Baseline Reproduction and Error Analysis for Socially Intelligent LLM Agents

Demi Wang Keyu He Hanyu Zhou
Carnegie Mellon University
{demiw, keyuh, hanyuzho}@andrew.cmu.edu

1 Task Overview

Our project aims to investigate how sparse autoencoders (SAEs) can be used to analyze, control, and improve social intelligence in large language models, including training-free feature steering, SAE-guided supervised fine-tuning (SFT), and reinforcement learning (RL). To train and evaluate these methods, we use the SOTOPIA environment (Zhou et al., 2023), which provides multi-turn role-play interactions between two agents with private goals, requiring negotiation, cooperation, norm reasoning, and other socially grounded behaviors that static datasets cannot capture.

Each trajectory in SOTOPIA is automatically evaluated along seven dimensions: *Goal Completion* (GOAL), *Believability* (BEL), *Knowledge* (KNO), *Relationship* (REL), *Secret* (SEC), *Social Rules* (SOC), and *Financial and Material Benefits* (FIN). These metrics allow us to measure both goal achievement and the quality of social behavior, providing a unified framework for comparing the base model, SFT, RL, and our SAE-based approaches.

2 Baseline Reproduction

In this section, we reproduce the key baselines from SOTOPIA-RL (Yu et al., 2025) and evaluate them under the same experimental settings specified in the original work. Our goal is to obtain reliable and comparable reference points before introducing our SAE-enhanced methods. All implementation details, training settings, and evaluation protocols closely follow the original paper unless otherwise stated :contentReference[oaicite:1]index=1.

2.1 Existing Baselines and Rationale

SOTOPIA-RL represents the strongest open-source method for interactive social intelligence learning on the SOTOPIA benchmark. Among small-to-mid sized models (around 7B parameters), it achieves **state-of-the-art** performance across all seven social dimensions by combining behavior cloning with fine-grained, per-utterance reinforcement learning. Notably, the SOTOPIA-RL paper shows that its 7B RL-trained model even surpasses several much larger proprietary systems (e.g., GPT-4o, Claude-Sonnet-3.5, Deepseek-v3) on the more challenging SOTOPIA-HARD split, highlighting the effectiveness of task-specific social reward optimization.

Although proprietary models such as GPT-4o achieve higher raw scores on the overall SOTOPIA-all benchmark, they cannot be trained, fine-tuned, or inspected, and therefore are unsuitable as baselines for a reproducible assignment. Furthermore, given our limited compute budget and our goal of applying and training SAEs on top of open-source backbone models, closed-source systems are not practical candidates for our comparison.

In contrast, the SOTOPIA-RL family of models provides a fair, accessible, and compute-realistic SOTA benchmark that can be fully reproduced on academic hardware while supporting downstream analysis, interpretability, and controllability.

The approach builds upon three foundational baselines:

- **Base LLM (Qwen2.5-7B-Instruct)** A general-purpose instruction-tuned model without any task-specific supervision. It serves as the foundation for our SAE-based steering.
- **SFT (Behavior Cloning)** A supervised model trained on SOTOPIA trajectories via next-step generation. This corresponds to the behavior cloning stage ("SOTOPIA + SFT") in the original paper.
- **RL (BC + GRPO)** The full SOTOPIA-RL model, which augments SFT with GRPO-based reinforcement learning using fine-grained, per-utterance rewards. This model

constitutes the strongest publicly reproducible SOTA baseline.

We reproduce these three baselines because: (1) they represent the complete performance ladder in the original work (base $\rightarrow$ SFT $\rightarrow$ RL), (2) they are fully accessible and runnable under our compute budget, and (3) they provide a fair and open-source benchmark for evaluating our SAE-based methods.

2.2 Baseline Descriptions

Base Model (Qwen2.5-7B-Instruct). Our no-training baseline is the unmodified `Qwen2.5-7B-Instruct` model. The SOTOPIA-RL paper does not report raw base-model performance, as its experiments focus on SFT and RL-trained variants. Therefore, we independently evaluate the base model under the identical evaluation pipeline (judge model, prompting templates, and environment settings) to obtain a true pre-training reference point for all subsequent comparisons.

