QM-ToT: A Medical Tree of Thoughts Reasoning Framework for Quantized Model

Zongxian Yang[1[0009-0002-6230-4301]] and Jiayu Qian[1[0009-0007-7148-2843]] and Kay Chen Tan[2[0000-0002-6802-2463]] and Hau-San Wong[3[0000-0002-1530-7529]] and Yulong Chen[1[0000-0003-1821-3330]] and Haoyu Zhang[3[0009-0004-0616-0096]] and Zhi-An Huang[1[0000-0001-9974-148X]]*

[1] City University of Hong Kong(Dongguan), Dongguan, China
[2] Hong Kong Polytechnic University, Hong Kong, China
[3] City University of Hong Kong, Hong Kong, China
huang.za@cityu-dg.edu.cn

Abstract. Large language models (LLMs) have achieved substantial progress in biomedical question answering, but resource-constrained medical deployment often requires low-bit quantization, which amplifies performance degradation on complex clinical reasoning tasks. We propose Quantized Medical Tree of Thought (QM-ToT), a structured tree-ofthought framework that decomposes medical problems into hierarchical reasoning paths and selects answers through dual-evaluation feedback. On MedQA-USMLE, QM-ToT improves INT4-quantized LLaMA2-70b from 34% to 50.25% and LLaMA-3.1-8b from 58.77% to 69.49%. We further propose Reflection-ToT, a QM-ToT-based data distillation method. Using 1,000 questions, Reflection-ToT reaches 66.44% accuracy on LLaMA3.1- 8B, exceeding QwQ-based CoT distillation with 10,178 questions (65.01%) and a QwQ-based matched-pair setting with 1,514 pairs (64.73%). This work demonstrates the potential of structured tree-based reasoning with evaluator feedback for improving biomedical question answering under low-precision deployment, while also highlighting the need to account for evaluator dependence and inference cost.

Keywords: Model Quantization, Medical Question Answering, Healthcare Application, Large Language Model

1 Introduction

Large language models (LLMs) have advanced many AI tasks [12, 15, 50], yet specialized biomedical benchmarks still expose important gaps [39]. The first challenge is reasoning complexity: medical QA requires terminology understanding, clinically grounded inference, and answers consistent with professional standards; even strong models such as OpenAI-o1[1] benefit from specialized reasoning strategies [45]. The second challenge is privacy. Local deployment can reduce risks from sensitive medical

* Corresponding author.
[1] https://openai.com/o1/

data [7, 23], and quantization lowers memory and cost [26], but it can degrade output quality [20]. Fig. 1 shows this FP16-to-INT4 drop on MedQA-USMLE [17], motivating methods that preserve robustness under efficient deployment. QM-ToT addresses these challenges by decomposing medical problem-solving into tree-structured paths and using evaluator feedback on factual accuracy and logical validity (Fig. 2). On MedQA-USMLE, it improves over chain-of-thought self-consistency (CoT-SC): Qwen2.5-72b rises from 70.25% to 74.25%, LLaMA-2-70b from 27.75% to 50.25%, and LLaMA-3.1-8b from 59.19% to 69.49%. We also stratify questions into easy,medium, and hard levels to analyze where ToT and evaluator feedback help.

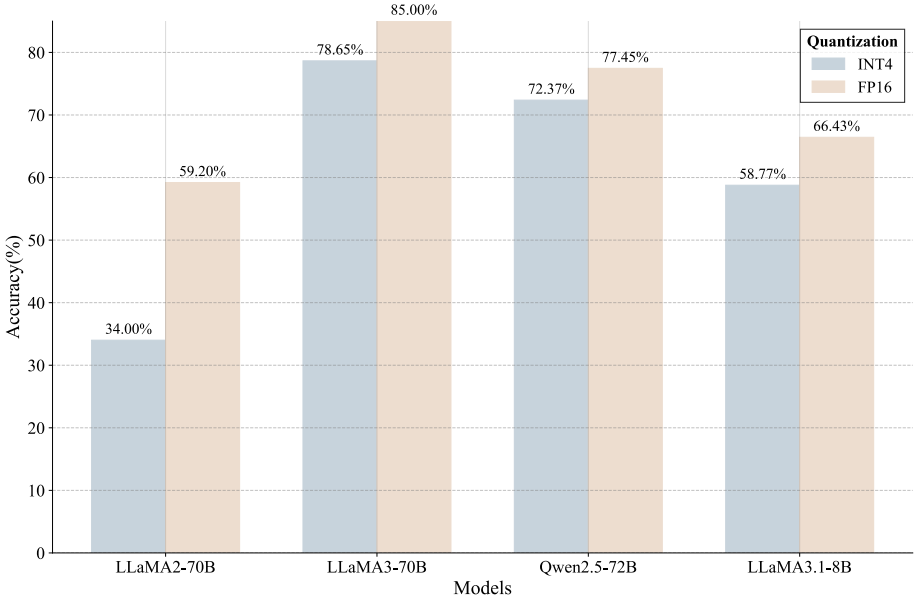

Fig. 1. FP16 vs INT4 quantization performance comparison. Performance gap between FP16 and INT4 quantization for LLaMA3-70b, LLaMA2-70b, Qwen2.5-72b, and LLaMA3.1-8b models on the MedQA-USMLE dataset.

We further use ToT to construct synthetic medical reasoning data through Reflection-ToT (Fig. 3), which refines short ToT-generated CoT into reflectionaugmented long CoT for student training. Reflection-ToT achieves 66.44% accuracy using 1,000 questions and 1,514 preference pairs, outperforming QwQ-based CoT distillation with 10,178 questions (65.01%) and a QwQ-based matched-pair setting with 1,514 pairs (64.73%).

We summarize our contributions as follows:

- QM-ToT combines a path-based ToT planner with a dual-evaluator module, achieving an average accuracy gain of 9.14% over CoT-SC on INT4-quantized models.
- Reflection-ToT converts short CoT into long o1-style CoT with reflection. Using only 9.8% of the questions, it improves over traditional CoT distillation from the o1-aligned QwQ model 3 , showing the potential of ToT for medical data distillation.

