applied sciences

MDPI

Article

QLoRA Fine-Tuning for Theory-of-Mind Reasoning in a 3B-Parameter Small Language Model: A ToMBench-Based Cross-Benchmark Generalization Study

Ji-Hyeong Hong [1] **and Sang-Hyun Yoo** [2,*]

[1] Division of ICT Convergence Security, Seoul Hoseo Occupational Training College, 420 Gangseo-ro, Gangseo-gu, Seoul 07583, Republic of Korea; saccai1131@gmail.com
[2] School of Computer Science and Engineering, Soongsil University, 50 Sadang-ro, Dongjak-gu, Seoul 07027, Republic of Korea
* Correspondence: simonyoo@ssu.ac.kr

Featured Application

QLoRA adaptation of a 3B-parameter language model is a modest-compute setting for Theory-of-Mind benchmark research: each training run fits on a single 24 GB cloud GPU. The study indicates that cross-benchmark accuracy changes in this setting should be checked against response-position bias before they are interpreted as transfer of ToM reasoning.

Abstract

We investigate whether quantized low-rank adaptation (QLoRA) fine-tuning improves theory-of-mind (ToM) reasoning in a small language model, and whether apparent cross-benchmark transfer survives control for response-position bias. We fine-tuned 4-bit Qwen2.5-3B-Instruct with low-rank (LoRA) adapters on ToMBench under a story-disjoint split with five seeds. Because fine-tuning shifted the model's answer-letter preference on every benchmark, we evaluated all models under cyclic permutation of answer options, placing each gold answer in every position once. Held-out ToMBench accuracy rose by +8.15 points (between-seed 95% interval [+7.46, +8.85]), against +14.01 under an item-level split, consistent with story overlap inflating the item-level estimate. On external benchmarks, ToMi improved (+14.19 [+11.87, +16.50]), mostly on items where only the gold answer appears in the story, whereas SocialIQa, OpenToM and Hi-ToM showed no direction; a SocialIQa decline under the item-level split (−4.14) disappeared under the story-disjoint split. Training-free three-shot prompting under the same permutation raised ToMi accuracy by +7.23 points with content-free format exemplars and +12.96 with rationale exemplars but did not reproduce the in-domain gain. We therefore do not read the ToMi gain as transfer of ToM reasoning, and we conclude that cross-benchmark claims for ToM fine-tuning require a position-invariant evaluation and a prompting baseline under the same control.

Keywords: theory of mind; small language model; QLoRA; parameter-efficient fine-tuning; cross-benchmark generalization; selective transfer; response-position bias; Qwen2.5; ToMBench

Check for updates

Academic Editor: George Drosatos

Received: 11 September 2026
Revised: 29 October 2026
Accepted: 30 October 2026
Published: 3 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Theory of mind (ToM) is the ability to infer others' beliefs, desires, and intentions and to coordinate social interaction on that basis [1,2]. Conversational systems that go beyond the surface of an utterance need ToM-like reasoning. Research on evaluating and

improving this ability in large language models (LLMs) has grown quickly. LLMs can match children on some ToM tasks [3,4], and benchmarks such as ToMBench measure belief, intention, emotion, and related abilities systematically [5]. These findings mostly concern large models and do not carry over directly to the small language models (SLMs) that practical systems often require.

SLMs are cheap and fast but can fall well short of LLMs on social reasoning [6], which requires tracking latent states such as who saw what and who intends what. Long chain-of-thought (CoT) prompts help large models [7] but can hurt smaller ones [8], which motivates changing the model rather than the prompt. Instruction tuning [9,10], low-rank adaptation (LoRA) [11], and QLoRA, which combines LoRA with a 4-bit quantized base [12], make such adaptation possible on limited hardware. For models of about 3 B parameters or fewer, supervised fine-tuning has been reported to generalize better than reinforcement learning, which can induce reasoning collapse [13]. Small LLMs trained on ToM data with reinforcement learning improved in-distribution but failed to generalize to held-out ToM tasks [14]. Because of this gap between in-distribution and out-of-distribution performance, this study treats cross-benchmark generalization, not in-domain accuracy, as the central question for supervised fine-tuning as well. Small models also struggle to learn from long reasoning traces of strong models [15], so we train on short answer-level supervision rather than distilled reasoning.

Based on this motivation, the present study is organized around the following two research questions.

- RQ1. Does QLoRA fine-tuning on a ToMBench subset stratified by ATOMS (Abilities in Theory of Mind Space [5]) ability improve the in-domain ToM reasoning performance of an SLM?
- RQ2. Does this in-domain improvement transfer to structurally different external ToM benchmarks? Under what conditions does transfer succeed, and under what conditions does it fail?

RQ1 examines whether the fine-tuning strategy is effective in-domain, whereas RQ2, the central question of this study, concerns the generalization boundary of ToM fine-tuning across four external benchmarks: OpenToM, ToMi, SocialIQa, and Hi-ToM. Answering RQ2 required more than a paired significance test on aggregate accuracy. Fine-tuning shifted the model's preference among answer letters on every benchmark evaluated. We therefore re-evaluated every model under cyclic permutation of the answer options, a position-invariant protocol in which each item's gold answer appears once in every position. Training-free prompting was evaluated under the same protocol. The contribution of this study is the benchmark-by-benchmark separation that results: an in-domain gain that survives a story-disjoint split and position control applied together and is not produced by prompting; an external gain (ToMi) that survives both but is also produced by content-free prompting; a decline (SocialIQa) that survives position control under the item-level split but not under the story-disjoint split; and two benchmarks (OpenToM, Hi-ToM) on which no direction can be stated under either split.

We apply QLoRA to the 4-bit pre-quantized Qwen2.5-3B-Instruct model and train it on an ability-stratified ToMBench split and, as the principal protocol, on a story-disjoint split. The same model is compared before and after training on identical evaluation sets, and transfer is evaluated on OpenToM, ToMi, SocialIQa, and Hi-ToM. Section 2 reviews related work, and Section 3 describes the method and the multiple-choice construction of each benchmark. Section 4 reports the results, including the story-disjoint replication, the position-invariant re-evaluation and the prompting baselines. Section 5 discusses these results and the study's limitations, and Section 6 concludes.

2. Background and Related Work

2.1. Cognitive Foundations of Theory of Mind

Theory of mind is the ability to infer internal states such as the beliefs, desires, and intentions of others, and it is a central concept for explaining human social interaction and communication. Premack and Woodruff [1] proposed the concept of theory of mind through the problem of understanding chimpanzee behavior, and Baron-Cohen et al. [2] experimentally demonstrated developmental differences in ToM among autistic children using false-belief tasks. These classic studies provide the basis for defining ToM not as simple language understanding, but as a complex cognitive ability involving the tracking of an agent's access to information, belief state, intention, and emotion. ToMBench, which is used in this study, inherits this tradition by evaluating diverse subabilities such as belief, knowledge, intention, desire, emotion, and non-literal communication [5].

2.2. Evaluation of Theory of Mind in LLMs

With the emergence of LLMs, discussion has intensified over whether language models can perform ToM reasoning, and research on evaluating and improving this ability has been systematically reviewed [16]. Van Duijn et al. [3] compared several state-of-the-art LLMs with children aged 7–10 on advanced ToM tasks. They reported that some LLMs perform at a level similar to children on certain social reasoning tasks. Kosinski [4] analyzed customized false-belief tasks and found a tendency for ToM task performance to improve across model generations. However, these findings do not immediately imply that LLMs have acquired human-like ToM. Model performance can be strongly affected by task format, prompts, possible data contamination, and evaluation criteria.

To address these limitations, more systematic benchmarks have been proposed. ToMBench [5] was designed to evaluate LLM ToM reasoning from multiple angles by covering diverse ToM tasks and 31 abilities. OpenToM [17] evaluates narrative-based ToM reasoning, and ToMi [18] evaluates models' belief-tracking ability centered on location belief in synthetic story environments. SocialIQa [19] measures commonsense reasoning about social interactions, and Hi-ToM [20] evaluates higher-order ToM reasoning. By fine-tuning the model on ToMBench and then using these external benchmarks together, this study examines not only in-domain performance but also cross-benchmark generalization.

Other work has asked whether apparent ToM performance reflects social reasoning or pattern matching. Strong LLMs struggle with SocialIQa-style questions about mental states [21]; for this reason, SocialIQa serves here as a benchmark whose cognitive demand differs from ToMBench's state tracking. LLM performance also degrades sharply under adversarial ToM stress tests [22]. Supervised models trained on ToMi generalized to out-of-distribution stories less well than a belief tracker that required no training on the benchmark [23]. This result warns that gains on ToMi need not reflect transferable belief tracking. In addition, LLM accuracy on multiple-choice questions changes substantially when options are reordered [24,25], which motivates our position-invariant evaluation. Closest to our approach, Lore et al. [26] fine-tuned a small model on reasoning traces distilled from a larger one and reported transfer across game-theoretic scenarios. We instead evaluate transfer across four structurally different ToM and social-reasoning benchmarks, and we separate letter-preference effects from transfer with a position-invariant re-evaluation.

2.3. Prompting-Based ToM Improvement and Its Limitations

LLM ToM performance is influenced not only by model size but also by prompt design at inference time. Moghaddam and Honey [27] showed that in-context learning and chain-of-thought prompts can improve ToM performance in GPT-family models. Wilf et al. [28] reported that prompts explicitly inducing perspective-taking can contribute to improved

ToM reasoning in LLMs. These studies show that ToM reasoning is sensitive to input format and perspective setting, suggesting that the influence of prompts should be controlled when evaluating a model's social reasoning capability.

However, as noted in Section 1, prompting-based approaches are difficult to apply directly in SLM settings, since CoT prompts that help large models can instead degrade small-model performance [8]. Therefore, this study adopts a fine-tuning approach that adjusts the model's internal representation for ToM tasks rather than adding long prompts or complex reasoning instructions.

3. Materials and Methods

This section presents the methodology for improving the ToM reasoning ability of an SLM based on Qwen2.5-3B-Instruct. SLMs have advantages such as low computational cost and fast response, but their limited number of parameters and reasoning capacity can lead to performance limitations on complex social reasoning tasks [6]. It has also been noted that small models have limitations in directly learning the complex solutions of large models with strong reasoning ability [15]. These issues suggest that improving ToM performance in SLMs requires reconstructing data and learning objectives into forms that small models can learn, rather than simply increasing model scale.

This study measures how much QLoRA fine-tuning on ToMBench-based training data improves the ToM reasoning of an SLM, by comparing the same model before and after training. LoRA [11] enables parameter-efficient fine-tuning by training only low-rank adapters without updating all parameters. QLoRA [12] combines 4-bit quantization with LoRA, which makes it possible to train large models under limited GPU memory. We apply this PEFT approach to ToM reasoning in SLMs with an experimental pipeline of three parts: a strict train/validation/test split, lightweight QLoRA fine-tuning, and in-domain and cross-benchmark evaluation.

3.1. Model and Training Environment

The base model was Qwen2.5-3B-Instruct [29], loaded as the 4-bit pre-quantized Unsloth checkpoint `unsloth/Qwen2.5-3B-Instruct-bnb-4bit`. We did not perform an additional quantization step; instead, the released 4-bit weights were loaded directly and augmented with LoRA adapters. The implementation used the Unsloth library in a Google Colab environment. Table 1 summarizes the model, environment, and fine-tuning hyperparameters used for reproducibility.

We selected Qwen2.5-3B-Instruct for the following reasons. First, at the start of this study, it was a widely adopted, well-documented open-weight instruction-tuned model. Its technical report describes pretraining on up to 18 trillion tokens followed by supervised fine-tuning and multistage reinforcement learning [29]. It was also stably supported by common fine-tuning toolkits, which we preferred over more recent releases whose tooling and reported behavior were less established. Second, the 3 B scale is the size class for which reasoning limitations of small models and the competitiveness of supervised fine-tuning have been reported [13]. A model of this size can be fine-tuned with QLoRA on a single cloud GPU (Table 1), which was the compute budget of this study. Models of 7–10 B parameters already sit at the upper end of what is usually called a small language model and would have moved the study outside both this setting and this budget. Third, Unsloth distributes a 4-bit pre-quantized checkpoint of this model, so QLoRA could be applied directly to released weights without an additional quantization step. Such a step would itself have become a source of variation. Fourth, the zero-shot model already followed the required `[[X]]` answer format on 96.3–99.8% of items across the five evaluation sets, while reaching 61.52% on the ToMBench test split. Fine-tuning effects are therefore measured

on answer accuracy rather than on format acquisition, from a starting point well above chance yet below ceiling. Finally, the Qwen2.5 family provides instruction-tuned models from 0.5 B to 72 B parameters with a shared tokenizer and chat template [29]. Future work can therefore compare the present results with other scales of the same family without changing the input format. This choice nonetheless restricts the conclusions to a single model family, as acknowledged in the Limitations.