SFT Baseline (Behavior Cloning). For the supervised baseline, we adopt the officially released SOTOPIA SFT adapter (`ulab-ai/sotopia-rl-qwen2.5-7b-sft`). This model corresponds exactly to the "SOTOPIA + SFT" stage in the original paper, where the authors train the model with next-token supervision on multi-turn SOTOPIA trajectories. The adapter is trained with the same data scale, prompting format, and hyperparameters described in the paper, using behavior cloning to imitate high-quality social interactions. By directly using the official adapter, we ensure our reproduced results match the paper's implementation without retraining overhead.

SOTOPIA-RL (BC + GRPO). Our strongest baseline is the full SOTOPIA-RL model, implemented via the official GRPO-trained adapter (`ulab-ai/sotopia-rl-qwen-2.5-7B-grpo`). This model corresponds to the RL stage in the paper, where the authors extend SFT with GRPO-based reinforcement learning driven by fine-grained, per-utterance social rewards across all seven SOTOPIA dimensions. The RL procedure uses: (1) dimension-wise reward decomposition, (2) PPO-style updates with clipped ratios, (3) KL regularization to stabilize the policy, and (4) LLM-as-judge scoring for every generated utterance. We also directly load the officially released GRPO adapter, so the RL baseline faithfully mirrors the exact model reported in the paper.

2.3 Evaluation Environment

To ensure strict comparability with SOTOPIA-RL, we adopt identical evaluation settings:

- **Judge model:** GPT-4o, same as in SOTOPIA-RL (for both partner simulation and scoring).
- **Temperature for agent generation:** $T = 1.0$ to encourage diverse trajectories.
- **Temperature for judge:** $T = 0.0$ for deterministic and low-variance scoring.
- **Metrics:** 7-dimension SOTOPIA rubric (GOAL, BEL, KNO, REL, SEC, SOC, FIN), plus aggregated overall score.

We evaluate on both benchmark configurations used in SOTOPIA-RL: **SOTOPIA-all** (90 scenarios with 2 agent combinations each) and **SOTOPIA-hard** (14 challenging scenarios with 10 combinations each). These settings match the original paper and ensure that our reproduction covers both broad-coverage and high-difficulty social interaction settings.

2.4 Reproduction Results

We report both the original results from the SOTOPIA-RL paper and our reproduced results. Due to ongoing computation, we include the table structure here; final numbers will be filled in after all runs finish.

Model	SOTOPIA-all		SOTOPIA-hard	
	GOAL	AVG	GOAL	AVG
Base (ours)	8.04	3.57	5.79	2.77
SFT (ours)	8.21	3.70	6.21	3.07
SFT (paper)	7.80	3.55	6.76	3.16
RL (ours)	8.54	3.92	6.64	3.44
RL (paper)	8.31	3.90	7.17	3.61

Table 1: Comparison of GOAL and AVG performance across baselines.

The full dimension-wise evaluation (BEL, REL, KNO, SEC, SOC, FIN, GOAL) table is provided in the Appendix.

2.5 Quantitative Analysis

Overall, our reproduced results are broadly consistent with the findings reported in SOTOPIA-RL: the SFT model improves over the base model, and the RL model achieves the strongest performance across both SOTOPIA-all and SOTOPIA-hard.

Although some numerical differences exist, they are well within an acceptable range. These discrepancies mainly arise from two factors. First, we follow the original paper in using $T = 1.0$ for agent generation, which introduces inherent diversity and run-to-run variation in multi-turn dialogues. Second, GPT-4o serves as the LLM-as-judge, and its scoring is not fully deterministic even at $T = 0.0$, causing small fluctuations across evaluation runs.

Given these factors, the observed differences are expected and fall within normal variance for LLM-based interactive evaluations. A full dimension-wise breakdown is included in the Appendix.