– We study tree-based reasoning for quantized biomedical multiple-choice QA with dual evaluator feedback, extending prior ToT evidence from mathematical reasoning [48] and coding [41].

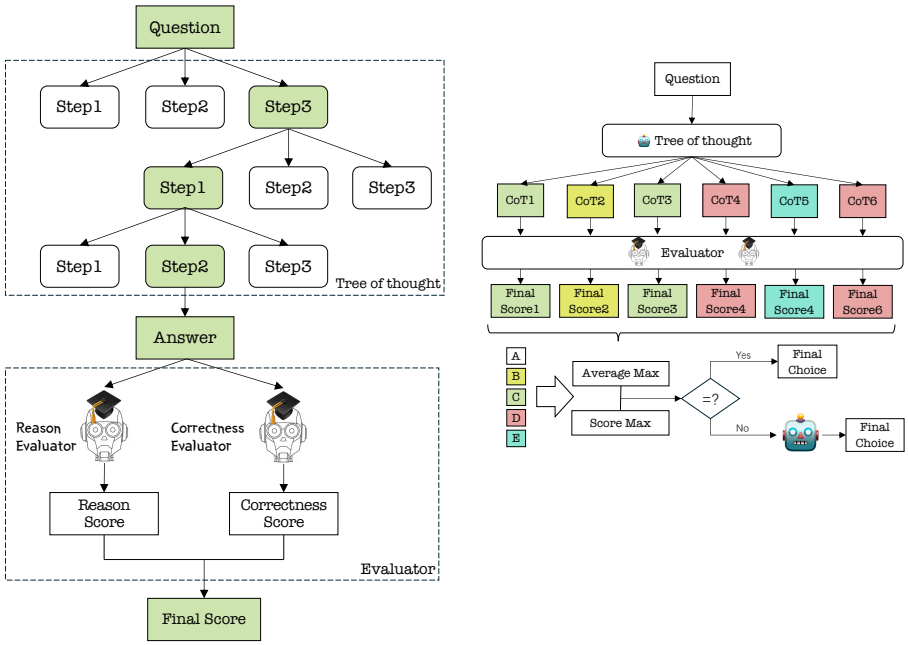

Fig. 2. QM-ToT workflow. Left: tree-based reasoning expands a question into branching paths, and reasoning/medical-fact evaluators combine their scores into a final quality score. Right: the decision workflow compares the highest average score and highest individual score; matching selections are accepted directly, while disagreements trigger re-evaluation.

2 Related Works

2.1 Artificial Intelligence Biomedical Tasks

Biomedical tasks include discriminative settings such as QA, feature extraction [51], diagnosis [14], classification [24], multi-objective optimization [42], concept prediction [43], and related prediction tasks [31], as well as generative settings such as summarization [38,40], generation [25,47], and simplification [28]. LLMs support both general clinical assistance [16] and specialized medical models trained with domain data, such as MedPaLM-2 [36] and Med-Gemini [33]. Related data-driven models in traffic prediction, edge networks, and biomedical classification further show the expanding role of AI in complex real-world systems [2, 49].

2.2 Tree of Thought Reasoning

Tree of thought frames LLM reasoning as search over branching problem-solving paths, echoing Newell's view of problem solving as search through combinatorial spaces [27]. Yao et al. [48] introduced ToT as an LLM problem-solving paradigm, and later work showed gains in mathematics [48], coding [41], and domain-specific data generation [8].

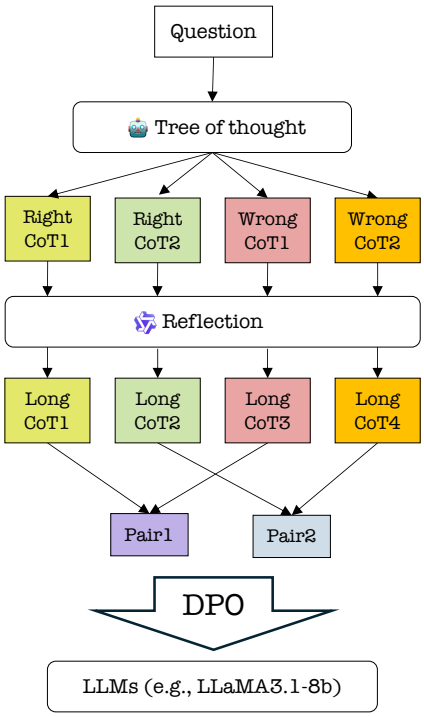

Fig. 3. Reflection-ToT: a data distillation method driven by ToT. Short CoT generated by ToT without reflection is refined by the teacher model (eg. Qwen2.5-72b or LLaMA3- 70B) to produce long CoT. Correct and incorrect long CoT are randomly paired to form Direct Preference Optimization (DPO) training data.

2.3 The Rise of LLM-as-a-Judge

Recent LLMs such as GPT-4 [1] and Qwen2.5 [46] show strong instruction following, query understanding, generation, and causal learning [55]. This enables the "LLM-as-a-judge" paradigm [52], where strong models evaluate, rank, or select candidates across alignment [4], retrieval [19], reasoning [21], and broader LLM lifecycle tasks [18], including self-evolution [37].

2.4 Quantization for Efficient LLM Deployment

LLM reasoning is computationally demanding, and quantization reduces storage and latency by mapping floating-point parameters to integer representations. Representative methods include GPTQ [9], AWQ [22], and GGUF[2] . Although quantization can degrade performance, quantized high-parameter models often outperform smaller unquantized models under similar memory budgets [20], making them attractive for local deployment. Work on quantized sampled-data

systems, secure consensus control, and lightweight authentication similarly emphasizes robustness under resource constraints [10, 34, 35].