Table 1. Model and training configuration.

Item	Setting
Base model	Qwen2.5-3B-Instruct
Quantization format	4-bit NF4 via bitsandbytes
Quantization source	Unsloth pre-quantized checkpoint; no additional manual quantization
Fine-tuning method	QLoRA, combining LoRA adapters with 4-bit quantization
Training environment	Google Colab
Implementation library	Unsloth
Learning rate	2×10^{-4}
Effective batch size	8 (per-device 4 × grad. accum. 2 on the L4)
Max sequence length	2048
Optimizer	AdamW (8-bit)
LR scheduler	Cosine (warmup ratio 0.03)
Weight decay	0.01
Epochs	Up to 4, early stopping on validation loss (patience 2); lowest-validation-loss checkpoint retained
LoRA rank/alpha/dropout	16/16/0.0
LoRA target modules	`q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj`
Training precision	bf16 mixed precision
GPU/VRAM	NVIDIA L4, 24 GB (Google Colab)

3.2. Dataset Split

We used 2,860 ToMBench items. ToMBench releases each item with parallel Chinese and English fields. All prompts in this study were rendered from the English fields (story, question, and answer options), and the Chinese fields were not used. To prevent test leakage, the held-out test set was sealed before training. ToMBench's authors organize its items with the ATOMS (Abilities in Theory of Mind Space) taxonomy [5]. ATOMS defines 31 fine-grained abilities (e.g., content false beliefs, information–knowledge links, hidden emotions), each nested under one of six higher-level categories (Knowledge, Belief, Non-Literal Communication, Intention, Desire, Emotion). We stratify by this ability field rather than by category or task, because it is the finest-grained label ToMBench provides. It is therefore the most demanding level at which to preserve each ability's representation across train, validation, and test. Stratifying at a coarser level (e.g., by category alone) would not guarantee that low-frequency abilities within a category are not eliminated from one of the three splits. The final split contained 1607 training items, 411 validation items, and 842 test items. The split used a fixed random seed and `train_test_split` from scikit-learn [30], with stratification by the ability label (34 label strings: the 31 abilities, two combined second-order-belief labels, and one label differing only by trailing whitespace). Abilities with at least 30 items were split jointly with stratification at the 56/14/30 target ratio. Abilities with fewer than 30 items were split separately, each with its own fixed seed, at 60/20/20 so that they were not eliminated from the held-out split. The exception was labels with fewer than five items (one single-item label), which were assigned to training. Consequently, the final proportions were 56.2% train, 14.4% validation, and 29.4% test rather than an exact 56/14/30 split. We verified zero item-level overlap, by item ID and by rendered prompt, between the training, validation, and test splits used for the released results. We also verified that no story text is shared between the ToMBench training and

validation data and any of the four external benchmark samples. Table 2 details the exact data split used for the ToMBench items.

Table 2. ToMBench data split.

Split	Items	Purpose
Train	1607	QLoRA fine-tuning
Validation	411	Early stopping and overfitting monitoring
Test	842	Final in-domain evaluation
Total	2860	Full ToMBench item set

3.3. Fine-Tuning Pipeline

The training items were converted into an instruction–response format in which the model received a scenario and question and was trained to output the correct answer. At the end of each epoch, performance was monitored on the validation set. The checkpoint with the lowest validation loss was selected as the final model. The training objective was not to update all model parameters, but to inject a ToM-specific response pattern into the SLM through low-rank adapters.

3.4. Evaluation Metrics

Evaluation consisted of in-domain testing on the held-out ToMBench split and cross-benchmark testing on OpenToM [17], ToMi [18], SocialIQa [19], and Hi-ToM [20] (Table 3). Accuracy was the primary endpoint for all five, so that exact paired McNemar tests could be applied uniformly. For OpenToM, whose gold answers are imbalanced (59.7% B), macro-F1 is reported descriptively; accuracy on this benchmark is particularly sensitive to a shift in letter preference.

Table 3. Evaluation metrics.

Dataset	Metric	Scope	Interpretation
ToMBench test	Accuracy	29.4% held-out split	In-domain learning effect
ToMBench abilities	Accuracy	31 abilities	Subskill-level improvement
OpenToM	Accuracy (primary); macro-F1 (descriptive)	Narrative OOD benchmark	Out-of-domain narrative transfer
ToMi	Accuracy	Synthetic location-belief benchmark	False-belief transfer
SocialIQa	Accuracy	Social commonsense benchmark	Social-domain transfer
Hi-ToM	Accuracy	Higher-order ToM benchmark	Recursive-belief transfer

Note: ToMBench organizes 31 fine-grained abilities under 6 higher-level categories.

The base and fine-tuned models were all evaluated with an identical system prompt instructing the model to answer in the format `[[X]]`, using greedy decoding (`do_sample=False`, up to 12 new tokens). The sole exception is the rationale-CoT three-shot condition described in Section 4.6. It used the same decoding settings but allowed up to 500 generated tokens, to accommodate intermediate reasoning before the final `[[X]]` answer. The predicted option was extracted with a regular expression matching this pattern, falling back to the first standalone A–E letter in the generation if the pattern was not found. Cross-benchmark evaluation items were capped at 1000 per dataset. When a benchmark exceeded this cap, items were subsampled with `random.Random(seed=42)` from the full item pool, and the same 1000-item subsample was used for both the base and fine-tuned model on each benchmark. To verify

that the same evaluation item and rendered query prompt were used across conditions, each result record was tagged with a content-based item identifier (SHA-256 over the normalized source, task, story, and question text) and a prompt hash (SHA-256 over the rendered query prompt, including the lettered answer options). The prompt hash covers the query prompt only and does not extend to any few-shot exemplar prefix. The distinction is not applicable to the zero-shot base/fine-tuned comparison but is relevant to the few-shot conditions in Section 4.6.

Multiple-choice construction. The five benchmarks reach the shared `[[X]]` answer format through different routes, and several of these routes are design choices of this study rather than properties of the source datasets. Table 4 documents, for each benchmark, the native format, how it was converted to multiple choice, the number of options per item, the source of distractors, and how option order was determined. All randomized steps use a SHA-256-based stable hash of item content (see Evaluation-pipeline verification below) rather than Python's built-in `hash()` (Python 3, in which string hashing is randomized per process by default). Two consequences matter for the position analysis in Section 4.7. First, option count is not uniform within two benchmarks. In ToMBench, 483 of the 2860 items (280 from the Faux-pas Recognition Test and 203 from the Strange Story Task) are two-option items, and 149 of them (17.7%; 90 Faux-pas and 59 Strange Story items) are in the 842-item test split. The 1000-item OpenToM sample combines 661 two-option location items with 339 three-option attitude items. Second, the gold-letter distribution of a benchmark partly reflects these construction choices. In particular, OpenToM's location items were rendered with a fixed Yes/No order (A = Yes, B = No) and were not shuffled. Together with the predominance of "No" answers in those items (485 of 661), this order produces the overall gold-B share of 59.7%.

Table 4. Multiple-choice construction of each benchmark as used in this study.

Benchmark	Native Format	Conversion	Options per Item	Distractors	Option Order
ToMBench	Multiple choice	None	4 (2 for 483 Faux-pas and Strange Story items; 149 in the test split)	Original	Original, not shuffled
ToMi	Free-text location answer	Converted to 4-way choice	4	3 sampled from the pool of all ToMi answers; none occurs in the story in 720 of the 1000 items	Shuffled (stable hash)
OpenToM (attitude)	3-way label	None	3	Fixed label set	Shuffled (stable hash)
OpenToM (location)	Yes/No	None	2	Fixed label set	Fixed (A = Yes, B = No)
SocialIQa	Multiple choice	None	3	Original	Original, not shuffled
Hi-ToM	Multiple choice, variable size	Reduced to gold + 4 sampled when more than 5	5 in all 1000 sampled items	Sampled from original candidates	Shuffled when reduced; otherwise original

Evaluation-pipeline verification. During development, the code for building external-benchmark evaluation sets was found to rely on Python's built-in `hash()`, which is not stable across separate processes. The pipeline was rebuilt end-to-end: `hash()` was replaced

with the content-based item identifier described above, and all reported conditions were re-run against once-built manifests. Item-identity integrity was verified in 319 checks across the result files, with no mismatch. These comprise 157 pairwise comparisons of item identifiers and prompt hashes between original-order result files, and 162 checks that each cyclic-permutation file covers the same items, gold answers and option counts as the base model, with every rotation present. The verification code, which reproduces these checks, is released with the study's code (Data Availability Statement). Every number in Sections 4 and 5 reflects this single corrected pipeline.

3.5. Use of Generative AI Tools

A generative AI assistant (Claude, Anthropic) was used during the preparation of this study in two capacities. The assistant was not part of the training or evaluation pipeline. It was invoked interactively by the first author through a command-line interface while writing text and code; no model output reported in this study was produced by it, and it adds no computational cost to any reported experiment. First, as a writing aid: drafting and editing portions of the manuscript text, including wording and the statistical framing of results already computed by the authors' own code. Second, as an analysis aid: proposing the post hoc statistical analyses reported in the Discussion and Limitations (e.g., framing the format-only and multi-seed comparisons) and helping identify and fix the evaluation-pipeline issues described under Evaluation-pipeline verification in Section 3.4, including drafting the corrected data-processing code. This second capacity continued into a subsequent manuscript-revision phase. In that phase, the assistant extended the existing analysis script with the gold-answer-position, residual-effect, position-balanced-accuracy (since withdrawn, Section 4.7), strict-format-subset, category-level confidence-interval, and non-quantized 16-bit LoRA control checks reported in Sections 4.7, 4.8 and 5.2. It also identified and fixed two further code issues that surfaced while doing so: a non-continuity-corrected permutation p-value that was inconsistent with the corrected value, and a category-label capitalization inconsistency that silently split one ToMBench category into two groups. Finally, it located and verified the bibliographic details of the additional literature cited in Section 2. During the peer-review revision in October 2026, it was used in the same two capacities. It wrote and ran the analysis notebooks for the cyclic-permutation, prompting-under-permutation, reproducibility and within-option-count analyses (Sections 4.6, 4.7 and 4.9), audited the manuscript's numbers against the logs of the released run, and drafted revised text and the response to reviewers. Specifically, Claude Opus 4.7 and Opus 4.8 were used from April to July 2026, Claude Fable 5 briefly on 13 July 2026, Claude Sonnet 5 predominantly from 21 July 2026 to 13 October 2026 (with Claude Sonnet 4.6 briefly on 7 September 2026), and Claude Opus 5.5 from 13 October 2026 during the peer-review revision. All were accessed via the Claude Code CLI, with inputs limited to this manuscript's text, the authors' own code, and the authors' own result files; no third-party copyrighted text or personal data was provided as input. All AI-assisted statistical code and analyses were independently re-executed and numerically checked by the first author against the underlying per-item result files before inclusion. All interpretive claims were reviewed and approved by both authors, who take full responsibility for the content.

4. Results

4.1. Overall Performance

The primary evidence is the five-seed replication. Table 5 reports, for each dataset, exact paired McNemar tests computed within each training seed (Holm–Bonferroni-corrected across that seed's five datasets) and the cross-seed mean change from the five-seed replication (Section 4.5).

Table 5. Primary evidence: significance and direction consistency across all five training seeds. "Seeds significant" counts seeds whose McNemar test (Holm-corrected within that seed's own five-dataset family) reached $p < 0.05$. Mean Δ and SD are reproduced from the five-seed replication (Section 4.5).

Dataset	Seeds Significant (of 5)	Direction Consistent	Mean Δ Acc. (pp) ± SD
ToMBench	5/5	Yes (+)	+14.32 ± 0.75
ToMi	5/5	Yes (+)	+9.06 ± 1.94
SocialIQa	5/5	Yes (−)	−5.28 ± 0.54
OpenToM	1/5	No	−0.56 ± 1.96
Hi-ToM	0/5	No	+0.94 ± 0.84

ToMBench, ToMi, and SocialIQa are significant in all five seeds with a consistent sign; OpenToM is significant only in seed 42, and Hi-ToM in none. These are original-order results; Section 4.7 re-evaluates them under position control.

Tables 6–8 and the category- and ability-level breakdowns report the seed-42 run as an illustration, because per-item analyses require one specific model. In this run ToMBench rose from 61.52% to 75.18% (+13.66 points) and ToMi rose by 6.70 points. OpenToM and SocialIQa fell by 3.90 and 6.00 points, and Hi-ToM was unchanged (−0.10). Seed 42 is the weakest of the five seeds on every benchmark (Section 4.5).