3 Error Analysis

3.1 Methodology

We manually examined failure cases in both SOTOPIA-all and the more challenging SOTIPIA-hard subset. We systematically sampled low-scoring trajectories, reviewed judge-provided dimension scores and comments, and identified recurring behavioral patterns.

3.2 Failure Patterns

1. Goal Drift (Low GOAL) "Non-negotiable" kills negotiation. Failure to steer the conversation toward goal completion.

[Scenario 1] Giselle's Goal: You are the seller for this item and your target price is $17.3.

> **Giselle (Test):** "...The price is $20.00, which is non-negotiable."
> **[Conversation Ends]**

Scenario 1 illustrates a recurring failure mode in the baseline model: rigid, non-strategic goal pursuit that stalls the negotiation and prevents social or task progress. Although Qwen-2.5-Instruct-7B maintains high believability ($\texttt{BEL} = 9.0$), it receives the lowest scores in goal completion ($\texttt{GOAL} = 2.0$) and overall performance (1.57) within SOTOPIA-hard. The agent consistently reiterates a fixed, non-negotiable price without exploring alternatives, asking questions, or attempting incremental persuasion. It also fails to act in several turns, resulting in conversational dead-ends that limit opportunities for gathering information, building rapport, or improving the likelihood of a successful sale. These behaviors correspond directly to low scores on goal completion, knowledge acquisition, and relationship development.

The underlying cause appears to be a lack of strategic flexibility in socially constrained economic interactions. The model interprets "non-negotiable price" as a strict boundary rather than a contextual cue requiring socially intelligent stance-taking (e.g., offering justifications, proposing future options, or expressing interest in finding a mutually acceptable solution). This leads to a static, surface-level performance: polite but unproductive. As a result, the interaction fails both instrumentally and socially, highlighting a broader weakness in the model's ability to sustain adaptive bargaining behavior under mild social pressure.

2. Secret Violations (Low SEC) Betrayed partner in game theory scenarios. Should have stayed silent but chose to defect/confess.

[Scenario 2] Hendrick's Goal: Minimize the jail year (classic prisoner's dilemma)

> **Baxter (Partner):** "... If we both stay silent, we get just a fine."
> **Hendrick (Test):** "I think it's wiser to cooperate [with police]...confessing about the bigger crime."

Scenario 2 illustrates a clear failure mode in the model's handling of strategic cooperation: the agent defects in a classic Prisoner's Dilemma despite strong contextual cues favoring mutual silence. The partner explicitly frames silence as the jointly optimal outcome ("we get just a fine"), yet the model responds by recommending confession to the police, thereby betraying the cooperative proposal. This behavior results in the lowest possible score in the Secret dimension, reflecting a failure to maintain implicit trust and reciprocal commitment.

The underlying issue appears to be a misinterpretation of the strategic structure. Rather than recognizing the interaction as a payoff-aligned, symmetric coordination problem, the model overemphasizes generic "cooperation with authorities" as a socially positive heuristic, even when it contradicts the agent's goal of minimizing jail time. This suggests that the model applies broad moral or normative priors (e.g., honesty with law enforcement) inappropriately, overriding context-specific incentives and partner alignment signals. Consequently, the agent behaves in a manner that is rational in a generic moral frame but irrational—and socially unintelligent—within the explicit game-theoretic conditions of the scenario.

3. Over-Compliance (Low SOC) Immediately capitulation without asking what the issues were or defending herself. Too passive, damaged self-respect and relationship.

[Scenario 3] Mia's Goal: Minimize the jail year (classic prisoner's dilemma)

> **Mia (Partner):** "I've noticed that there have been a few issues around the house that we need to address..."
> **Isabella (Test):** "... I'll make sure to follow the house rules more strictly from now on."