3 Method

3.1 ToT for Quantized Medical Reasoning

QM-ToT follows a structured loop: Planning through path decomposition, Execution via thought generation, and Reflection through dual-evaluation feedback. This controlled tree search lets quantized LLMs explore diverse reasoning paths and mitigate degradation through feedback-guided selection. Formally, the generation of a reasoning path can be defined as,

$$s_0 = D_\pi(Q), \tag{1}$$

where Q represents the medical problem description with specific problem statements and requirements, D_π denotes the path decomposition function parameterized by the LLM_π , and s_0 is the initial reasoning path.

Built upon the initial path, the subsequent reasoning paths can be achieved as,

$$s_1 = D_\pi(Q + s_0), \tag{2}$$

where s_1 is derived from both Q and the preceding path s_0, preserving problem context. This iterative process continues, ultimately leading to a final solution:

$$a_Q = s_k = D_\pi(Q + \textstyle\sum_{t=0}^{k} s_t), \tag{3}$$

where a_Q is the final solution and s_k is the cumulative reasoning sequence constrained by maximum length k.

QM-ToT uses an evaluation model W (DeepSeek-V3[3] here) to execute paths and assess validity. Algorithm 1 summarizes the procedure.

[2] https://github.com/ggerganov/ggml

[3] https://github.com/deepseek-ai/DeepSeek-V3

3.2 Result Evaluator

For multiple-choice medical QA, self-consistency can be unreliable when misleading information appears in the query. We therefore score each generated chain with DeepSeek-V3 using reasoning coherence (r) and medical correctness (c), combined into a final score (f_s):

$$fs = \alpha \cdot exp(r) + (1-\alpha) \cdot exp(c) \quad 0 \leq \alpha \leq 1. \tag{4}$$

For each option $x \in \{A, B, C, D, E\}$, we compute both the average score $Avg(fs)_x$ and maximum score$\max(fs)_x$. If their best options match, the answer is:

$$a = argmax_x(Avg(fs)_x). \tag{5}$$

If they differ, a judge model (J_π) compares the two candidates:

$$a = J_\pi\big(argmax_x(Avg(fs)_x), argmax_{x'}(\max(fs)_{x'})\big). \tag{6}$$

Algorithm 1 QM-ToT Reasoning with Maximum Length

Input Q: Medical problem description, k: Maximum length of reasoning paths
Output Solution set a_Q

$S \leftarrow \{\}$ ▷ Initialize the set of reasoning paths
$s_0 \leftarrow D_\pi(Q)$ ▷ Generate the initial path
S.add(s_0)
$i \leftarrow 1$ ▷ Path index
while not $W(Q, S)$ **and** $i \leq k$ **do**
▷ While the problem is not solved and within max length
  for each possible next path s_i from $D_\pi(Q, S)$ **do**
    $S' \leftarrow S \cup \{s_i\}$ ▷ Create a new branch
    if $W(Q, S')$ **then**
      $S \leftarrow S'$ ▷ Update current reasoning paths
      $i \leftarrow i + 1$ ▷ Increment the path index
      break ▷ Move to the next level of the tree
    else
      $S' \leftarrow S' \setminus \{s_i\}$ ▷ Backtrack if invalid paths
    end if
  end for
end while

if $W(Q, S)$ **then**
  return the final path in S as a_Q
else
  return Failure to Solve ▷ If k is exceeded
end if

3.3 Reasoning and Bias Boundaries

To analyze ToT reasoning on MedQA, we adapt the reasoning boundary (RB) [6], which quantifies the maximum difficulty level at which a model can reason effectively:

$$\mathcal{B}\_K_1(t|m) = \sup\_d \{ d \mid Acc(t|d,m) \geq K_1 \}, \quad (7)$$

Where $Acc(t|d,m)$ is accuracy at difficulty d and K_1 is a threshold. For multiplechoice MedQA questions, we also define a bias boundary Ω as the random-guessing baseline (20–25%). Based on CoT-SC performance (Fig. 4), questions are grouped as Easy ($Acc \geq 90\%$), where reasoning is reliable; Medium ($\Omega < Acc < 90\%$), where reasoning is meaningful but imperfect; and Hard ($Acc \leq \Omega$), where failure or bias dominates. This hierarchy enables systematic analysis across question complexity.

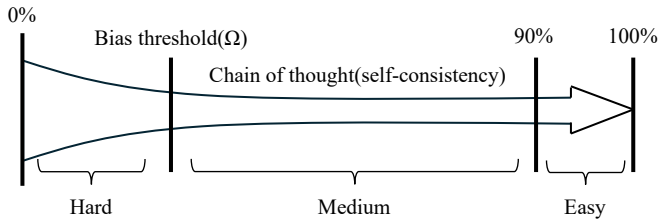

Fig. 4. Difficulty classification of the dataset based on CoT-SC accuracy

3.4 Data Distillation for Data Generation

Because ToT explores divergent solution paths, it naturally generates both correct and incorrect CoT, matching the needs of DPO [32]. With a reflector R driven by model μ, the distilled dataset T is:

$$T = \sum_{Q=1}^{n} R\left(M\left(a_Q\right)\right), \quad (8)$$

where Q represents the question in MedQA-USMLE and M represents the random match policy to create the right-against-wrong sample pairs.

4 Experiments

We evaluate on MedQA-USMLE with 1,272 questions. For 70B-class models, computational cost limits evaluation to a random 400-question MiniTest; the 8B model is evaluated on the full set. We report average accuracy and exclude runs with no answer, so 70B accuracies need not be multiples of 0.25. All experiments use Ollama6 with default settings and GGUF Q4_K_M (INT4) quantization to simulate local hospital deployment. We set α = 0.6 in Eq. (4) and analyze sensitivity below.

4.1 Performance Comparison Analysis

We compare ToT across LLaMA2-70b, LLaMA3-70b, and Qwen2.5-72b on MiniTest, and LLaMA3.1-8b on the full MedQA-USMLE test set.

Table 1 shows that QM-ToT generally improves over CoT-AVG, CoT-SC, and ToT. LLaMA2-70b rises from 34% with CoT-AVG to 50.25% with QM-ToT, and LLaMA3.1-8b rises from 58.77% to 69.49%. LLaMA3-70B is the exception: QM-ToT obtains 79.70%, close to CoT-SC at 80.00%; given the 400-question MiniTest and overlapping confidence intervals, this difference is statistically indistinguishable rather than evidence of degradation.