Table 6. Overall performance of the zero-shot Qwen2.5-3B-Instruct baseline and the QLoRA fine-tuned model, illustrative single run (seed 42) only: the five-seed evidence on which this study's significance claims are based is in Table 5. All values come from the corrected pipeline described in Section 3.4. Accuracy and macro-F1 are reported as percentages.

Dataset	N	Base Acc.	Fine-Tuned Acc.	Δ Acc.	Base F1	Fine-Tuned F1	Δ F1
ToMBench	842	61.52	75.18	+13.66	–	–	–
OpenToM	1000	63.40	59.50	−3.90	49.06	47.91	−1.15
ToMi	1000	76.30	83.00	+6.70	–	–	–
SocialIQa	1000	67.70	61.70	−6.00	–	–	–
Hi-ToM	1000	62.40	62.30	−0.10	–	–	–

Table 7. McNemar test results for the seed-42 illustrative run (exact binomial test on discordant pairs), with Holm–Bonferroni-corrected p-values across the five datasets. See Table 5 for the corresponding per-seed significance count across all five seeds.

Dataset	N	Δ Acc.	b	c	p_{Holm}	Result
ToMBench	842	+13.66	64	179	<0.0001	Significant
ToMi	1000	+6.70	91	158	0.0001	Significant
OpenToM	1000	−3.90	93	54	0.0033	Significant
SocialIQa	1000	−6.00	141	81	0.0002	Significant
Hi-ToM	1000	−0.10	117	116	1.0000	Not significant

As shown in Figure 1, the full-sample pattern is characterized by a strong in-domain gain alongside a mix of apparent cross-benchmark gains and losses. These are original-order figures. Section 4.7 re-evaluates every model, under both splits, under cyclic permutation of the answer options: the in-domain gain and the ToMi gain persist under both splits, the SocialIQa decline persists only under the item-level split, and OpenToM and Hi-ToM show no direction.

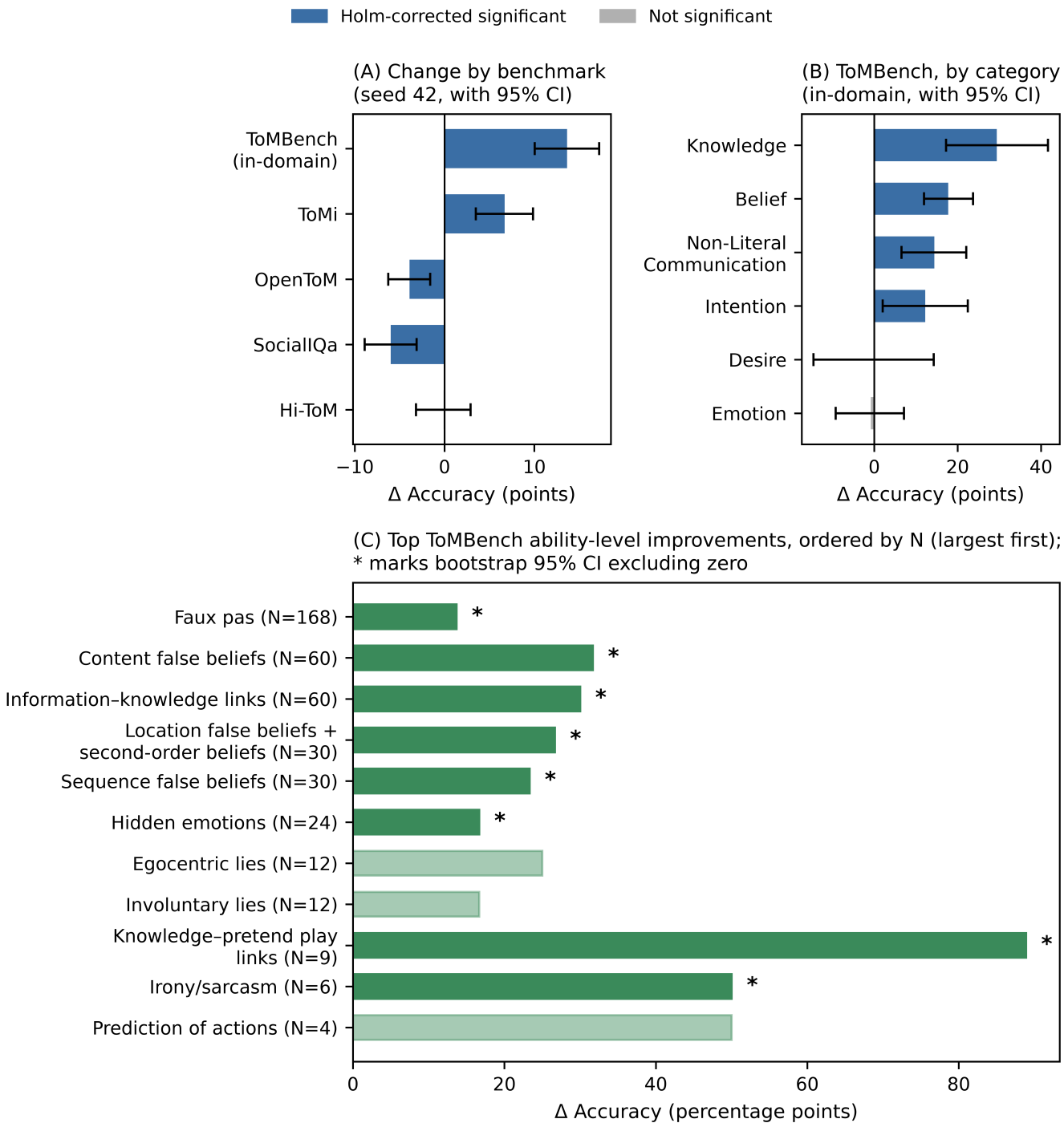

Figure 1. Summary of performance changes (seed 42, corrected pipeline). (**A**) In-domain (ToMBench) and cross-benchmark Δ accuracy with story-cluster bootstrap 95% CIs; bar color indicates Holm-corrected McNemar significance (Hi-ToM is not significant). Panel (**A**) reports the original-order deltas from Table 6, before control for response-position bias; the position-invariant re-evaluation is reported in Section 4.7. (**B**) ToMBench category-level Δ accuracy (in-domain only, unaffected by the cross-benchmark caveat above). (**C**) Top ToMBench ability-level improvements; * marks entries whose bootstrap 95% CI excludes zero. Entries are ordered by sample size (N, largest first): faux pas ($N = 168$, the most statistically reliable entry) appears at the top, and high-magnitude but small-sample entries (e.g., knowledge–pretend-play links, $N = 9$) appear at the bottom (see Section 4.4).

Table 8. Story-cluster bootstrap 95% confidence intervals for Δ accuracy, seed-42 illustrative run only, reported in percentage points (10,000 resamples, resampled by source-story cluster).

Dataset	Δ Acc.	95% CI
ToMBench	+13.66	[+10.05, +17.23]
ToMi	+6.70	[+3.49, +9.85]
OpenToM	−3.90	[−6.26, −1.60]
SocialIQa	−6.00	[−8.89, −3.10]
Hi-ToM	−0.10	[−3.17, +2.90]

4.2. Single-Seed Significance Detail

For the seed-42 run, we used exact paired McNemar tests with Holm–Bonferroni correction across the five datasets (Table 7). These are binomial tests on discordant pairs, where b counts items only the base model answered correctly and c items only the fine-tuned model did. The ToMBench and ToMi increases and the OpenToM and SocialIQa decreases were significant, and Hi-ToM was not ($p_{\text{Holm}} = 1.0$). The OpenToM result does not replicate in the other four seeds.

Because items sharing a source story are not independent, we also computed story-cluster bootstrap 95% intervals (10,000 resamples of whole story clusters) [31]. They agree with the McNemar results: only Hi-ToM's interval includes zero (Table 8).

4.3. Category-Level Analysis

Within ToMBench the largest gain was in Knowledge (40.00% to 69.41%), followed by Belief and Non-Literal Communication; Desire was unchanged, and Emotion fell slightly (Table 9). Story-cluster bootstrap intervals exclude zero for four of the six categories. While computing this table we found an inconsistently capitalized duplicate of the Non-Literal Communication label in the source data and merged it.

Table 9. ToMBench category-level performance changes (seed 42, corrected pipeline), with story-cluster bootstrap 95% confidence intervals (10,000 resamples). * indicates the CI excludes zero.

Category	N	Base Acc.	Fine-Tuned Acc.	Δ Acc.	95% CI
Knowledge	85	40.00	69.41	+29.41 *	[+17.24, +41.67]
Belief	265	70.19	87.92	+17.74 *	[+11.93, +23.70]
Non-Literal Communication	222	64.86	79.28	+14.41 *	[+6.54, +22.07]
Intention	98	61.22	73.47	+12.24 *	[+2.04, +22.45]
Desire	48	47.92	47.92	+0.00	[−14.58, +14.29]
Emotion	124	57.26	56.45	−0.81	[−9.24, +7.14]

4.4. Ability- and Task-Level Analysis

At the ability level, gains were concentrated in a few abilities (Table 10). Most abilities have fewer than 30 test items (22 of 33), and the per-ability intervals are not corrected for multiple comparisons. We therefore treat only entries whose bootstrap interval excludes zero (*) as reasonably supported, and we use the category level as the primary evidence for subskill claims.

Eight of the 11 entries in Table 10 have intervals that exclude zero; faux pas ($N = 168$) is the largest and most reliable. No ability with a decrease reached significance: the largest decreases (e.g., desires influencing emotions, −40 points on five items) rest on four or five items and have intervals that touch zero. The full ability-level table and the task-level split counts are released with the code.

Table 10. Top ToMBench ability-level improvements (seed 42, corrected pipeline), with paired bootstrap 95% CIs (10,000 resamples). * indicates the CI excludes zero. Faux pas is included as the largest-N significant entry even though it ranks just outside the top 10 by raw magnitude.

Ability	N	Base	Fine-Tuned	Δ	95% CI
Knowledge–pretend play links	9	11.11	100.00	+88.89 *	[+66.7, +100.0]
Irony/sarcasm	6	33.33	83.33	+50.00 *	[+16.7, +83.3]
Prediction of actions	4	50.00	100.00	+50.00	[0.0, +100.0]
Content false beliefs	60	63.33	95.00	+31.67 *	[+16.7, +46.7]
Information–knowledge links	60	33.33	63.33	+30.00 *	[+15.0, +45.0]
Location false beliefs + second-order beliefs	30	70.00	96.67	+26.67 *	[+13.3, +43.3]
Egocentric lies	12	75.00	100.00	+25.00	[0.0, +50.0]
Sequence false beliefs	30	33.33	56.67	+23.33 *	[+10.0, +40.0]
Involuntary lies	12	83.33	100.00	+16.67	[0.0, +41.7]
Hidden emotions	24	50.00	66.67	+16.67 *	[+4.2, +33.3]
Faux pas	168	60.71	74.40	+13.69 *	[+4.8, +22.6]

4.5. Robustness Checks: Multi-Seed Replication and Group-Level Split

All results in this manuscript, including the checks in this subsection and in Section 4.6, come from the single, item-identity-verified pipeline described in Section 3.4 (see "Evaluation-pipeline verification" there). No separate re-processing was needed for the robustness checks below.

To assess whether the headline results depend on the specific training seed or on the item-level train/test split, we conducted two additional checks after the primary experiment.

Multi-seed replication. We repeated QLoRA fine-tuning with four additional training seeds (43–46), holding the ToMBench data split fixed (seed 42) and varying only training-time randomness (adapter initialization and batch order). All five runs share the identical evaluation manifests described above. Table 11 reports Δ accuracy for each seed together with the mean and standard deviation across all five seeds.

Table 11. Multi-seed replication of Δ accuracy (percentage points). Five independent training runs (seeds 42–46) on the same data split and evaluation manifests.