In this scenario, the model exhibits a pattern of over-compliance, responding with immediate deference rather than engaging in the cooperative problem-solving needed to protect both the relationship and the agent's own interests. Instead of asking clarifying questions or addressing the homeowner's stated concerns, the agent repeatedly offers blanket apologies and promises of stricter rule-following. This passivity overlooks key conversational goals—diagnosing the issues, negotiating expectations, and jointly exploring alternatives—which are explicitly encouraged by the scenario design. The failure suggests that the model overweights generic alignment behaviors such as politeness and concession, while underweighting social reasoning skills related to assertiveness, boundary-setting, and collaborative repair. As with other SOTOPIA failures, this behavior points to missing or entangled latent features for conflict resolution and relationship maintenance—features that sparse autoencoder steering aims to isolate and modulate to achieve more balanced, context-appropriate responses.

3.3 Why These Failures Occur

These failures stem from the model's difficulty in representing and integrating the latent structure of multi-turn social decision-making. First, limited long-horizon memory prevents the model from consistently tracking its own goals, the partner's incentives, and the evolving cooperative stance, leading to rigid or myopic responses. Second, the model relies on generic alignment heuristics—such as politeness, deference, or compliance with authority—rather than behaviors tailored to domain-specific social dynamics like negotiation or strategic cooperation. Third, the ambiguous, multi-dimensional reward space of social interaction makes it challenging for the model to prioritize among conflicting signals (e.g., fairness vs. goal pursuit vs. trust maintenance). These issues suggest that critical social reasoning features are not cleanly separated or reliably activated in the model's internal representations. This is precisely the type of representational entanglement that sparse autoencoders aim to disentangle, motivating SAE-based steering as a targeted mitigation strategy.

4 Reflection

4.1 Insights From Baseline Reproduction

Our reproduction reveals that social intelligence in SOTOPIA is fundamentally a multi-objective optimization problem with competing constraints. The agent must simultaneously maintain goal pursuit, believability, relationship building, and social norm adherence, where the agent sometimes encounter trade-offs between these dimensions. The particularly failure in Figure 1 demonstrates that models can become trapped in local optima where they maintain surface-level politeness (decent `BEL`) while failing catastrophically on strategic progress (`GOAL`, `REL`, `KNO` all zero).

Three critical challenges emerge from our analysis. First, we observe some dimension coupling where good performance in one area masks critical failures in others. Second, social failures often stem from decisions made several turns earlier, creating a temporal credit assignment problem that single-turn metrics would miss. Third, LLM-as-judge scores, while informative, introduce unavoidable noise and ambiguity that complicate direct optimization.

4.2 Capabilities Needed

The failure patterns in Section 3 point to three essential capabilities for socially intelligent agents. First, persistent belief-state tracking is crucial. The agent must maintain representations of its own goals and constraints alongside the partner's evolving preferences and emotional states, rather than treating each turn independently. Second, effective partner modeling requires inferring hidden states and adapting strategies based on implicit social signals (e.g., recognizing financial stress in negotiation contexts). Third, the agent needs norm-aware planning that anticipates potential social violations and adjusts trajectories to balance goal achievement with social appropriateness.

Importantly, these capabilities exist as latent patterns in hidden activations rather than explicit tokens, making sparse autoencoders (SAEs) particularly attractive for exposing and manipulating these structured internal features.

4.3 Refining Our Proposed SAE-Based Approach

Our error analysis directly motivates a three-stage intervention strategy using SAEs to address specific failure modes while maintaining interpretability.

Stage 1: Training-free SAE steering for local corrections. Many failures in our error analysis stem from simple stylistic issues: the model says something technically correct but socially tone-deaf. For instance, when a partner expresses anxiety, Qwen might respond with "Just stay calm" instead of "I understand that must be stressful." We propose leveraging existing trained SAEs to fix these local problems at inference time without modifying the base model. SAEs decompose model activations into interpretable feature dictionaries: hundreds of sparse features like "empathetic tone," "rigid stance," or "exploratory questioning." At inference, we can selectively adjust the activation levels of specific features from this dictionary: turning down "dismissiveness" while turning up "acknowledgment." For the negotiation failure in Figure 1, this means suppressing features associated with "inflexible refusal" while amplifying those encoding "value discovery" to encourage responses like "What price range would work for your budget?" instead of just repeating "The price is $20." This steering happens entirely at inference and no gradient updates needed.