Fig. 6 shows how different paths for the same question can diverge after a shared first step and receive different scores. The low-scoring path emphasizes incorrect diagnostic features in step 3, while the correct path shows that QM-ToT can decompose a problem and explore multiple perspectives.

Table 1. Accuracy of Quantized (int4) LLMs on MedQA-USMLE. CoT-AVG represents the average accuracy achieved across all questions when CoT reasoning is applied. "ToT": QM-ToT framework without the evaluator module.

Models	CoT-AVG	CoT-SC	ToT	QM-ToT
LLaMA2-70b	34%	27.75%	35.25%	50.25%
LLaMA3-70b	78.65%	80.00%	76.72%	79.70%
Qwen2.5-72b	72.37%	70.25%	73.00%	74.25%
LLaMA3.1-8b	58.77%	59.19%	60.77%	69.49%

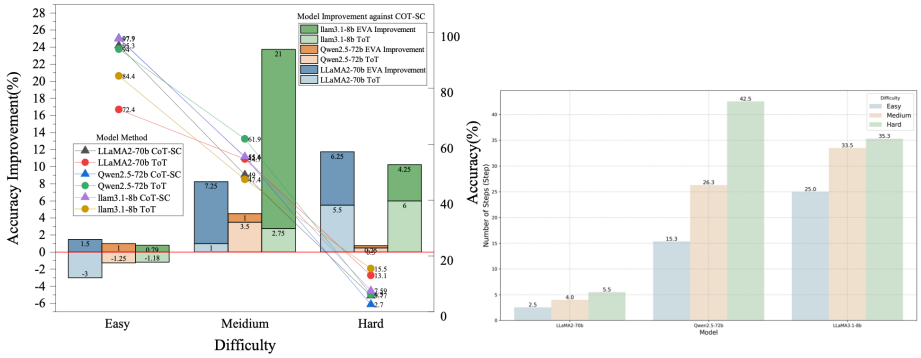

Fig. 5. Difficulty-level results and path scaling. Left: CoT-SC, ToT, and evaluator improved QM-ToT accuracy across difficulty levels. Right: average QM-ToT path counts by difficulty.

4.2 Effectiveness Analysis across Difficulty Levels

We analyze ToT across easy, medium, and hard questions using the boundary definition in Section 3.3. Fig. 5 shows that ToT improves over CoT-SC mainly on medium and hard questions, while gains are minimal or negative on easy questions. The evaluator module improves all models across difficulty levels, with LLaMA3.1-8b gaining especially on medium questions (21%). Overall, QM-ToT is most useful when reasoning is complex rather than already reliable.

Fig. 5 shows that average path count increases with difficulty, indicating adaptive path scaling: QM-ToT allocates more reasoning paths to complex clinical cases and fewer to simpler ones. Together with the INT4 degradation in Fig. 1, this suggests that quantization reduces final accuracy but does not eliminate the model's ability to scale reasoning effort with task difficulty.

4.3 Ablation and Robustness Studies

Statistical reliability. Table 2 reports 95% Clopper–Pearson confidence intervals. Since 70B-class models use 400 questions, their intervals are about ±4–5 points; thus LLaMA3-70B's CoT-SC (80.00%) and QM-ToT (79.70%) results should be treated as statistically indistinguishable.

[Question]

A four years old boy is brought to the physician because of a progressive rash for 2 days. The rash started behind the ears and now involves the trunk and extremities. Over the past 4 days, he has had mild sore throat, red, itchy eyes, and headache. He was born at term and has been healthy except for recurrent episodes of tonsillitis and occasional asthma attacks. Two weeks ago, he was treated for tonsillitis that resolved with penicillin therapy. He immigrated with his family from Brazil 3 weeks ago… Which of the following is the most likely diagnosis?
A: Measles, B: Rubella, C: Allergic drug reaction, D: Scarlet fever, E: Erythema infectiosum

[CoT]

The patient presents with a progressive rash that started behind the ears and now involves the trunk and extremities, which is a characteristic feature of measles. Additionally, the patient's symptoms such as mild sore throat, red, itchy eyes, and headache are consistent with measles. The postauricular and suboccipital lymphadenopathy further supports this diagnosis. Measles was also considered in the differential based on the child's immigration status from Brazil where measles is still prevalent. Allergic drug reaction, rubella, scarlet fever, and erythema infectiosum do not fully account for the patient's symptoms.

[QM-ToT Low Score Solution]

Step1 : Review the patient's history for any prodromal symptoms or preceding illnesses that could suggest a viral etiology, such as recent tonsillitis treated with antibiotics, which may be a clue to superinfection or post-infectious complications. **Step2** : Given the patient's symptoms of mild sore throat, red itchy eyes, headache, and fever, evaluate if there is a pattern consistent with an allergic reaction, such as an allergy to antibiotics that were recently administered for tonsillitis. **Step3** : Consider the patient's rash distribution, which started behind the ears and spread to the trunk and extremities. This pattern is more typical of certain viral exanthems rather than allergic reactions.
Conclusion: The provided information in step 3 indicates that the patient's rash distribution, which started behind the ears and spread to the trunk and extremities, is more typical of certain viral exanthems rather than allergic reactions. This pattern is a crucial clue for determining the diagnosis. Additionally, the previous steps (steps 1 and 2) have already ruled out an allergic reaction as the most likely cause. The current information builds upon this analysis by providing further context about the rash distribution.