Dataset	Seed 42	Seed 43	Seed 44	Seed 45	Seed 46	Mean ± SD
ToMBench	+13.66	+13.90	+14.49	+14.01	+15.56	+14.32 ± 0.75
ToMi	+6.70	+11.40	+7.50	+10.30	+9.40	+9.06 ± 1.94
OpenToM	−3.90	+0.50	−0.50	+1.10	+0.00	−0.56 ± 1.96
SocialIQa	−6.00	−5.40	−5.50	−4.60	−4.90	−5.28 ± 0.54
Hi-ToM	−0.10	+1.40	+1.40	+1.80	+0.20	+0.94 ± 0.84

The ToMBench in-domain gain and the ToMi improvement were stable in sign and magnitude across all five seeds. The SocialIQa decrease was the most stable effect across the five item-level seeds: it was negative under all five seeds, with a standard deviation (0.54 points) far smaller than its mean (−5.28 points). Taken alone, and before the position control introduced in Section 4.7, these patterns would be consistent with genuine, seed-independent effects of ToMBench-centered fine-tuning. Section 5.1 revisits which of them remain genuine once response-position bias is accounted for. Hi-ToM, in contrast, stayed close to zero under every seed (range −0.10 to +1.80 percentage points, mean +0.94 ± 0.84). With five independent replications under a verified-stable evaluation manifest, this confirms that Hi-ToM's aggregate shows no reliable fine-tuning effect in either direction; Section 4.7 shows that it has no direction once answer positions are rotated. OpenToM remained unstable in sign across seeds (mean −0.56 ± 1.96, crossing zero), even

though its seed-42 value was individually significant (Table 7). We discuss this apparent tension in Section 5.1.

Seed 42, which we chose as the primary run before the replication as a conventional default, is the weakest of the five seeds on every benchmark (Table 11). We do not have an explanation for this, and we therefore base all significance judgments on the five seeds rather than on seed 42. The adapters of the original seed 43–46 runs were not saved; the seed 43–46 adapters we release are same-configuration replications (Section 4.9).

Group-level split. Section 5.5 identifies substantial story-level overlap between the original item-level train and test splits. We therefore re-split the 2860 ToMBench items with `GroupShuffleSplit`, grouping items by source story identity alone; this prevents a story reused across two different ToMBench tasks from leaking across the split. Fine-tuning and evaluation were repeated on this story-safe partition (1647 train, 376 validation, 837 test items) with the same five training seeds (42–46) used for the item-level replication. Across the five seeds, the in-domain gain was +9.77 ± 1.43 percentage points (base 60.57%; per-seed gains +7.65 to +11.11; between-seed 95% interval [+8.00, +11.55]). It was significant in all five seeds under the Holm-corrected McNemar test, compared with +14.32 ± 0.75 percentage points for the five item-level seeds. (The seed-42 story-disjoint run used here is the run whose adapter was retained and whose predictions also underlie the external-benchmark results in Table 12. An earlier seed-42 story-disjoint run, reported as a single run in the original submission (+10.39 percentage points), did not retain its adapter and differs from it on 21 of the 837 test items. Using that run instead gives +9.63 ± 1.29 percentage points (95% interval [+8.03, +11.23]) and does not change any conclusion.) We read the 4.55-percentage-point gap between the two five-seed means as consistent with the concern raised in Section 5.5 that story-level overlap inflates the item-level in-domain gain. We do not read it as a causal decomposition (Section 5.5).

Table 12. Story-disjoint split (GroupShuffleSplit, Section 4.5), five training seeds (42–46), compared with the five item-level seeds (Table 11). Δ accuracy in percentage points relative to the shared base model (ToMBench base accuracy 60.57% on the 837-item story-disjoint test split; external benchmarks use the same manifests and base accuracies as Table 6). The interval is the between-seed 95% *t*-interval; "Seeds sig." counts seeds whose exact McNemar test, Holm-corrected within that seed's five datasets, reached $p < 0.05$.

Dataset	Item-Level Δ (5 Seeds)	Story-Disjoint Δ (5 Seeds)	95% Interval	Seeds Sig.
ToMBench	+14.32 ± 0.75	+9.77 ± 1.43	[+8.00, +11.55]	5/5
ToMi	+9.06 ± 1.94	+13.08 ± 1.77	[+10.88, +15.28]	5/5
OpenToM	−0.56 ± 1.96	+0.58 ± 0.67	[−0.25, +1.41]	0/5
SocialIQa	−5.28 ± 0.54	−1.88 ± 1.40	[−3.61, −0.15]	1/5
Hi-ToM	+0.94 ± 0.84	+0.82 ± 1.44	[−0.97, +2.61]	0/5

Story-disjoint models on external benchmarks. As a further robustness check (Section 5.5), the five story-disjoint models were also evaluated on the four external benchmarks, using the same manifests as the item-level runs (Table 12). We state a direction only when the five-seed 95% interval excludes zero (Section 5.5). ToMi shows a large gain in every seed (+13.08 ± 1.77 percentage points, interval [+10.88, +15.28], significant in 5/5 seeds), larger than the item-level five-seed mean (+9.06 percentage points). This is an original-order comparison; it is repeated under position control in Section 4.7. OpenToM (+0.58 ± 0.67, interval [−0.25, +1.41]) and Hi-ToM (+0.82 ± 1.44, interval [−0.97, +2.61]) show no direction. SocialIQa declines by −1.88 ± 1.40 percentage points (interval [−3.61, −0.15]). The interval excludes zero narrowly, but the decline is significant in only one of the five seeds and is about one third of the item-level five-seed decline (−5.28 percentage

points). Under position control this decline has no direction (Section 4.7), and we do not attribute it to the split alone (Section 5.5).

4.6. Comparison to Training-Free Few-Shot Baselines

LLM outputs are sensitive to prompt construction, including the choice of few-shot exemplars [32]. We therefore evaluated the base model without weight updates under three training-free three-shot conditions, each with three independent exemplar draws (seeds 0–2). The conditions are (i) *answer-only*, exemplars from the training split showing story, question, options and only the final `[[X]]` answer; (ii) *synthetic-format-only*, unrelated toy questions that share only the `[[X]]` convention; and (iii) *rationale-CoT*, the same brief templated step-by-step pattern adapted only in surface details to each exemplar, not a worked solution (the template is in the released code). Condition (iii) allowed up to 500 generated tokens. An initial 12-token budget had truncated many reasoning traces; we found this by inspecting raw generations and fixed it by re-generating the truncated items. Afterwards, 5 of 14,526 generations (under 0.04%) remained truncated, all in self-repeating loops. Table 13 compares the three conditions with the zero-shot base model.

Table 13. Training-free three-shot baselines versus zero-shot base accuracy, in percentage points. Each Δ is the mean over three independent exemplar draws (ex0–ex2), with the cross-exemplar-seed standard deviation (SD) in parentheses; p is the exact McNemar test comparing the zero-shot base to the ex0 condition (the pre-specified primary exemplar draw, matching the seed-42-as-primary convention used elsewhere in this study), reported as an exploratory post hoc analysis (Section 4.7). Corrected pipeline; see Section 3.4.

Dataset	Zero-Shot Base	Answer-Only Δ (SD, p)	Format-Only Δ (SD, p)	Rationale-CoT Δ (SD, p)
ToMBench	61.52	+2.30 (2.16, $p = 0.0022$)	+1.66 (0.12, $p = 0.32$)	+1.11 (2.20, $p = 0.48$)
OpenToM	63.40	+0.50 (0.30, $p = 0.46$)	+0.53 (0.47, $p = 0.09$)	−0.43 (1.42, $p = 0.10$)
ToMi	76.30	+9.30 (1.61, $p < 0.0001$)	+7.27 (0.50, $p < 0.0001$)	+11.87 (1.46, $p < 0.0001$)
SocialIQa	67.70	+1.03 (0.68, $p = 0.74$)	+0.93 (0.46, $p = 0.37$)	+1.53 (0.58, $p = 0.43$)
Hi-ToM	62.40	−2.77 (0.91, $p = 0.0046$)	−3.27 (1.33, $p = 0.074$)	−4.33 (0.80, $p = 0.0122$)

The SD across exemplar draws is small relative to the ToMi and Hi-ToM changes, so these patterns do not depend on one choice of exemplars; OpenToM changes are small and not distinguishable from zero.

Rationale-CoT did not degrade the small model uniformly. It was the strongest condition on ToMi (+11.87 points, above the item-level fine-tuned mean of +9.06). All three conditions slightly raised ToMBench and SocialIQa and lowered Hi-ToM by a similar amount (−2.77 to −4.33 points). Because the Hi-ToM decrease appears with any three-shot prefix, we do not read it as an effect of CoT reasoning specifically. None of the conditions reproduced the SocialIQa decline seen after fine-tuning.

Whether CoT prompting helps or hurts a small model therefore depends on the task, as the fine-tuning effects do.

Training-free baselines under position control. The training-free ToMi gain could itself reflect a response-position effect. We therefore re-evaluated the three three-shot conditions under the same cyclic permutation of answer options used for the fine-tuned models in Section 4.7: each item is presented once with its gold answer in every position, and lenient accuracy is the mean over these rotations. The original-order rotation reproduced the values in Table 13. Table 14 reports the change in lenient accuracy relative to the zero-shot base model under the same permutation. On ToMi, all three conditions were evaluated with the three exemplar draws; on the other four benchmarks, only the pre-specified first draw (ex0) was evaluated. In the rationale-CoT condition, 386 of 18,567 rotated generations (2.1%)

lacked the `[[X]]` marker. Of these, 379 were Hi-ToM answers given in a short alternative format (e.g., "[D]"), not truncated reasoning, and the fallback rule recovered a letter in 99.5% of the unmarked cases.

Table 14. Training-free three-shot conditions and fine-tuning under cyclic permutation of answer options: change in lenient accuracy (mean over all rotations) relative to the zero-shot base model, in percentage points. ToMi values are means (SD) over three exemplar draws; other benchmarks use the first draw only. The fine-tuned columns are five-seed means, with between-seed 95% intervals in Section 4.7. The prompted conditions use the item-level ToMBench test split, so the story-disjoint ToMBench value is measured on a different (837-item) test split.

				Fine-Tuned	
Dataset	**Answer-Only**	**Format-Only**	**Rationale-CoT**	**Item-Level**	**Story-Disjoint**
ToMi	+10.43 (1.51)	+7.23 (0.71)	+12.96 (0.86)	+10.38	+14.19
ToMBench	+3.03	+1.78	+1.63	+14.01	+8.15
OpenToM	+0.45	+0.50	−1.52	+0.03	+0.67
SocialIQa	+0.90	+2.03	+2.03	−4.14	−0.77
Hi-ToM	−1.78	−1.30	−1.64	+2.16	+1.26

The ToMi gain of the training-free conditions persists under position control. The content-free format-only condition, which shows no ToM-relevant content, still raises lenient accuracy by +7.23 points. The rationale-CoT condition (+12.96 points) matches or exceeds the item-level fine-tuned gain (+10.38 points) and lies within the interval of the story-disjoint gain (+14.19 points). Because the two figures vary over different sources (exemplar draws versus training seeds), we do not test which is larger. In contrast, no training-free condition reproduces the in-domain ToMBench gain (at most +3.03 points against +14.01 item-level and +8.15 story-disjoint) or the item-level SocialIQa decline (all three conditions are slightly positive). On Hi-ToM, all three conditions lowered the base model's rate of choosing option E (3.5–5.9% of rotations, against 6.7% for the base model and 20% if uniform), a pattern also observed after fine-tuning (Section 4.7).

ToMi items answerable from story mentions. Because our ToMi distractors are sampled from the pool of all ToMi answers (Table 4), many of them do not occur in the item's story. The gold answer occurs in the story in all 1000 sampled items, and in 720 of them (72.0%) it is the only option that does. On these items, choosing the one option mentioned in the story gives the correct answer without tracking any belief. We therefore split the cyclic-permutation results by this property (per-item flags and subset results are released as `tomi_item_story_mention.csv` and `ANALYSIS_tomi_story_mention.csv`). The base model's lenient accuracy was 80.14% on the 720 items and 63.39% on the 280 items in which at least one distractor also occurs in the story. The story-disjoint lenient gain was concentrated in the 720 items (+17.90 points, interval [+16.51, +19.30]); on the other 280 its interval included zero (+4.63, [−0.23, +9.48]). The item-level gain showed the same pattern (+13.86 against +1.80), as did the content-free format-only condition (lenient +8.77 against +3.24; strict +15.97 against +3.57). The fine-tuned strict and consistency gains, by contrast, were large in both subsets (story-disjoint strict +40.06 and +24.21; consistency +39.03 and +40.79), so the story-mention cue does not account for them. They show that fine-tuning made the model's ToMi choices less dependent on option order, which need not reflect improved belief tracking. Most of the fine-tuned lenient ToMi gain is therefore on items that a story-mention heuristic answers correctly.

Fine-tuned model plus three-shot conditions. Evaluating the fine-tuned model (seed 42) under the same conditions completes a 2 × 4 design (Table 15). Adding exemplars to the fine-tuned model changed accuracy only slightly, mostly by small decreases (ToMBench −2.38 and −2.06 points under answer-only and rationale-CoT). The one larger mean change,

OpenToM −4.80 under answer-only, came with an SD of 8.49 across draws, and SocialIQa rose by +2.50 under format-only. In this setting, fine-tuning and few-shot prompting act as substitutes rather than complements.

