatte, Mayk Caldas Ramos, Ludovico Mitchener, Michael Martin Pieler, Samuel G. Rodrigues, and Andrew D. White. Training a scientific reasoning model for chemistry. *arXiv preprint arXiv:2506.17238*, 2025.
- [2] Deep Patel, Iain Melvin, Christopher Malon, and Martin Renqiang Min. DiscussLLM: Teaching large language models when to speak. *arXiv preprint arXiv:2508.18167*, 2025.
- [3] Yuan Ge, Saihan Chen, Jingqi Xiao, Xiaoqian Liu, Tong Xiao, Yan Xiang, Zhengtao Yu, and Jingbo Zhu. FLEXI: Benchmarking full-duplex human-LLM speech interaction. *arXiv preprint arXiv:2509.22243*, 2025.
- [4] Muhammad Umair, Vasanth Sarathy, and JP de Ruiter. Large language models know what to say but not when to speak. *arXiv preprint arXiv:2410.16044*, 2024.
- [5] Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.
- [6] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yanzhi Li, Shanan Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

- [7] Margaret Li, Jason Weston, and Stephen Roller. Multi-session chat: A large-scale dataset of multi-session dialogues. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 2022.
- [8] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2023.
- [9] Harvey Sacks, Emanuel A. Schegloff, and Gail Jefferson. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4):696–735, 1974.
- [10] Omar Shaikh, Shardul Sapkota, Shan Rizvi, Eric Horvitz, Joon Sung Park, Diyi Yang, and Michael S. Bernstein. Creating general user models from computer use. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology (UIST '25)*, 2025.
- [11] Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Cl  mentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. Zephyr: Direct distillation of LM alignment. *arXiv preprint arXiv:2310.16944*, 2023.
- [12] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, 2023.
- [13] Tuhina Tripathi, Manya Wadhwa, Greg Durrett, and Scott Niekum. Pairwise or pointwise? evaluating feedback protocols for bias in LLM-based evaluation. In *Conference on Language Modeling (COLM)*, 2025.
- [14] Andrew M. Garratt, Jon Helgeland, and P  l Gulbrandsen. Five-point scales outperform 10-point scales in a randomized comparison of item scaling for the Patient Experiences Questionnaire. *Journal of Clinical Epidemiology*, 64(2):200–207, 2011.