[QM-ToT High Score Solution]

Step1 : Review the patient's history for any prodromal symptoms or preceding illnesses that could suggest a viral etiology, such as recent tonsillitis treated with antibiotics, which may be a clue to superinfection or post-infectious complications. **Step2** : Examine the patient's rash for specific characteristics that might point towards a particular diagnosis. For example, is it non-confluent and maculopapular as described, or are there other features such as blisters or petechiae? **Step3**: Note that the rash is described as non-confluent and maculopapular, which is characteristic of certain viral infections such as measles. However, the presence of postauricular and suboccipital lymphadenopathy may also suggest a diagnosis like rubella.
Conclusion: The patient's symptoms, particularly the non-confluent and maculopapular rash, along with prodromal symptoms such as fever, sore throat, red itchy eyes, and headache, are consistent with viral infections like measles or rubella. However, the presence of postauricular and suboccipital lymphadenopathy suggests a diagnosis like rubella. Considering these characteristics, we can proceed to evaluate the options.

Fig. 6. Example of different solution to the #44 question. All the solutions are generated by LlaMA3.1-8b-INT4. The red text indicates the reasoning process that led to incorrect results, while the green text shows the reasoning process that led to correct results.

Table 2. 95% confidence intervals for QM-ToT accuracy. Sensitivity to α. Table 3 varies the coefficient balancing reasoning coherence and medical correctness in Eq. (4). Using both scores outperforms either alone, and α = 0.6 is close to the best setting within the stable 0.4–0.7 range.

Model A	Accuracy	*n*	95% CI
LLaMA2-70B	50.25%	400	[45.24%, 55.01%]
LLaMA3-70B	79.70%	400	[75.47%, 83.35%]
Qwen2.5-72B 7	74.25%	400	[69.67%, 78.23%]
LLaMA3.1-8B	69.49%	1272	[66.88%, 71.94%

Table 3. Accuracy under different α values.

Model	0.0	0.2	0.4	0.5	0.6	0.8	1.0
LLaMA2 -70B	48.02%	49.09%	50.19%	49.75%	50.25%	49.05%	47.60%
LLaMA3 -70B	78.50%	79.80%	79.73%	80.10%	79.70%	79.47%	78.28%
Qwen2.5 -72B	72.01%	72.91%	73.52%	74.18%	74.25%	73.40%	72.01%
LLaMA3 .1-8B	68.56%	68.88%	68.20%	69.32%	69.49%	68.15%	67.15%

Table 4. Ablation of the two-stage answer selection strategy.

Statistic	LLaMA2-70B	LLaMA3-70B	Qwen2.5-72B	LLaMA3.1-8B
Agree rate	53%	72%	72%	68%
Disagree rate	47%	28%	30%	32%
Avg-Only	43.55%	78.69%	73.54%	68.13%
Score-Only	45.67%	77.84%	72.60%	67.41%
Two-Stage	50.25%	79.70%	74.25%	69.49%
Judge win rate	62%	75%	72%	70%

Two-stage answer selection. Table 4 shows that comparing the highest average score option with the highest individual-score option outperforms Avg-Only and Score-Only selection. Disagreements occur in 28%–35% of cases, and the judge resolves 62%–75% correctly, so the second stage is not redundant.

FP16 control experiment. Table 5 separates general QM-ToT benefits from quantization-specific effects. QM-ToT also improves FP16 LLaMA3.1-8B, but the gain is much smaller than under INT4, suggesting that structured search and evaluator feedback are especially beneficial for quantized models.

Self-evaluator baseline. Table 6 replaces DeepSeek-V3 with the student model. Self-Eval QM-ToT improves over CoT-SC for LLaMA2-70B, Qwen2.5-72B, and LLaMA3.1-8B, showing that the ToT structure itself contributes; DeepSeek-V3 adds further gains, especially for weaker models.

CoT-SC with DeepSeek selector. Table 7 tests whether DeepSeek-V3 selection alone explains the gain. CoT-SC with DeepSeek improves over plain CoT-SC in most cases but remains below full QM-ToT, indicating that tree-structured exploration adds benefit beyond external selection.

Table 5. FP16 control experiment on LLaMA3.1-8B.

Precision	CoT-SC	QM-ToT	Improvement
FP16	66.43%	68.09%	+1.66%
INT4	59.19%	69.49%	I+10.30%

Table 6. Self-evaluator baseline for QM-ToT.

Model	CoT-SC	Self-Eval QM-ToT	DeepSeek QM-ToT	Self-Eval Gain	DeepSeek Extra Gain
LLaMA2-70B	27.75%	38.50%	50.25%	+10.75%	+11.75%
LLaMA3-70B	80.00%	78.25%	79.70%	-1.75%	+1.70%
Qwen2.5-72B	70.25%	71.50%	74.25%	+1.25%	+2.80%
LLaMA3.1-8B	59.19%	63.11%	69.49%	+3.92%	+6.38%

Table 7. CoT-SC with DeepSeek selector baseline.

Model	CoT-SC	CoT-SC + DeepSeek	QM-ToT	ToT Extra Gain
LLaMA2-70B	27.75%	42.00%	50.25%	+8.25%
LLaMA3-70B	80.00%	78.50%	79.70%	+1.20%
Qwen2.5-72B	70.25%	72.50%	74.25%	+1.75%
LLaMA3.1-8B	59.19%	64.32%	69.49%	+5.17%

Table 8. Comparison of Reflection-ToT and CoT Approaches

	Baseline	CoT	CoT (Matched Pairs)	Reflection-ToT
Teacher Model	-	QwQ	QwQ	LLaMA3-70B
Question Used ↓	-	10178	10178	10000
Generated Data ↓	-	6659 (samples)	1514 (pairs)	1514 (pairs)
Accuracy↑	61.01%	65.01%	64.73%	66.44%

Multi-seed stability. Table 8 repeats QM-ToT with three seeds. The standard deviation remains around 1 percentage point, suggesting stable performance across sampling variation.