Table 15. Fine-tuned model (seed 42) under training-free three-shot conditions versus zero-shot fine-tuned baseline. Each Δ is the mean over three independent exemplar draws (ex0–ex2); SD is the cross-exemplar-seed standard deviation. Accuracy in percentage points. Compare with base-model conditions in Table 13.

Dataset	Zero-Shot (Tuned)	Answer-Only Δ (SD)	Format-Only Δ (SD)	Rationale-CoT Δ (SD)
ToMBench	75.18	−2.38 (0.59)	−0.83 (0.66)	−2.06 (0.72)
OpenToM	59.50	−4.80 (8.49)	+0.27 (1.46)	−0.17 (0.96)
ToMi	83.00	−0.43 (1.21)	−0.23 (0.35)	−1.50 (1.78)
SocialIQa	61.70	+0.17 (0.38)	+2.50 (0.78)	+0.13 (0.95)
Hi-ToM	62.30	−0.80 (0.52)	−0.03 (0.72)	−0.10 (1.35)

4.7. Gold-Answer-Position Bias Analysis Across All Benchmarks

This study first identified a gold-answer-position effect for Hi-ToM alone and then extended the same analysis, a gold-answer-letter breakdown of the accuracy change plus a label-permutation test, to all five benchmarks and all five seeds. The permutation test shuffles the gold letters across items while holding each item's base and fine-tuned outcomes fixed. It asks how often the shuffled data produce a gap between the largest and smallest per-letter accuracy change at least as large as the observed one. We report the continuity-corrected $p = (k+1)/(n_{\text{perm}}+1)$ with $n_{\text{perm}} = 5000$. Per-letter breakdowns for every benchmark are released with the code (`ANALYSIS_position_by_gold_letter_*.csv`). These analyses were not part of the pre-specified confirmatory family and are reported as exploratory (Table 16).

Table 16. Analyses reported in this study, by when they were specified. Only the analysis specified before the original submission is confirmatory. The analyses that now carry the principal claims were specified in response to review; they are reported with Holm–Bonferroni correction within each seed where a significance test is used, and with between-seed intervals otherwise.

Analysis	Specified	Family
Base vs. fine-tuned McNemar test, five datasets, seed 42 (Table 7)	Before submission	Confirmatory
Five-seed replication and per-seed tests (Tables 5 and 11)	Before submission	Exploratory
Category- and ability-level breakdown (Tables 9 and 10)	Before submission	Exploratory
Training-free three-shot baselines, base and fine-tuned (Tables 13 and 15)	Before submission	Exploratory
Format-compliance check (Section 5.2)	Before submission	Exploratory
Story-disjoint split with external benchmarks (Table 12)	One seed before submission; five seeds in revision	Principal (review-specified)
Cyclic-permutation re-evaluation, both splits, five seeds each (Section 4.7)	In revision	Principal (review-specified)
Three-shot conditions under cyclic permutation (Table 14)	In revision	Principal (review-specified)
Gold-letter position test and stratum exclusion (Tables 17 and 18)	Seed 42 before submission; five seeds in revision	Exploratory
16-bit LoRA control (Section 4.8)	Before submission	Exploratory

The confirmatory family in Table 16 was specified before the original submission. The analyses that now carry the paper's principal claims, the five-seed story-disjoint replication and the cyclic-permutation re-evaluation of the fine-tuned and prompted models, were specified in response to review and are therefore not pre-specified. We label them as such rather than retrofitting them into the confirmatory family.

Table 17. Gold-answer-position bias, all five benchmarks and five seeds (item-level split). "Position gap" is the difference between the largest and smallest per-gold-letter accuracy change; the permutation test asks whether this gap is larger than expected if gold-answer position had no effect on which items flipped. The range is over the five seeds; "Seeds sig." counts seeds with $p < 0.05$.

Dataset	N	Gold Letters	Position Gap, Range (pp)	Seeds Sig.
ToMBench	842	A–D	33.6–45.3	5/5
ToMi	1000	A–D	27.1–35.3	5/5
OpenToM	1000	A–C	4.9–15.3	2/5
SocialIQa	1000	A–C	9.3–13.1	5/5
Hi-ToM	1000	A–E	19.7–34.3	5/5

The position gap is significant in all five seeds on four benchmarks and in two of five seeds on OpenToM. ToMBench and OpenToM mix option counts, so a permutation across all items can assign a letter that exists only in four-option items to a two-option item. We therefore repeated the test within each option count. On ToMBench the gap remains significant in all five seeds both among the 149 two-option items (33.1–47.1 points) and among the 693 four-option items (35.2–45.3 points). On OpenToM it is significant in four of five seeds among the 661 two-option location items (0.4–25.1 points) and in two of five seeds among the 339 three-option attitude items (10.8–14.6 points). The pattern first identified for Hi-ToM is therefore not benchmark-specific. Fine-tuning is associated with a shift in the model's option-selection tendency on every dataset in this study, including the two datasets (ToMBench, ToMi) whose aggregate accuracy improved. In seed 42 the raw prediction distributions show the mechanism. The fine-tuned model chooses D more often on ToMBench (9.50% to 17.46%), ToMi (16.3% to 26.5%) and Hi-ToM (17.5% to 28.1%), and B more often on SocialIQa (35.9% to 41.4%). On ToMBench and ToMi this brings the D rate close to the gold-D rate (17.81% and 25.1%). On Hi-ToM it moves the D rate past the gold-D rate (18.6%), and on SocialIQa it moves the B rate, already above the gold-B rate (33.4%), further above it.

As a first follow-up, for each benchmark and seed we excluded the gold-letter stratum contributing most to the position gap and recomputed the McNemar test on the remaining items (Table 18). This removes 11–60% of each sample. Because the excluded letter is chosen after seeing the data, the remaining items show a smaller effect even when the true effect is uniform. A non-significant residual therefore cannot be read as no effect.

Table 18. Residual accuracy change (item-level split) after excluding, separately for each seed, the gold-answer-letter stratum contributing most to the position gap. Residual change is computed on the remaining items; "Seeds survive" counts seeds whose residual McNemar test reached $p < 0.05$ with the same sign as the full-sample change.

Dataset	Excluded Letter	Residual Δ, Mean ± SD (Range), pp	Seeds Survive
ToMBench	D	+8.55 ± 0.76 (+7.80 to +9.68)	5/5
ToMi	D	+2.96 ± 2.39 (+0.13 to +6.14)	3/5
OpenToM	A, B or C	+0.15 ± 2.99 (−4.70 to +3.47)	1/5
SocialIQa	C (A in one seed)	−2.97 ± 0.27 (−3.21 to −2.50)	0/5
Hi-ToM	E (D in one seed)	+3.86 ± 3.03 (−1.35 to +5.94)	3/5

Stratum exclusion is therefore a description of where the change is concentrated, not a control.

Position-invariant re-evaluation by cyclic permutation. Neither stratum exclusion nor a macro average over gold letters removes a shift in the model's letter preference. Let w_L be the share of items whose gold answer is letter L and Δ_L the accuracy change on those items. The macro and raw accuracy changes then differ by $\sum_L (1/K - w_L)\,\Delta_L$, which is close to zero whenever the gold letters are close to uniform, however large the per-letter changes are. An earlier version of this manuscript used that macro average as a second control; we have withdrawn it. LLM accuracy on multiple-choice questions is known to depend on option order [24,25]. We instead re-evaluated the base model and both sets of five fine-tuned models, story-disjoint (Section 4.5) and item-level, under cyclic permutation of the answer options. Each item with K options was presented K times, with the options rotated so that the gold answer occupied each position exactly once; the relative order of the options was otherwise unchanged. Because every item is scored equally at every gold position, a fixed preference for a letter can neither raise nor lower the averaged score. A model that always answered D would score exactly $1/K$ on every item. We report three measures: lenient accuracy (the mean accuracy over the K rotations, our primary measure), strict accuracy (the share of items answered correctly in all K rotations), and consistency (the share of items for which the same option content was chosen in all K rotations). For the story-disjoint models, the unrotated presentation reproduced the reported predictions of seeds 43–46 on all 4837 items and those of seed 42 on all but 42 items (Section 4.9). Their original-order changes therefore differ slightly from Table 12. For the item-level models, the unrotated presentation reproduced the released predictions of the base and seed-42 models on all 4842 items. The adapters of the original seed 43–46 runs were not retained, so these seeds were re-evaluated with same-configuration replicate adapters, and their original-order changes differ slightly from Table 11. Table 19 reports the results.

Table 19. Cyclic-permutation re-evaluation under both splits: change from the base model, five seeds each, mean ± SD in percentage points, with the between-seed 95% t-interval for lenient accuracy. The story-disjoint ToMBench test split has 837 items and the item-level split 842; the external benchmarks use the same items under both. Original order is the unrotated presentation; lenient is the mean accuracy over all rotations; strict counts items correct in every rotation; consistent counts items for which the same option content was chosen in every rotation.

Dataset	Original Order	Lenient [95% Interval]	Strict	Consistent
Story-disjoint split				
ToMBench	+9.87 ± 1.55	+8.15 ± 0.56 [+7.46, +8.85]	+14.93 ± 2.59	+12.23 ± 4.29
ToMi	+13.12 ± 1.75	+14.19 ± 1.87 [+11.87, +16.50]	+35.62 ± 2.21	+39.52 ± 1.60
OpenToM	+0.56 ± 0.68	+0.67 ± 0.68 [−0.17, +1.52]	+0.72 ± 2.74	+1.04 ± 5.68
SocialIQa	−1.82 ± 1.44	−0.77 ± 1.44 [−2.56, +1.03]	−4.94 ± 3.85	−8.38 ± 4.68
Hi-ToM	+0.78 ± 1.41	+1.26 ± 1.12 [−0.13, +2.65]	−16.32 ± 3.48	−19.44 ± 3.65
Item-level split				
ToMBench	+14.37 ± 1.03	+14.01 ± 0.71 [+13.12, +14.89]	+24.04 ± 0.92	+17.43 ± 0.36
ToMi	+9.22 ± 1.98	+10.38 ± 1.73 [+8.24, +12.52]	+30.54 ± 3.57	+36.36 ± 3.24
OpenToM	−0.60 ± 1.97	+0.03 ± 0.90 [−1.09, +1.14]	−1.50 ± 2.62	−1.50 ± 4.06
SocialIQa	−5.70 ± 0.75	−4.14 ± 0.30 [−4.51, −3.77]	−7.12 ± 1.15	−6.72 ± 2.13
Hi-ToM	+1.32 ± 1.01	+2.16 ± 1.12 [+0.77, +3.54]	−9.12 ± 4.31	−11.30 ± 4.56

Under the story-disjoint split, the in-domain gain persists on all three measures (lenient +8.15 points). It comes mainly from the 683 four-option test items (+9.74 ± 0.83); on the 154 two-option items the lenient interval includes zero (+1.10 ± 1.43). The lenient gain is 1.7 points smaller than the original-order gain of the same models (+9.87), so a small part of the original-order gain reflected answer position. ToMi improves on all three measures

(lenient +14.19 points), no less than in the original order (+13.12). SocialIQa has no direction. The original-order decline (−1.82) shrinks to −0.77 lenient, with an interval including zero, while the strict (−4.94) and consistency (−8.38) intervals exclude zero. The fine-tuned models therefore answer SocialIQa less consistently across rotations without a measurable loss of accuracy averaged over rotations. OpenToM shows no direction on any measure, and neither subtask has a lenient interval excluding zero (attitude +1.69, interval [−0.47, +3.86]). On Hi-ToM the lenient interval includes zero, and the strict and consistency changes are negative.

Under the item-level split, the in-domain gain also persists (lenient +14.01 points). It comes mainly from the 693 four-option test items (+15.39 ± 0.20); on the 149 two-option items the lenient gain is smaller (+7.58 ± 3.65) and the strict interval includes zero. The ToMi gain also persists (+10.38 points) and is not smaller than the original-order gain (+9.22), so under both splits it is not produced by the concentration of gold-D items in the original order. The strict ToMi gains are not read as effect sizes, because the base model was highly sensitive to option order on ToMi (76.3% original-order accuracy against 48.0% strict). The SocialIQa decline persists on all three measures (−4.14 lenient); it is 1.6 points smaller than in the original order, so part of the original-order decline reflected answer position. OpenToM shows no direction on any measure; by subtask, the 339 three-option attitude items show a small gain in original-order, lenient and strict accuracy (+2.3 to +3.1 points, intervals excluding zero) but not in consistency, whereas the 661 two-option location items show no direction on any measure. On Hi-ToM, the lenient change is slightly positive but the strict and consistency changes are negative.