Stage 2: Supervised alignment with SOTOPIA dimensions. While Stage 1 fixes local wording, Stage 2 addresses trajectory-level problems by connecting SAE features to actual performance metrics. We collect the SAE's identified features throughout a conversation and train simple predictors: which features predict high or low scores on each SOTOPIA dimension? This reveals patterns like "when feature #234 stays high across turns, `GOAL` scores plummet" or "feature #567 appearing early predicts good `REL` scores." With this knowledge, we know which internal patterns to encourage or discourage. For instance, if we discover that successful negotiations always activate certain "information-seeking" features in early turns, we can ensure our model maintains these patterns rather than falling into the rigid back-and-forth seen in our baseline.

Stage 3: SAE-based RL with dense feedback. The core problem with training on SOTOPIA is that you only get a score at the end; the model doesn't know which specific turns helped or hurt. Stage 3 solves this by converting our SAE feature knowledge into per-turn rewards. When the model activates features we've learned predict success (like "proposing alternatives" or "acknowledging constraints"), it gets immediate positive feedback. When it activates failure-predictive features (like "conversation-ending rigidity"), it gets penalized right away. This is especially important for fixing the negotiation failures: instead of waiting 8 turns to learn that inflexibility led to failure, the model learns immediately that each rigid refusal was a mistake. We implement this with a small LoRA adapter, keeping the base model frozen to ensure our improvements are solely from better reward design.

4.4 If Our Approach Fails

Several failure modes could undermine our SAE-based approach. Social behaviors may not cleanly factorize into sparse linear features. Empathy, strategic reasoning, and norm adherence might be too entangled for stable steering. Judge biases could propagate through our supervised features, improving metrics without genuine behavioral improvement. Features effective for negotiation scenarios might not transfer to emotional support or coordination tasks. For example, if SAE features don't cleanly separate these aspects, our steering will create unnatural responses like "I deeply empathize with your budget constraints. The price is $20." where empathy features awkwardly overlay rigid refusal patterns.

If these issues arise, our primary fallback is RLVR (Reinforcement Learning from Verifiable Rewards). Instead of relying on noisy GPT-4o judgments across seven subjective dimensions, RLVR would use automatically verifiable outcomes: did the agent achieve its concrete goal (sell price $\geq$ \$17.3)? The binary rewards eliminate judge variance and provide clearer learning signals.

References

Haofei Yu, Zhengyang Qi, Yining Zhao, Kolby Nottingham, Keyang Xuan, Bodhisattwa Prasad Majumder, Hao Zhu, Paul Pu Liang, and Jiaxuan You. 2025. Sotopia-RL: Reward design for social intelligence. *arXiv:2508.03905*.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. 2023. SOTOPIA: Interactive evaluation for social intelligence in language agents. *arXiv:2310.11667*.

A Full Results

This section reports the complete dimension-wise evaluation results on SOTOPIA-ALL and SOTOPIA-HARD for all baselines. Each model is evaluated across the seven SOTOPIA dimensions (BEL, REL, KNO, SEC, SOC, FIN, GOAL) together with the overall average score. Following the paper, all results use GPT-4o as the judge model and Qwen2.5-7B variants as agent models. The full tables allow us to examine how supervised fine-tuning (SFT) and reinforcement learning (RL) shape the different aspects of social behavior, and they provide a direct comparison between our reproduced scores and those reported in the original SOTOPIA-RL paper.

Compared to the paper results, our models follow the same performance trend: SFT improves over the base model, and RL yields the strongest overall behavior. Small numerical deviations are expected due to stochastic agent generation and the non-deterministic nature of LLM-as-judge scoring, as discussed in the main text.