4.4 Performance of Reflection-ToT Distillation

We use Llamafactory [53] to train LoRA [11] adapters for LLaMA3.1-8B on MedQA-USMLE-test, with MedQA-USMLE-dev/train used for distillation data. DPO uses

LLaMA3-70B as teacher, learning rate 5e-5, batch size 16, cut-off length 3k, one epoch, and 1,000 sampled training questions; SFT uses QwQ with the same hyperparameters. Reflection-ToT refines short ToT CoT into extended o1-style CoT with reflection and summarization (Fig. 7). Table 9 shows that it uses only 1,000 questions to construct 1,514 preference pairs and achieves 66.44% accuracy, compared with 65.01% from full QwQ-CoT distillation over 10,178 questions and 64.73% from QwQ matched pairs. This suggests data efficiency, though teacher and construction differences require cautious interpretation.

Table 9. QM-ToT accuracy across three random seeds.

Model	Seed 1	Seed 2	Seed 3	Mean ± Std
LLaMA2-70B	51.25%	48.50%	51.00%	49.90 ± 1.30%
LLaMA3-70B	80.10%	80.50%	78.50%	79.60 ± 1.00%
Qwen2.5-72B	74.25%	73.00%	75.50%	74.30 ± 1.30%
LLaMA3.1-8B	69.82%	68.47%	70.18%	69.40 ± 0.90%

[Long CoT with Reflection]

Okay, so I've got this question about a 4-year-old boy with a progressive rash. Let me think... the rash started behind the ears and now involves the trunk and extremities. Over the past 4 days, he's had a mild sore throat, red itchy eyes, and a headache. **Hold on...** these symptoms are starting to sound familiar.

I remember from textbooks that a rash starting behind the ears and spreading to the trunk and extremities is a classic presentation. This brings up the idea of viral exanthems, like measles or rubella. The patient also has postauricular and suboccipital lymphadenopathy, which is a big clue.

Let's break it down step by step. First, the rash is described as a non-confluent, pink maculopapular rash. Maculopapular rashes can be associated with a variety of infections, but the pattern of starting behind the ears and spreading is very specific.

Next, the patient's travel history from Brazil and unknown immunization status is important. Brazil still has cases of measles and rubella, so these are definitely on the differential.

Now, let's look at the symptoms in more detail. He has a mild sore throat, red itchy eyes, and a headache. These are consistent with both measles and rubella. But the key here is the postauricular and suboccipital lymphadenopathy, which is more characteristic of rubella.

The patient also has a history of recurrent tonsillitis and occasional asthma attacks, but these aren't directly related to the current presentation. The recent tonsillitis treated with penicillin that resolved is a good sign, **but it doesn't rule out a viral infection.**

Ah, now I see it. The combination of the rash pattern, the lymphadenopathy, and the travel history strongly points to rubella. Measles would typically present with more severe symptoms and a different pattern of rash spread. Scarlet fever would have a more confluent rash and more pronounced throat symptoms. An allergic drug reaction would be more localized and less systemic. Erythema infectiosum (fifth disease) would have a more distinctive "slapped cheek" appearance.

In conclusion, to tie it all together, the most likely diagnosis is rubella. The patient's symptoms, travel history, and physical exam findings all align with this. The answer is B.

Fig. 7. Example of right long CoT to the #44 question from Reflection-ToT. The bold text demonstrate the reflection thinking process. The green text shows the reasoning process that led to correct results.

5 Limitations

QM-ToT improves medical QA accuracy by expanding and evaluating multiple reasoning paths, but this increases inference cost over CoT-SC. Based on path count and prompt length, the estimated cost ranges from about 2.8× to 20.5× depending on model and difficulty, so future work should optimize path pruning, early stopping, and evaluator invocation. QM-ToT also depends on an external evaluator, although self-evaluator and DeepSeek-selector ablations show that both tree search and evaluator quality contribute, the method is not fully self-contained.

In addition, difficulty stratification is based on CoT-SC accuracy, which is useful for diagnosing CoT-SC failures but not an independent difficulty label; future work should include question category, USMLE step, expert annotation, or modelindependent item statistics. Finally, 70B-class results use a 400-question MiniTest subset, so small differences of a few percentage points should be interpreted cautiously despite confidence intervals and multi-seed results.

6 Conclusion

We introduced QM-ToT, a structured tree-of-thought framework for feedbackguided clinical reasoning in quantized LLMs. The results show that evaluator guided structured reasoning can improve reliability under low-precision deployment, though with higher inference-token cost. QM-ToT dynamically allocates reasoning paths by task complexity and can generate useful reasoning data, supporting both local biomedical deployment and ToT-driven low-cost data distillation. Future work includes Monte Carlo tree search [5], reinforcement learning [30], neural architecture search [54], black-box optimization [13], evolutionary computation [44], and fine-tuning on the generated reasoning chains to build biomedical deep-thinking models.

Acknowledgments. This work was supported in part by the National Nature Science Foundation of China under Grant No. 62572413, in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2025A1515012944, in part by the Guangdong and Hong Kong Universities '1+1+1' Joint Research Collaboration Scheme.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Ali, A., Sharafian, A., Naeem, H.Y., Zakarya, M., Wu, Z., Bai, X.: Advanced computational models for urban traffic flow prediction: A comprehensive review and future directions. Computer Science Review **60**, 100886 (2026)
3. Anthropic: Claude. https://claude.ai/, accessed: 2026-01-22

4. Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., et al.: Constitutional ai: Harmlessness from ai feedback. arXiv preprint arXiv:2212.08073 (2022)
5. Browne, C.B., Powley, E., Whitehouse, D., Lucas, S.M., Cowling, P.I., Rohlfshagen, P., Tavener, S., Perez, D., Samothrakis, S., Colton, S.: A survey of monte carlo tree search methods. IEEE Transactions on Computational Intelligence and AI in Games (2012). https://doi.org/10.1109/TCIAIG.2012.2186810
6. Chen, Q., Qin, L., Wang, J., Zhou, J., Che, W.: Unlocking the capabilities of thought: A reasoning boundary framework to quantify and optimize chain-of-thought. arXiv preprint arXiv:2410.05695 (2024)
7. Dhar, N., Deng, B., Lo, D., Wu, X., Zhao, L., Suo, K.: An empirical analysis and resource footprint study of deploying large language models on edge devices. In: Proceedings of the 2024 ACM Southeast Conference. pp. 69–76 (2024)
8. Feng, X., Wan, Z., Wen, M., McAleer, S.M., Wen, Y., Zhang, W., Wang, J.: Alphazero-like tree-search can guide large language model decoding and training. arXiv preprint arXiv:2309.17179 (2023)
9. Frantar, E., Ashkboos, S., Hoefler, T., Alistarh, D.: Gptq: Accurate posttraining quantization for generative pre-trained transformers. arXiv preprint arXiv:2210.17323 (2022)
10. Haider, Z.A., Fayaz, M., Zhang, Y., Ali, A.: Advanced hyperelliptic curve-based authentication protocols for secure internet of drones communication. ICCK Transactions on Advanced Computing and Systems **1**(4), 1–16 (2024)
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. arXiv preprint arXiv:2106.09685 (2021)
12. Hu, Y., Liu, R., Zhang, J., Huang, Z.A., Song, L., Tan, K.C.: Heterogeneous structured federated learning with graph convolutional aggregation for mri-based mental disorder diagnosis. In: 2024 International Joint Conference on Neural Networks (IJCNN). pp. 1–8 (2024). https://doi.org/10.1109/IJCNN60899.2024.10651026
13. Huang, B., Wu, X., Zhou, Y., Wu, J., Feng, L., Cheng, R., Tan, K.C.: Exploring the true potential: Evaluating the black-box optimization capability of large language models. arXiv preprint arXiv:2404.06290 (2024)
14. Huang, Z.A., Hu, Y., Liu, R., Xue, X., Zhu, Z., Song, L., Tan, K.C.: Federated multi-task learning for joint diagnosis of multiple mental disorders on mri scans. IEEE Transactions on Biomedical Engineering **70**(4), 1137–1149 (2023). https://doi.org/10.1109/TBME.2022.3210940
15. Huang, Z.A., Zhu, Z., Yau, C.H., Tan, K.C.: Identifying autism spectrum disorder from resting-state fmri using deep belief network. IEEE Transactions on Neural Networks and Learning Systems **32**(7), 2847–2861 (2021). https://doi.org/10. 1109/TNNLS.2020.3007943
16. Jiang, X., Fang, Y., Qiu, R., Zhang, H., Xu, Y., Chen, H., Zhang, W., Zhang, R., Fang, Y., Ma, X., et al.: Tc–rag: Turing–complete rag's case study on medical llm systems. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 11400–11426 (2025)
17. Jin, D., Pan, E., Oufattole, N., Weng, W.H., Fang, H., Szolovits, P.: What disease does this patient have? a large-scale open domain question answering dataset from medical exams. Applied Sciences **11**(14), 6421 (2021)
18. Li, D., Jiang, B., Huang, L., Beigi, A., Zhao, C., Tan, Z., Bhattacharjee, A., Jiang, Y., Chen, C., Wu, T., et al.: From generation to judgment: Opportunities and challenges of llm-as-a-judge. arXiv preprint arXiv:2411.16594 (2024)

19. Li, D., Yang, S., Tan, Z., Baik, J.Y., Yun, S., Lee, J., Chacko, A., Hou, B., Duong-Tran, D., Ding, Y., et al.: Dalk: Dynamic co-augmentation of llms and kg to answer alzheimer's disease questions with scientific literature. arXiv preprint arXiv:2405.04819 (2024)
20. Li, S., Ning, X., Wang, L., Liu, T., Shi, X., Yan, S., Dai, G., Yang, H., Wang, Y.: Evaluating quantized large language models. arXiv preprint arXiv:2402.18158 (2024)
21. Liang, T., He, Z., Jiao, W., Wang, X., Wang, Y., Wang, R., Yang, Y., Shi, S., Tu, Z.: Encouraging divergent thinking in large language models through multi-agent debate. arXiv preprint arXiv:2305.19118 (2023)
22. Lin, J., Tang, J., Tang, H., Yang, S., Chen, W.M., Wang, W.C., Xiao, G., Dang, X., Gan, C., Han, S.: Awq: Activation-aware weight quantization for on-device llm compression and acceleration. Proceedings of Machine Learning and Systems **6**, 87–100 (2024)
23. Liu, L., Yang, X., Lei, J., Shen, Y., Wang, J., Wei, P., Chu, Z., Qin, Z., Ren, K.: A survey on medical large language models: Technology, application, trustworthiness, and future directions. arXiv preprint arXiv:2406.03712 (2024)
24. Liu, R., Huang, Z.a., Jiang, M., Tan, K.C.: Multi-lstm networks for accurate classification of attention deficit hyperactivity disorder from resting-state fmri data. In: 2020 2nd International Conference on Industrial Artificial Intelligence (IAI). pp. 1–6 (2020). https://doi.org/10.1109/IAI50351.2020.9262176
25. Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H., Liu, T.Y.: Biogpt: generative pre-trained transformer for biomedical text generation and mining. Briefings in bioinformatics **23**(6), bbac409 (2022)
26. Melin, E.L., Torek, A.J., Eisty, N.U., Kennington, C.: Precision or peril: Evaluating code quality from quantized large language models. arXiv preprint arXiv:2411.10656 (2024)
27. Newell, A., Shaw, J.C., Simon, H.A.: Report on a general problem solving program. In: IFIP congress. vol. 256, p. 64. Pittsburgh, PA (1959)
28. Ondov, B., Attal, K., Demner-Fushman, D.: A survey of automated methods for biomedical text simplification. Journal of the American Medical Informatics Association **29**(11), 1976–1988 (2022)
29. OpenAI: Chatgpt. https://chat.openai.com/, accessed: 2026-01-22
30. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P.F., Leike, J., Lowe, R.: Training language models to follow instructions with human feedback. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems. vol. 35, pp. 27730–27744. Curran Associates, Inc. (2022), https://proceedings.neurips.cc/paper_files/paper/2022/file/ b1efde53be364a73914f58805a001731-Paper-Conference.pdf
31. Qian, J., Lu, Y., Xie, W., Lai, Z., Wang, M., Li, X.: Cpsr-clip: Conditional promptinduced style reconstruction for zero-shot domain adaptation. IEEE Transactions on Multimedia (2025). https://doi.org/10.1109/TMM.2025.3599044
32. Rafailov, R., Sharma, A., Mitchell, E., Manning, C.D., Ermon, S., Finn, C.: Direct preference optimization: Your language model is secretly a reward model. Advances in neural information processing systems **36**, 53728–53741 (2023)
33. Saab, K., Tu, T., Weng, W.H., Tanno, R., Stutz, D., Wulczyn, E., Zhang, F., Strother, T., Park, C., Vedadi, E., et al.: Capabilities of gemini models in medicine. arXiv preprint arXiv:2404.18416 (2024)
34. Samy, S.A., Durairaj, P., Ali, A., Wu, Z., Li, J., Bai, X.: Secure fuzzy-based h∞ consensus control for nonlinear multi-agent systems under quantized sampled-data mechanism and its applications. Engineering Applications of Artificial Intelligence **162**, 112715 (2025)