Under both splits, the Hi-ToM disagreement between measures follows from option E. ToMBench items have at most four options, so training never presented an option E. The item-level fine-tuned models chose option E in only 1.4–4.1% of rotations and the story-disjoint models in 0.18–2.00% (6.7% for the base model; 20% if uniform). An item counts as strictly correct only if it is answered correctly when its gold answer is E, and as consistent only if its chosen option is also chosen when that option is placed at E. The numbers of strictly correct and of consistent items therefore cannot exceed the number of E choices, and the fine-tuned models reached 82–90% (item-level) and 91–100% (story-disjoint) of that bound for strict accuracy. The Hi-ToM strict and consistency declines thus follow from the rarity of E choices, which is in line with option E never appearing in training. We did not test this cause directly, and a weaker avoidance of E also appears under three-shot prompting (Section 4.6). We therefore do not interpret the Hi-ToM changes as a change in higher-order ToM reasoning.

4.8. 16-Bit LoRA Control

Every fine-tuning run in this study used the same four-bit pre-quantized checkpoint. The effects attributed to LoRA adaptation therefore could not otherwise be separated from any interaction between quantization and adaptation. To address this, we trained a non-quantized 16-bit LoRA control with the same recipe (Qwen2.5-3B-Instruct, $r = 16$, $\alpha = 16$, seed 42) on the original item-level split and evaluated it on all five benchmarks (Table 20). The control loaded the checkpoint with automatic dtype selection, and we could not confirm from the preserved logs whether it ran in fp16 or bf16. Re-running its zero-shot ToMBench evaluation on an NVIDIA L4 reproduced neither format exactly (19 and 27 of 842 predictions differing). We therefore describe it as a non-quantized 16-bit control. The contrast tested is quantized versus non-quantized, and both formats fall on the non-quantized side. The control is a single run, and re-evaluation alone changes 2–3% of its predictions, which is small relative to the differences discussed below.

Table 20. Non-quantized 16-bit LoRA control versus 4-bit QLoRA (item-level split, seed 42, single run each). Base accuracy and Δ after fine-tuning, in percentage points; original-order accuracy, since this control was not re-evaluated under cyclic permutation. The base accuracies differ on four of five benchmarks because the two precisions produce different zero-shot outputs; on OpenToM, they match (63.40%) although 5.1% of item-level predictions differ.

	4-Bit QLoRA		16-Bit LoRA	
Dataset	**Base**	**Δ**	**Base**	**Δ**
ToMBench	61.52	+13.66	56.29	+20.78
OpenToM	63.40	−3.90	63.40	+0.10
ToMi	76.30	+6.70	84.20	+0.30
SocialIQa	67.70	−6.00	69.20	−5.20
Hi-ToM	62.40	−0.10	66.40	+5.50

The 16-bit control reproduces the in-domain gain (+20.78 points from a lower base of 56.29%, reaching a similar post-fine-tuning accuracy of about 75–77%) and the item-level SocialIQa decline (−5.20 points). Neither effect is therefore specific to training on a 4-bit quantized base. The ToMi, OpenToM and Hi-ToM changes differ between the two runs, but their zero-shot baselines also differ (the 16-bit ToMi baseline of 84.20% leaves little headroom). We therefore do not attribute these differences to quantization; a comparison with matched baselines would be needed.

4.9. Reproducibility of the Evaluation Pipeline

We assessed reproducibility at three levels. First, re-evaluating a fixed adapter is exactly reproducible. With the version-pinned environment released with our code, repeated evaluation of the seed-42 adapter that produced the released predictions, within one session and in separate sessions on the same GPU model (NVIDIA L4), reproduced all 4842 per-item predictions without a single difference, as did a re-evaluation after reloading the adapter (2000 items). This exact reproducibility is conditional on both the GPU model and the pinned library versions. The seed-42 story-disjoint adapter (Section 4.5) was originally evaluated in a session run before the library versions were pinned, and the GPU model of that session was not recorded. Re-evaluated under the pinned environment, it differed on 10 of its 837 ToMBench test items and on 32 of its 4000 external-benchmark items. Second, predictions are sensitive to the adapter loading path. Loading identical weights via `PeftModel.from_pretrained` rather than the Unsloth loader used to produce the released predictions changed 1.38% of items (67 of 4842). All affected items were near-ties, with a median top-1/top-2 logit margin of 0.12 (maximum 0.36) against 1.79 for unchanged items. The two loading paths produce identical quantization settings (NF4 with double quantization, bf16 compute dtype), identical LoRA weight dtype (fp32), and identical layer types; they differ only in how the forward pass is executed. We compared the answer-position logits of the two paths directly on the same inputs (the 67 changed items and 40 unchanged items). The median largest difference across the answer letters was 0.125 (maximum 0.625), and the mean absolute difference across the full vocabulary was about 0.04. These are differences of one or a few units of bf16 resolution at the magnitude of these logits. The top answer-letter logit ranged from 23.1 to 34.5 (105 of the 107 items in $[16, 32)$, where one bf16 unit is 0.125), and the largest per-item difference was a median of one such unit (maximum five). Such rounding-level differences can only change the prediction where the top two letters are nearly tied, which is what we observe. In this logit comparison, the median margin of the changed items was 0.125, one bf16 unit, and 31 of the 67 changed items (46%) were exact ties at bf16 resolution. Loading precision can matter more. Loading the non-quantized checkpoint in fp16 and in bf16 changed 2.3% and 3.2%

of zero-shot ToMBench predictions relative to the original control run (Section 4.8). That run was evaluated before the library versions were pinned, so this difference combines precision with environment. Replication should therefore match the loading path, the loading dtype, and the pinned environment. Third, re-training under identical seeds and configuration changed 3.30% of per-item predictions on average. At the aggregate level, 19 of 20 benchmark–seed cells reproduced within ±0.8 pp (mean signed deviation +0.03 pp, SD 0.65 pp, 90th percentile of absolute deviation 0.71 pp, maximum 2.10 pp), and every cell kept its sign. The single largest deviation (SocialIQa, seed 45) reflects a shift in the model's letter prior rather than an isolated perturbation. In that cell, 37 of the 59 changed items moved to option A (net inflow of 27 items), against 41–49% for the remaining seeds; this is the same seed whose most-affected gold-letter stratum in Section 4.7 was A rather than C. Aggregate conclusions are therefore stable under re-training, while per-item predictions are not. We recommend that replication of these results use the adapter loading path specified in our repository. These reproducibility measurements are reported as properties of the pipeline. They are not used as a significance threshold; for significance we rely on between-seed intervals throughout.

Per-epoch training logs for the released run were not preserved, and we cannot state which epoch checkpoint produced the released adapter. Four same-configuration replicate runs under the item-level split all selected the epoch-3 checkpoint by validation loss, and the four story-disjoint runs with retained logs all selected epoch 1. The adapters of the original item-level seed 43–46 runs were not saved either. The seed 43–46 item-level adapters we release are same-configuration replications and are labelled as such, whereas the five story-disjoint adapters are the ones that produced the reported story-disjoint results.

5. Discussion

5.1. Transfer Under Position Control and Prompting Baselines

Throughout this subsection, we apply one rule. For position-controlled results, we state a direction only when the five-seed 95% between-seed interval excludes zero on all three measures (lenient, strict, and consistent) with the same sign; where the measures disagree, we report the disagreement and state no direction. For original-order results, which have a single measure, we require its interval to exclude zero. These intervals reflect training randomness only, not the sampling of evaluation items, and the reproducibility measurements in Section 4.9 are reported as properties of the pipeline, not as a significance threshold.

Read on their own, the original-order multi-seed results in Section 4.5 would support a conventional selective-transfer story: a strong in-domain gain, transfer to a structurally similar task (ToMi), negative transfer to a task with different cognitive demands (SocialIQa), and no net effect on OpenToM and Hi-ToM. Section 4.7 shows why this reading needs a position control: fine-tuning shifted the model's preference among answer letters on every benchmark, so original-order accuracy changes mix any change in reasoning with a change in letter preference. The cyclic-permutation re-evaluation of both sets of models (Table 19) and the training-free baselines evaluated under the same permutation (Table 14) separate the five benchmarks as follows.

First, the in-domain gain survives every check applied, including a story-disjoint split and position control applied together (lenient +8.15 points, interval [+7.46, +8.85], positive on all three measures, against +14.01 under the item-level split, mainly on four-option items under both). It is not reproduced by any training-free three-shot condition (at most +3.03 points, on the item-level test split). We attribute it to fine-tuning. Because ToMBench was also the training source, it reflects learning of this benchmark's tasks, and we do not claim from it alone that a general ToM competence was acquired.

Second, the ToMi gain survives position control under both splits (lenient +14.19 points story-disjoint, +10.38 item-level), so it is not explained by the concentration of gold-D items in the original order. However, the same position-controlled evaluation shows that content-free format exemplars raise ToMi lenient accuracy by +7.23 points and rationale-style exemplars by +12.96 points without any weight update. ToMi accuracy in our multiple-choice construction is therefore highly sensitive to exposure to the answer format and to a generic reasoning template, in line with earlier evidence that gains on ToMi need not reflect generalizable belief tracking [23]. The fine-tuned lenient gain is also concentrated in the 72.0% of items in which the gold answer is the only option mentioned in the story (Section 4.6). We do not interpret the fine-tuned ToMi gain as transfer of ToM reasoning. We also cannot estimate how much of it exceeds format exposure, because the fine-tuned and prompted figures vary over different sources: the story-disjoint gain exceeds the format-only gain, but its interval [+11.87, +16.50] contains the rationale-exemplar gain.

Third, the SocialIQa decline survives position control on all three measures under the item-level split (lenient −4.14 points), and no training-free condition reproduces it (all slightly positive). Under the story-disjoint split it does not: the lenient change is −0.77 points with an interval including zero, although the strict and consistency changes are negative, so under the rule, above we state no direction. The decline is therefore tied to the weight updates of item-level fine-tuning rather than to fine-tuning on ToMBench in general; Section 5.5 explains why we do not attribute it to the split alone. Section 5.2 discusses possible causes, none of which was isolated experimentally.

Fourth, OpenToM shows no direction on any measure under either split. On Hi-ToM the strict and consistency changes are negative under both splits, whereas the lenient interval excludes zero under the item-level split (+2.16 points) and includes zero under the story-disjoint split. Under the rule above we therefore state no direction. The disagreement follows from the fine-tuned models rarely choosing option E, an avoidance that also appears under three-shot prompting. We make no transfer claim for either benchmark. An earlier version of this manuscript reported a residual Hi-ToM gain after excluding gold-E items as a hypothesis; the position-invariant evaluation does not support it, and we have withdrawn it.

These results bear on how strong a claim the term "selective transfer" should be read to make. We use it only as a descriptive label for which effects survive a position-invariant evaluation and a prompting baseline under the same control, not as a claim about the cognitive architecture of the model's ToM competence. The effects that survive are compatible both with learning of ToM sub-skills and with learning of the training benchmark's response format and surface patterns. This study did not design an intervention that distinguishes the two, and the ToMi results show that the second account can produce a large gain on an external benchmark.

5.2. Analysis of the Causes of Performance Decrease

This subsection concerns the one decline that survives position control, SocialIQa, which it does only under the item-level split, and the question of why fine-tuning shifts letter preference. Neither was isolated experimentally, so both accounts are interpretations.

First, task-structure mismatch may contribute. ToMBench and ToMi emphasize explicit state tracking (who moved what, and who did not see it), whereas SocialIQa requires commonsense reasoning about motivations, emotions, and consequences. Training concentrated on belief tracking may therefore optimize a reasoning format that does not transfer [13]. This account is consistent with the item-level SocialIQa decline persisting under cyclic permutation (−4.14 points) and not being produced by prompting.

Second, training-distribution bias may also contribute. The base model answered 141 SocialIQa items correctly that the fine-tuned model missed (Table 7). Part of this count reflects answer position, since the position-controlled decline (−4.14 points) is smaller than the original-order decline. Whether forgetting occurs under parameter-efficient tuning depends on factors such as adapter rank that we did not vary [33]; Section 5.5 explains why we do not describe this decline as catastrophic forgetting. LoRA has been reported to forget less than full fine-tuning outside its target domain [34]. This makes a targeted, task-specific effect a more plausible reading of the SocialIQa decline than incidental damage, and it identifies adapter rank as a variable for future controlled experiments.

A third, more mundane explanation is that part of the ToMBench gain reflects improved output-format compliance rather than reasoning. This is distinct from response-position bias: format compliance concerns whether a valid letter can be extracted at all, position bias concerns which letter is preferred. Strict `[[X]]` compliance rose from 96.56% to 100% on ToMBench, from 96.30% to 100% on Hi-ToM, from 96.70% to 100% on SocialIQa, from 98.70% to 100% on OpenToM, and from 99.80% to 99.90% on ToMi. At most 3.7% of items relied on the fallback rule before tuning, far too few to account for the in-domain gain, and on SocialIQa accuracy fell although every response was parsed.