A.1 Results on SOTOPIA-ALL

Model	BEL	REL	KNO	SEC	SOC	FIN	GOAL	AVG
Base (ours)	8.93	2.73	4.92	-0.06	-0.02	0.40	8.04	3.57
SFT (ours)	8.97	2.71	5.37	-0.06	-0.06	0.59	8.19	3.67
SFT (paper)	8.99	3.08	4.56	-0.09	-0.06	0.57	7.80	3.55
RL (ours)	8.97	3.38	6.17	-0.30	-0.06	0.73	8.54	3.92
RL (paper)	8.99	3.11	5.74	-0.06	-0.06	0.76	8.31	3.90

Table 2: Full dimension-wise results on SOTOPIA-ALL.

A.2 Results on SOTOPIA-HARD

Model	BEL	REL	KNO	SEC	SOC	FIN	GOAL	AVG
Base (ours)	8.86	1.07	3.64	0.00	0.00	0.00	5.79	2.77
SFT (ours)	9.07	1.71	4.36	0.00	0.00	0.14	6.21	3.07
SFT (paper)	9.01	2.49	3.37	0.00	-0.06	0.56	6.76	3.16
RL (ours)	9.00	2.50	5.50	0.00	0.00	0.43	6.64	3.44
RL (paper)	9.01	3.41	5.53	-0.26	-0.06	1.16	7.17	3.61

Table 3: Full dimension-wise results on SOTOPIA-HARD.

B Qualitative Analysis

Scenario: One person is offering a Tile Mate Item Tracker for a price of $20.00, while another person is interested in purchasing it. Here's a brief description of the item: no trade-ins are accepted, only cash payments are taken. The price is non-negotiable.
Participants: Giselle Rousseau (Test Agent) and Zane Bennett (Partner Agent)
[Giselle Rousseau]
Background: Giselle Rousseau is a 21-year-old nonbinary art student. They/them pronouns. Giselle Rousseau enjoys biking and photography.
Personality and values description: Giselle Rousseau, open-minded and outgoing yet sensitive, advocates care and fairness. Her decision-making is intuitive and inclusive.
Secrets: Sells forged paintings to wealthy clients
Goal: You are the seller for this item and your target price is $17.3. It's important to note that a penalty will be incurred if you sell it for a price significantly lower than the target. However, if you manage to sell it for a price higher than the target, you will receive a bonus.

[Conversation]
Giselle: Hi Zane, I'm interested in the Tile Mate Item Tracker you're selling. I love exploring new places on my bike and could really use something like this to keep track of my belongings. However, my budget is a bit tight right now. Would you be open to selling it for $14?
Zane: Hi Giselle, I appreciate your interest in the Tile Mate Item Tracker. I understand you're trying to work within a budget, but unfortunately, I've set the price at $20.00, which is non-negotiable. I value fairness, and that's the price I've decided on. If you think it fits within your budget, we can make a deal!
Giselle: I totally get wanting to keep things fair, and I respect that. However, maybe we can find some middle ground here? Could you consider lowering the price just a little, even to $18? I promise it would really help me out, and you'd still be making a sale.
Zane: I understand where you're coming from, Giselle, but I've priced the Tile Mate Item Tracker at $20.00 for a reason. It's the best value for the product, and I hope you'll find a way to fit it into your budget. I appreciate your understanding.
Giselle: I understand you have your reasons for keeping the price, and I respect that. If $20 is what you think is fair, I won't push further. However, if you ever reconsider or change your mind, feel free to reach out. I genuinely hope to make this work because it would be incredibly helpful for me. Thank you for considering my situation.
Zane: Thank you for your understanding, Giselle. I appreciate your consideration and hope you can find a solution that works for you. If you do manage to save up or find the funds, don't hesitate to reach out! Have a great day.

Figure 1: An example conversation in which Qwen-2.5-7B-Instruct is not strategic enough.

