How Extractable Are Edge LLMs?

Anonymous ACL submission

Abstract

Edge large language models (Edge LLMs) are increasingly deployed on resource-constrained smart devices through aggressive compression and quantization. Although these techniques substantially reduce model size and deployment cost, they do not fully eliminate the risk of behavioral knowledge leakage through query-based interaction. Compared with cloud-based LLM extraction, extracting quantized Edge LLMs is fundamentally more challenging due to strict interaction budgets, limited model capacity, and quantization-induced perturbations. As a result, prior work often assumes that quantization and deployment constraints naturally hinder effective behavioral extraction, leaving the associated security risks underexplored. To systematically investigate this overlooked vulnerability, we propose CLIQ (**C**oordinated **L**inguistic **I**nstruction **Q**uerying), a structured semantic querying framework for analyzing query-based knowledge extraction from quantized Edge LLMs. CLIQ organizes instruction queries through semantic clustering, constructs low-redundancy query pools with broad semantic coverage, and enables efficient behavioral reconstruction under constrained interaction budgets. Extensive experiments on mainstream open-source Edge LLMs show that quantization alone does not provide effective protection against query-based extraction, highlighting a previously underexplored security risk in edge deployment.

1 Introduction

Large language models (LLMs) have demonstrated remarkable instruction-following and reasoning capabilities across a wide range of applications (Ouyang et al., 2022). As training frontier LLMs requires enormous computational and financial resources (Hoffmann et al., 2022), trained models have become valuable intellectual assets. Meanwhile, recent advances in model compression and efficient inference have accelerated the deployment of LLMs onto edge devices such as smartphones, laptops, and embedded AI systems (Zheng et al., 2025). Industry systems including Apple Intelligence, Gemini Nano, Samsung Galaxy AI, and Qualcomm mobile AI platforms increasingly rely on on-device foundation models for local inference and privacy-preserving interaction (Google, 2023; Samsung Electronics, 2024; Qualcomm, 2024). As LLM capabilities migrate from cloud infrastructure to personal devices, understanding the security implications of edge-deployed models becomes increasingly important.

To satisfy the strict memory and latency constraints of edge hardware, deployed models are typically compressed through aggressive low-bit quantization such as INT8 or INT4. Although quantization substantially reduces deployment cost, it also introduces discretization noise and output instability (Dettmers et al., 2023; Nagel et al., 2021). This creates an important distinction from traditional cloud-side model extraction settings, where attackers usually interact with full-precision models under relatively unconstrained query access (Hinton et al., 2015; Zhang et al., 2024). In contrast, edge-oriented interaction is inherently constrained: on-device inference incurs nontrivial latency and energy consumption, while practical deployment interfaces naturally limit the number of allowed interactions (Zheng et al., 2025). Consequently, existing extraction or distillation strategies that rely on large volumes of randomly sampled queries become substantially less effective in edge scenarios.

At first glance, these constraints may appear to provide a degree of natural protection against model extraction. However, quantization primarily perturbs output precision rather than removing the semantic knowledge encoded within model parameters (Gholami et al., 2021). The core challenge therefore shifts from *whether* extraction remains possible to *how* queries should be constructed under strict interaction budgets. In practice, instruc-