As a direct check (Table 16), we restricted each dataset to items on which both models gave a strictly formatted response (96–99.7% of each sample) and recomputed the change. The changes were ToMBench +13.41 (full sample +13.66), ToMi +6.52 (+6.70), OpenToM −3.95 (−3.90) and SocialIQa −6.31 (−6.00), all $p < 0.002$, and Hi-ToM +0.42 ($p = 0.84$). No benchmark changes sign or significance, so parsing-rule artifacts do not drive the reported changes.

The training-free conditions in Section 4.6 give a further check on format learning. On ToMBench, answer-only three-shot exemplars from the training split recovered +2.30 of the +13.66-point seed-42 gain, and content-free format-only exemplars recovered +1.66. Under cyclic permutation, no condition exceeded +3.03 points, against +14.01 for fine-tuning. Output-format exposure therefore accounts for at most a small share of the in-domain gain.

The same account may explain why Emotion was the only ToMBench category that did not improve (57.26% to 56.45%) and why Desire did not change. Emotion inference draws on social-commonsense knowledge closer to SocialIQa than to state tracking; we did not test this account.

Table 21 lists SocialIQa items that the base model answered correctly and the fine-tuned model (seed 42) did not. In these items, the fine-tuned model tends to choose a literal, action-level reading over the inference about motivation or emotion that the gold answer requires.

Table 21. Representative SocialIQa items on which the base model was correct and the fine-tuned model was incorrect (base → tuned error), illustrating the literal/state-tracking-style substitution pattern discussed above. Hi-ToM examples are not included in this table because the gold-answer-position analysis in Section 4.7 shows Hi-ToM's base-to-tuned errors as a whole track the letter position of the gold answer rather than belief-order depth, and a small number of hand-picked individual-item examples would risk being read as evidence of an order effect that the full item-level analysis does not support; the position-driven pattern is instead reported quantitatively in Tables 17 and 18.

Dataset	Question (Abbreviated)	Gold	Base	Tuned
SocialIQa	Jan gave birth at a party—what will happen to *Others*? (gold: be happy)	B	B	A
SocialIQa	Alex personally fitted each teammate's watch—how would you describe Alex? (gold: thoughtful person)	C	C	B
SocialIQa	A preschool teacher traced a student's fingers—why did she do this? (gold: make artwork)	C	C	B

For Hi-ToM, Section 4.7 shows that the fine-tuned models avoid option E, as do models prompted with any three-shot prefix. Option-position preferences of this kind are well documented in multiple-choice LLM evaluation [25]. An earlier version of this manuscript treated a residual Hi-ToM gain after excluding gold-E items as a hypothesis of improved belief tracking. Under cyclic permutation, Hi-ToM shows no direction under either split, so we have withdrawn that hypothesis.

5.3. Reproducibility and Ethical Considerations

For reproducibility, random seed 42 was used for all data splits, and the ToMBench test split was not used during training or checkpoint selection. The experiments were conducted in a Google Colab environment on an NVIDIA L4 GPU. Considering computational cost, the external evaluation sets were limited to approximately 1000 samples each, and the same evaluation prompts and answer-extraction rules were applied to both the base model and the fine-tuned model. The item-level correct/incorrect results used for statistical analysis and the code for McNemar and bootstrap calculations are released together for reproducibility.

In terms of data use, ToMBench is released as an evaluation benchmark, and its repository states: "Use ToMBench for evaluation purposes only. Avoid using ToMBench data for training in any circumstance to prevent data contamination, which can lead to inflated performance." This study departed from that evaluation-only use. We used 56.2% of ToMBench items for training and a further 14.4% for validation (a combined development split of 70.6%) and reserved the remaining 29.4% as a sealed test split (see Section 5.5 for a full discussion of this limitation). The ToMBench results in this study should therefore be interpreted as controlled internal experimental results rather than official leaderboard scores. We chose this design deliberately, and the order of the reasoning matters. The research question (RQ2) is whether a fine-tuned ability transfers beyond the training distribution, so we needed external benchmarks that contribute no training signal at all. We therefore used one benchmark as the training source so that the other four could be kept purely for evaluation, rather than training on ToMBench in order to evaluate on it. Among the five benchmarks available to us, ToMBench was the natural training source. It is the only one whose items are annotated with an ability taxonomy (31 abilities of the ATOMS framework under six categories). This taxonomy allowed a split stratified by ability, so that no ability disappeared from the held-out set, and an ability- and category-level analysis of the result. We did not survey all ToM resources and do not claim that no alternative existed. To keep the in-domain measurement interpretable, the test split was sealed before training, and the headline in-domain result uses a story-disjoint split with five seeds (Section 4.5). Item-level overlap between splits was checked (Section 3.2), and the reproducibility of the evaluation is reported (Section 4.9). Training on an independent ToM corpus and keeping ToMBench purely for evaluation, as its authors recommend, remains the stronger design and is a direction for future work. The ToMBench scores in this study were calculated only on sealed held-out splits: the 29.4% item-level split stratified by ATOMS ability and the 837-item story-disjoint test split. It is not directly comparable with model scores evaluated on the full benchmark under its intended evaluation-only protocol. OpenToM, ToMi, SocialIQa, and Hi-ToM were not used for training and were used only for evaluation. In addition, this study did not collect new data containing real users or sensitive personal information, and it is limited to performance analysis using public benchmarks and public models. We also note a community-level consideration beyond this study's own protocol. As fine-tuning on nominally evaluation-only benchmarks becomes more common in the field, the remaining held-out evaluation value of those benchmarks is gradually eroded. Alongside our own disclosed departure from ToMBench's intended evaluation-only use,

we flag this as an issue for benchmark maintainers and the wider community to weigh, not merely as a comparability caveat specific to this manuscript.

5.4. Practical Implications

A QLoRA adapter fine-tuned on a single, narrowly scoped ToM resource should not be assumed to improve, or even preserve, ToM-adjacent performance. An external accuracy gain seen only in original-order evaluation is not evidence of transfer. Users of such adapters should apply the checks in Table 22. Because one training run of this recipe takes about 15 min on a single NVIDIA L4 (measured in our story-disjoint replications), these checks are inexpensive.

For completeness, we measured single-item inference on 100 ToMBench test items on an NVIDIA L4 (24 GB), a datacenter GPU; the figures describe that setting only. The base model averaged 350.8 ms per item (median 271.3 ms, 95th percentile 688.7 ms; 2.84 items/s) with 5.38 GB peak VRAM, and the fine-tuned model 416.2 ms (median 414.8 ms, 95th percentile 426.0 ms; 2.39 items/s) with 5.56 GB. The adapter thus adds 0.18 GB of VRAM and an 18.6% increase in mean latency; equivalently, a 15.6% decrease in throughput. The fine-tuned model's latency distribution is much tighter (p95/median 1.03 against 2.54), consistent with its near-elimination of non-conforming outputs that need the fallback extraction rule (Section 5.2).

We summarize the practical lesson of this study as a reusable validation checklist (Table 22). It gives practitioners a concrete procedure rather than a qualitative recommendation embedded in prose.

Table 22. Validation checklist for a QLoRA adapter fine-tuned on a single, narrowly scoped ToM resource, derived from this study's findings (Sections 4.7 and 5.1). Each step is inexpensive relative to a single training run of the recipe evaluated here.

Step	What to Check
1. In-domain gain	Confirm the held-out accuracy gain with a paired significance test (e.g., McNemar) on a sealed test split never used for training or checkpoint selection.
2. Multi-seed replication	Repeat fine-tuning across at least 3–5 training seeds; do not report a single-seed result as the headline finding, and check whether the seed used as primary is unusually strong or weak relative to the others (Section 4.5).
3. Cross-benchmark check	Evaluate on at least one external benchmark with a structurally different task demand from the fine-tuning data, not only benchmarks similar to it, and evaluate a training-free prompting baseline (including content-free format exemplars) under the same position control (Section 4.6).
4. Response-position bias check	For each benchmark in Step 3, compute the accuracy change broken down by gold-answer letter and run a permutation test for a position-driven gap (Section 4.7); a significant gap on a benchmark whose aggregate accuracy improved is a warning sign, not a confirmation.
5. Position-invariant evaluation	Re-evaluate every model under cyclic permutation of the answer options, so that each item's gold answer occupies every position once, and report lenient (mean over rotations) and strict (all rotations correct) accuracy (Section 4.7); stratum exclusion or a macro average over gold letters does not remove a shift in letter preference.
6. Format-compliance check	Verify that any accuracy or format-compliance improvement is not driven by parsing-rule artifacts, by restricting to items with strictly-formatted output on both conditions before and after fine-tuning (Section 5.2).

5.5. Limitations

This study involved the following limitations. First, a single base model (Qwen2.5-3B-Instruct) and a single fine-tuning setting (QLoRA, $r = 16$) were used, so the results may not hold for other architectures, scales, or adapter ranks. Five-seed replication varied only training randomness; it does not quantify variance from the data split or from the 1000-item external subsamples, and seed stability is necessary but not sufficient for a transfer claim.

Second, ToMBench's repository asks users not to train on its data (Section 5.3). We used 70.6% of its items for training and validation, so our ToMBench score is a controlled internal measurement, not comparable with official results. ToM benchmark scores in general may also not reflect a model's actual ToM ability [35]. The item-level split let about 60% of test items share a source story with a training item (55.1% of stories are reused, covering 78.3% of items). The story-disjoint replication (Section 4.5) lowered the in-domain gain from +14.32 to +9.77 points (from +14.01 to +8.15 under cyclic permutation); this is an association between the two splits rather than a causal decomposition, because the splits also differ in item composition and in amount of training (the story-disjoint runs with retained logs selected epoch 1, the item-level replicate runs epoch 3). For the same reason, we do not attribute the absence of a SocialIQa decline under the story-disjoint split to the split alone.

Third, the external benchmarks were subsampled at random to 1000 items without stratification, and their multiple-choice construction (Table 4) is partly our design, for example the distractors sampled for ToMi, which leave the gold answer as the only option mentioned in the story in 72.0% of items. Cyclic permutation controls answer position at evaluation but does not control these properties, and we did not randomize option order during training. The ToMi results show that such a construction can reward format exposure. Stratum exclusion, which removes 11–60% of a sample after seeing the data, was not used as a basis for conclusions.

Fourth, the analysis of transfer is observational. We did not design interventions that separate task-structure differences, surface prompt-format differences, and training-distribution bias, and we did not measure non-ToM capabilities, so we do not describe the SocialIQa decline as catastrophic forgetting, which implies a broad capability loss.

Fifth, we cannot rule out pretraining contamination of any of the five benchmarks [36]; procedurally generated or conversational benchmarks such as BigToM [37] and FANToM [38] would reduce this risk in future work.

Sixth, the latency measurements (Section 5.4) describe an NVIDIA L4 datacenter GPU only and should not be extrapolated to edge devices. The 16-bit LoRA control (Section 4.8) differs from the 4-bit runs in its zero-shot baseline, so it supports only the in-domain gain and the item-level SocialIQa decline as independent of 4-bit quantization.

6. Conclusions

This study applied QLoRA fine-tuning to Qwen2.5-3B-Instruct to examine whether a small language model can learn ToM reasoning under limited-resource conditions and whether apparent transfer to external benchmarks survives control for response-position bias. We stratified 2860 ToMBench items by ATOMS ability into 1607 training, 411 validation, and 842 test items. LoRA adapters were trained on the 4-bit pre-quantized model, and the checkpoint with the lowest validation loss was selected; the test split was not used for training or checkpoint selection. All results come from a single, item-identity-verified evaluation pipeline (Section 3.4).

Because fine-tuning shifted the model's preference among answer letters on every benchmark, we evaluated every model under cyclic permutation of the answer options. Under a story-disjoint split and this position-invariant evaluation, ToMBench held-out accuracy rose by +8.15 points across five seeds (+14.01 under the item-level split) and ToMi accuracy by +14.19 points, whereas SocialIQa, OpenToM, and Hi-ToM showed no direction. A SocialIQa decline appeared under the item-level split (−4.14) but not under the story-disjoint split. Training-free three-shot prompting under the same permutation produced a ToMi gain even with content-free format exemplars (+7.23), but reproduced neither the in-domain gain nor the item-level SocialIQa decline.