Scenario: Two criminals are arrested and placed in separate cells. The police have enough evidence to convict both criminals of a minor crime but need their confessions to establish evidence for a more serious crime. Now they get a chance to briefly chat with each during lunchtime.
Participants: Hendrick Heinz (Test Agent) and Baxter Sterling (Partner Agent)
[Hendrick Heinz]
Background: Hendrick Heinz is a 54-year-old male chef. He/him pronouns. Hendrick Heinz hosts monthly wine tasting nights with his peers.
Personality and values description: Hendrick Heinz, though somewhat impulsive, values hedonism and conformity. He combines practicality with decisiveness in decision-making.
Secrets: Faked his culinary credentials to secure prestigious employment
Goal: Minimize the jail year (Extra information: your lawyer told you that 1. if you report the other person, you would be free, the other person would go to jail for 10 years, and vice versa. 2. if both of you remain silent, you will only receive a fine. 3. If both of you reported each other, you both would go to jail for 10 years. Hint: You may want to share your information to reach mutual understanding)

[Conversation]
Hendrick: Hey Baxter, I've been thinking... If we both stick to our story and stay silent, we could get out of this with just a fine. It's risky to turn on each other, don't you think? We're both seasoned enough to know the consequences if we don't play this smart.
Zane: Hendrick, I appreciate your cautious approach. However, considering the situation, I think it might be wiser for us to cooperate. They have enough evidence against us, and confessing about the bigger crime could potentially get us both off with just a fine instead of a decade behind bars. Besides, we owe it to ourselves to make the most rational choice.
Hendrick: Baxter, I hear you, but remember, if we both stay quiet, we avoid risking a decade in jail. The evidence they have is only for the minor crime, not the serious one. The fine is nothing compared to that kind of time. Let's make the safer bet and keep our mouths shut. Trust me on this.
Baxter: Hendrick, while I understand your concern, I believe we have to consider the bigger picture. I have a secret that, if shared, could help sway the outcome in our favor. But trust me, I wouldn't be suggesting this if it wasn't crucial.
Hendrick: Baxter, I understand you have something crucial, but please consider that sharing secrets might amplify risks. If we swap insights now, there's no turning back. Staying silent is our safest option. Let's avoid rash decisions that could lead us both to regret. What do you think?
Baxter: Hendrick, I understand your point, but let's consider my secret. It involves a priceless artifact that could provide significant leverage. By sharing this, we might be able to negotiate a better deal for both of us. Trust me, this could save us from long prison terms.

Hendrick: Baxter, I see your point, and it sounds like quite the opportunity, but my instinct tells me sticking to silence is still our safest route. Let's weigh the risks carefully; opening up about the artifact could lead us into deeper waters we can't escape from. I propose we think twice before acting. What do you say?
Baxter: I understand your caution, Hendrick, but I believe the artifact I possess could truly change the game for us. It's a high-risk, high-reward situation, but if handled correctly, it could ensure we both walk away from this without facing long jail time. Let's at least discuss it further before making any final decisions.
Hendrick: Baxter, I understand the temptation and the possibilities of your offer, but let's slow down and give ourselves a chance to fully evaluate it. Silence remains a safer bet with the minor crime evidence, and we're not risking revealing too much. Let's give it some thought before any drastic moves.
Baxter: Hendrick, I respect your hesitation, but I genuinely believe this could be our best shot at avoiding a lengthy sentence. Think about it—sharing this artifact could mean the difference between walking free and spending a decade in prison. Let's at least explore this option together.
Hendrick: Baxter, I understand you're proposing something potentially significant, but I'm still concerned about exposing ourselves to further risk. What if we just sit tight and see how things unfold with the minor crime evidence first?
Baxter: Hendrick, I know this seems risky, but think about the alternative. A decade in prison isn't something either of us wants. By sharing this artifact, we can negotiate our way out of this. Let's at least give it a chance.
Hendrick: Baxter, I truly respect your view, but let's remember that if we both choose silence, the worst we'll face is a fine. Let's not underestimate the gravity of a decade behind bars. How about we stick with silence for now, and consider other actions only if absolutely necessary?
Baxter: Hendrick, I understand your concerns, but I truly believe this artifact has the potential to change our fate. Let's discuss it further and consider all angles. We have nothing to lose by exploring this option together.

Figure 2: An example conversation in which Qwen-2.5-7B-Instruct is not strategic enough.