35. Sharafian, A., Ghandi, F., Ali, A., Ullah, I., Zhang, B., Bai, X.: Consensus tracking control of incommensurate fractional order multiagent systems using sliding mode for secure communication systems. Physica Scripta **100**(8), 085260 (2025)
36. Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Hou, L., Clark, K., Pfohl, S., Cole-Lewis, H., Neal, D., et al.: Towards expert-level medical question answering with large language models. arXiv preprint arXiv:2305.09617 (2023)
37. Sun, Z., Shen, Y., Zhang, H., Zhou, Q., Chen, Z., Cox, D.D., Yang, Y., Gan, C.: Salmon: Self-alignment with instructable reward models. In: The Twelfth International Conference on Learning Representations (2024)
38. Tang, L., Sun, Z., Idnay, B., Nestor, J.G., Soroush, A., Elias, P.A., Xu, Z., Ding, Y., Durrett, G., Rousseau, J.F., et al.: Evaluating large language models on medical evidence summarization. NPJ digital medicine **6**(1), 158 (2023)
39. Thirunavukarasu, A.J., Hassan, R., Mahmood, S., Sanghera, R., Barzangi, K., El Mukashfi, M., Shah, S.: Trialling a large language model (chatgpt) in general practice with the applied knowledge test: observational study demonstrating opportunities and limitations in primary care. JMIR Medical Education **9**(1), e46599 (2023)
40. Van Veen, D., Van Uden, C., Blankemeier, L., Delbrouck, J.B., Aali, A., Bluethgen, C., Pareek, A., Polacin, M., Reis, E.P., Seehofnerova, A., et al.: Clinical text summarization: adapting large language models can outperform human experts. Research Square (2023)
41. Wang, H., Liu, B., Zhang, Y., Chen, J.: Seed-cts: Unleashing the power of tree search for superiorperformance in competitive coding tasks. arXiv preprint arXiv:2412.12544 (2024)
42. Wang, Z., Liu, S., Chen, J., Tan, K.C.: Large language model-aided evolutionary search for constrained multiobjective optimization. In: International Conference on Intelligent Computing. pp. 218–230. Springer (2024)
43. Wang, Z., Lin, Z., Lin, W., Yang, M., Zeng, M., Tan, K.C.: Explainable molecular property prediction: Aligning chemical concepts with predictions via language models. arXiv preprint arXiv:2405.16041 (2024)
44. Wu, X., Wu, S.h., Wu, J., Feng, L., Tan, K.C.: Evolutionary computation in the era of large language model: Survey and roadmap. arXiv preprint arXiv:2401.10034 (2024)
45. Xie, Y., Wu, J., Tu, H., Yang, S., Zhao, B., Zong, Y., Jin, Q., Xie, C., Zhou, Y.: A preliminary study of o1 in medicine: Are we closer to an ai doctor? arXiv preprint arXiv:2409.15277 (2024)
46. Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., et al.: Qwen2. 5 technical report. arXiv preprint arXiv:2412.15115 (2024)
47. Yang, Z., Qian, J., Peng, Z., Zhang, H., Huang, Y.A., Tan, K., Huang, Z.A.: Medrefl: Medical reasoning enhancement via self-corrected fine-grained reflection. arXiv preprint arXiv:2506.13793 (2025)
48. Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T., Cao, Y., Narasimhan, K.: Tree of thoughts: Deliberate problem solving with large language models. Advances in Neural Information Processing Systems **36** (2024)
49. Yin, J., Ali, R., Li, L., Ma, W., Ali, A.: Edge network model based on double dimension. In: 2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC). pp. 341–346. IEEE (2018)
50. Zhang, Y., Huang, Z.A., Hong, Z., Wu, S., Wu, J., Tan, K.C.: Mixed prototype correction for causal inference in medical image classification. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 4377–4386 (2024)

51. Zhao, Y., Zhang, Y., Huang, Z.A., Yang, F., Duan, L., Yao, J.: Mining the associations between v(d)j gene segments and covid-19 disease characteristics. In: 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). pp. 608–613 (2021). https://doi.org/10.1109/BIBM52615.2021.9669824
52. Zheng, L., Chiang, W.L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al.: Judging llm-as-a-judge with mt-bench and chatbot arena. Advances in Neural Information Processing Systems **36**, 46595–46623 (2023)
53. Zheng, Y., Zhang, R., Zhang, J., Ye, Y., Luo, Z., Feng, Z., Ma, Y.: Llamafactory: Unified efficient fine-tuning of 100+ language models. arXiv preprint arXiv:2403.13372 (2024)
54. Zhou, X., Wu, X., Feng, L., Lu, Z., Tan, K.C.: Design principle transfer in neural architecture search via large language models (2024), https://arxiv.org/abs/ 2408.11330
55. Zhou, Y., Wu, X., Huang, B., Wu, J., Feng, L., Tan, K.C.: Causalbench: A comprehensive benchmark for causal learning capability of large language models. arXiv preprint arXiv:2404.06349 (2024)