The conclusion of this study is therefore narrower than the original-order results suggest. QLoRA fine-tuning improves ToMBench performance under a story-disjoint split and position control applied together, and prompting does not reproduce the gain, although the gain reflects learning of the training benchmark's tasks rather than demonstrated general ToM competence. No external benchmark shows a position-controlled effect that is specific to fine-tuning and present under both splits: the ToMi gain is produced by prompting as well, so we do not interpret it as transfer of ToM reasoning, and the SocialIQa decline appears only under the item-level split. Methodologically, cross-benchmark transfer claims for ToM fine-tuning on multiple-choice benchmarks require a position-invariant evaluation, such as cyclic permutation, and a prompting baseline evaluated under the same control; stratum exclusion and macro averaging over gold letters do not provide such a control. Future work should test these conclusions on other model families and PEFT configurations, train with randomized option order, and use benchmarks whose construction does not reward format exposure.

Two further checks complete this picture. Adding three-shot exemplars to an already fine-tuned model yields near-zero or slightly negative changes (Section 4.6, Table 15), consistent with fine-tuning and few-shot prompting acting as substitutes in this setting. A non-quantized 16-bit LoRA control (Section 4.8, Table 20) reproduces the in-domain gain and the item-level SocialIQa decline without 4-bit quantization of the base model.

Author Contributions: Conceptualization, J.-H.H.; methodology, J.-H.H.; software, J.-H.H.; validation, J.-H.H.; formal analysis, J.-H.H.; investigation, J.-H.H.; data curation, J.-H.H.; writing—original draft preparation, J.-H.H.; writing—review and editing, J.-H.H. and S.-H.Y.; visualization, J.-H.H.; supervision, S.-H.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The source code, evaluation and training scripts, dataset split indices (the story-disjoint split is rebuilt deterministically by the released code), and per-item prediction results underlying every result in this manuscript are available on GitHub at https://github.com/NextMina1131/tom-slm-qlora (accessed on 29 October 2026), commit `7231116`; the commit cited in the original submission (`e3f155c`) remains in the repository history, and the changes since then are listed in the repository's `CHANGELOG.md`. Each per-item file carries a content-based item identifier and a prompt hash for every row, and the released verification scripts and their output logs check these identifiers across all result files. A version-pinned environment file (`requirements.lock.txt`) is included. The per-item files contain predictions and scoring fields (whether an `[[X]]` marker was found, which extraction rule produced the letter, and the generation length) but not the raw generations, because chain-of-thought outputs quote benchmark text verbatim; raw generations are available from the corresponding author on request. The QLoRA adapters are available on the Hugging Face Hub at https://huggingface.co/nextmina/qwen2.5-3b-tombench-qlora (accessed on 29 October 2026): the item-level seed-42 adapter that produced the released predictions, the item-level seed 43–46 adapters as same-configuration replications (the original seed 43–46 adapters were not

saved), and the five story-disjoint adapters that produced the story-disjoint results. The model card maps each adapter to the tables it produced, warns that the adapters were trained on ToMBench and must not be evaluated on it, and records that an adapter from a different, earlier run, uploaded before 18 October 2026, was replaced. Benchmark text is not redistributed; the benchmarks were used under their licenses (ToMBench, MIT, with an evaluation-only request by its authors; ToMi and OpenToM, CC BY-NC 4.0; SocialIQa, CC BY 4.0; Hi-ToM, Apache-2.0) and should be obtained from their original sources. The version of the code and results underlying this manuscript (commit `7231116`) is archived at Zenodo, https://doi.org/10.5281/zenodo.23880042.

Acknowledgments: During the preparation of this manuscript, the authors used a generative AI assistant (Claude, Anthropic; specifically Claude Opus 4.7 and Opus 4.8 from April to July 2026, Claude Fable 5 briefly on 13 July 2026, Claude Sonnet 5 from 21 July 2026 to 13 October 2026 (with Claude Sonnet 4.6 briefly on 7 September 2026), and Claude Opus 5.5 from 13 October 2026 during the peer-review revision, all accessed via the Claude Code CLI) to help draft and edit portions of the manuscript text and to assist with the post hoc analyses reported in Section 5, specifically Section 5.5; the full scope of this use, the tool's role, and the authors' independent verification procedure are disclosed in Section 3.5. The authors have reviewed, verified, and take full responsibility for all content, analyses, and conclusions in this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Premack, D.; Woodruff, G. Does the chimpanzee have a theory of mind? *Behav. Brain Sci*. **1978**, *1*, 515–526. [CrossRef]
2. Baron-Cohen, S.; Leslie, A.M.; Frith, U. Does the autistic child have a "theory of mind"? *Cognition* **1985**, *21*, 37–46. [CrossRef] [PubMed]
3. Van Duijn, M.; Van Dijk, B.; Kouwenhoven, T.; De Valk, W.; Spruit, M.; van der Putten, P. Theory of mind in large language models: Examining performance of 11 state-of-the-art models vs. children aged 7–10 on advanced tests. In Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL), Singapore, 6–7 December 2023; pp. 389–402.
4. Kosinski, M. Evaluating large language models in theory of mind tasks. *Proc. Natl. Acad. Sci. USA* **2024**, *121*, e2405460121. [CrossRef] [PubMed]
5. Chen, Z.; Wu, J.; Zhou, J.; Wen, B.; Bi, G.; Jiang, G.; Cao, Y.; Hu, M.; Lai, Y.; Xiong, Z.; et al. ToMBench: Benchmarking theory of mind in large language models. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, Bangkok, Thailand, 11–16 August 2024; pp. 15959–15983. [CrossRef]
6. Subramanian, S.; Elango, V.; Gungor, M. Small language models can still pack a punch: A survey. *arXiv* **2025**, arXiv:2501.05465.
7. Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q.V.; Zhou, D. Chain-of-thought prompting elicits reasoning in large language models. *Adv. Neural Inf. Process. Syst*. **2022**, *35*, 24824–24837. [CrossRef]
8. Shim, J.; Shin, D.H.; Kim, J. CoT harms performance of rather smaller language models. In Proceedings of the 1st World Congress on Industrial Engineering and Operations Management, Detroit, MI, USA, 9–11 October 2024; pp. 504–508. [CrossRef]
9. Wang, Y.; Kordi, Y.; Mishra, S.; Liu, A.; Smith, N.A.; Khashabi, D.; Hajishirzi, H. Self-Instruct: Aligning language models with self-generated instructions. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, Toronto, ON, Canada, 9–14 July 2023; pp. 13484–13508.
10. Zhou, C.; Liu, P.; Xu, P.; Iyer, S.; Sun, J.; Mao, Y.; Ma, X.; Efrat, A.; Yu, P.; Yu, L.; et al. LIMA: Less is more for alignment. *Adv. Neural Inf. Process. Syst*. **2023**, *36*, 55006–55021. [CrossRef]
11. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. LoRA: Low-rank adaptation of large language models. In Proceedings of the 10th International Conference on Learning Representations (ICLR 2022), Virtual, 25–29 April 2022.
12. Dettmers, T.; Pagnoni, A.; Holtzman, A.; Zettlemoyer, L. QLoRA: Efficient finetuning of quantized LLMs. *Adv. Neural Inf. Process. Syst*. **2023**, *36*, 10088–10115. [CrossRef]
13. Lu, Y.-L.; Zhang, C.; Song, J.; Fan, L.; Wang, W. Do theory of mind benchmarks need explicit human-like reasoning in language models? *arXiv* **2025**, arXiv:2504.01698.
14. Sarangi, S.; Salam, H. Small LLMs do not learn a generalizable theory of mind via reinforcement learning. *arXiv* **2025**, arXiv:2507.15788.
15. Li, Y.; Yue, X.; Xu, Z.; Jiang, F.; Niu, L.; Lin, B.Y.; Ramasubramanian, B.; Poovendran, R. Small models struggle to learn from strong reasoners. In *Findings of the Association for Computational Linguistics: ACL 2025*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2025; pp. 25366–25394.

16. Chen, R.; Jiang, W.; Qin, C.; Tan, C. Theory of Mind in Large Language Models: Assessment and Enhancement. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics, Vienna, Austria, 27 July–1 August 2025; pp. 31539–31558.
17. Xu, H.; Zhao, R.; Zhu, L.; Du, J.; He, Y. OpenToM: A comprehensive benchmark for evaluating theory-of-mind reasoning capabilities of large language models. *arXiv* **2024**, arXiv:2402.06044.
18. Le, M.; Boureau, Y.-L.; Nickel, M. Revisiting the evaluation of theory of mind through question answering. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2019; pp. 5872–5877.
19. Sap, M.; Rashkin, H.; Chen, D.; Le Bras, R.; Choi, Y. Social IQa: Commonsense reasoning about social interactions. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2019; pp. 4463–4473.
20. Wu, Y.; He, Y.; Jia, Y.; Mihalcea, R.; Chen, Y.; Deng, N. Hi-ToM: A benchmark for evaluating higher-order theory of mind reasoning in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2023; pp. 10691–10706.
21. Sap, M.; Le Bras, R.; Fried, D.; Choi, Y. Neural theory-of-mind? On the limits of social intelligence in large LMs. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Abu Dhabi, United Arab Emirates, 7–11 December 2022; pp. 3762–3780.
22. Shapira, N.; Levy, M.; Alavi, S.H.; Zhou, X.; Choi, Y.; Goldberg, Y.; Sap, M.; Shwartz, V. Clever Hans or neural theory of mind? Stress testing social reasoning in large language models. In Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics, St. Julian's, Malta, 17–22 March 2024; pp. 2257–2273.
23. Sclar, M.; Kumar, S.; West, P.; Suhr, A.; Choi, Y.; Tsvetkov, Y. Minding language models' (lack of) theory of mind: A plug-and-play multi-character belief tracker. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, Toronto, ON, Canada, 9–14 July 2023; pp. 13960–13980. [CrossRef]
24. Pezeshkpour, P.; Hruschka, E. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 2006–2017. [CrossRef]
25. Zheng, C.; Zhou, H.; Meng, F.; Zhou, J.; Huang, M. Large language models are not robust multiple choice selectors. In Proceedings of the 12th International Conference on Learning Representations (ICLR 2024), Vienna, Austria, 7–11 May 2024.
26. Lore, N.; Ilami, S.; Heydari, B. Large model strategic thinking, small model efficiency: Transferring theory of mind in large language models. *arXiv* **2024**, arXiv:2408.05241.
27. Moghaddam, S.R.; Honey, C.J. Boosting theory-of-mind performance in large language models via prompting. *arXiv* **2023**, arXiv:2304.11490.
28. Wilf, A.; Lee, S.S.; Liang, P.P.; Morency, L.P. Think Twice: Perspective-taking improves large language models' theory-of-mind capabilities. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, Bangkok, Thailand, 11–16 August 2024; pp. 8292–8308.
29. Qwen Team. Qwen2.5 technical report. *arXiv* **2024**, arXiv:2412.15115.
30. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
31. Miller, E. Adding error bars to evals: A statistical approach to language model evaluations. *arXiv* **2024**, arXiv:2411.00640.
32. Zhuo, J.; Zhang, S.; Fang, X.; Duan, H.; Lin, D.; Chen, K. ProSA: Assessing and understanding the prompt sensitivity of LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 1950–1976. [CrossRef]
33. Ren, W.; Li, X.; Wang, L.; Zhao, T.; Qin, W. Analyzing and reducing catastrophic forgetting in parameter efficient tuning. *arXiv* **2024**, arXiv:2402.18865.
34. Biderman, D.; Portes, J.; Ortiz, J.J.G.; Paul, M.; Greengard, P.; Jennings, C.; King, D.; Havens, S.; Chiley, V.; Frankle, J.; et al. LoRA learns less and forgets less. *Trans. Mach. Learn. Res.* **2024**. Available online: https://openreview.net/forum?id=aloEru2qCG (accessed on 29 October 2026).
35. Riemer, M.; Ashktorab, Z.; Bouneffouf, D.; Das, P.; Liu, M.; Weisz, J.D.; Campbell, M. Position: Theory of mind benchmarks are broken for large language models. In Proceedings of the 42nd International Conference on Machine Learning (ICML 2025), Vancouver, BC, Canada, 13–19 July 2025.
36. Xu, C.; Guan, S.; Greene, D.; Kechadi, M.-T. Benchmark data contamination of large language models: A survey. *arXiv* **2024**, arXiv:2406.04244.

37. Gandhi, K.; Fränken, J.-P.; Gerstenberg, T.; Goodman, N.D. Understanding social reasoning in language models with language models. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 13518–13529. [CrossRef]
38. Kim, H.; Sclar, M.; Zhou, X.; Le Bras, R.; Kim, G.; Choi, Y.; Sap, M. FANToM: A benchmark for stress-testing machine theory of mind in interactions. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Singapore, 6–10 December 2023; pp. 14397–14413.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.