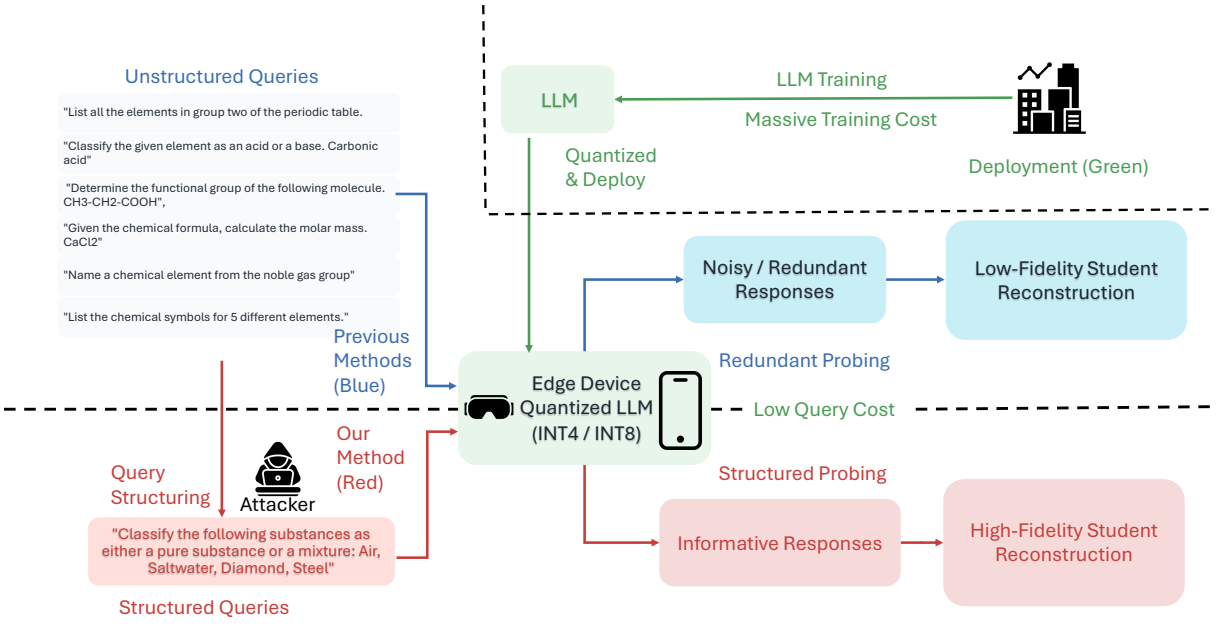

Figure 1: Overview of query-based knowledge extraction from quantized edge-deployed LLMs. Previous approaches (blue) rely on unstructured query interactions, which repeatedly probe overlapping semantic regions and produce noisy, low-information supervision under limited budgets. CLIQ (red) performs structured semantic probing through clustered instruction querying, constructing representative low-redundancy queries that enable substantially more effective behavioral extraction from quantized edge models.

tion queries sampled from natural distributions exhibit substantial semantic redundancy and repeatedly probe similar capability regions, leading to inefficient supervision and limited behavioral recovery (Shu et al., 2026). This suggests that query structure, rather than query quantity alone, may fundamentally govern extraction effectiveness in quantized edge settings.

Motivated by this observation, we propose **CLIQ** (**C**oordinated **L**inguistic Instruction **Q**uerying), a structured query construction framework designed for behavioral extraction from quantized edge LLMs. As illustrated in Figure 1, CLIQ first organizes candidate instructions through semantic clustering to uncover latent structure within the instruction space. It then constructs representative low-redundancy query sets for each semantic region, enabling broader semantic coverage under limited interaction budgets. By performing structured semantic probing instead of naive random querying, CLIQ efficiently recovers behavioral knowledge from quantized edge models despite strict deployment constraints.

To quantify extractability, we collect query-response pairs from quantized edge models and train compact student models to reproduce their behaviors through black-box interaction. If a student model can successfully approximate the target model using only a limited number of structured interactions, this indicates that substantial behavioral knowledge remains recoverable even after aggressive quantization. Extensive experiments across multiple architectures, quantization levels, and query budgets demonstrate that CLIQ consistently achieves significantly more effective behavioral extraction than conventional query strategies. Our results reveal that quantization alone does not provide effective protection against query-based extraction, highlighting a previously underexplored security risk in edge-deployed LLMs.

Contributions. This paper makes the following contributions:

(i) We identify a fundamental security risk in edge-deployed LLMs: aggressive quantization does not prevent query-based behavioral extraction, and substantial model knowledge remains recoverable under realistic interaction budgets. To the best of our knowledge, this is the first systematic study of extraction risks in quantized on-device LLMs.

(ii) We propose CLIQ, a coordinated linguistic instruction querying framework that performs structured semantic probing through represen-

tative low-redundancy query construction, enabling efficient behavioral extraction under strict edge interaction constraints.

(iii) We demonstrate across multiple architectures and quantization settings that structured query construction substantially improves extraction efficiency, showing that query design—rather than quantization itself—is the key factor governing extractability in edge LLMs.

2 Related Work

2.1 Knowledge Distillation for LLM

Knowledge distillation transfers knowledge from large teacher models to smaller students, with early work focusing on classification (Hinton et al., 2015) and sequence modeling (Kim and Rush, 2016), and recent studies extending to instruction-level (Honovich et al., 2023) and response-level distillation (Gu et al., 2024). These approaches typically rely on large query sets and assume inexpensive, high-quality teacher supervision.

Recent work explores data-free and API-based distillation, where students learn from black-box teacher outputs without ground-truth annotations (Zhang et al., 2023; Chen et al., 2024). However, these methods treat black-box access as a practical constraint rather than a security concern, implicitly assuming reliable supervision and ignoring the risk of capability leakage through query interaction.

2.2 Data-efficient and Query-efficient Distillation for LLM

Data-efficient distillation methods reduce supervision cost via uncertainty-based sampling (Du et al., 2025) or synthetic data generation (Honovich et al., 2023; Wang et al., 2023). These approaches focus on selecting informative queries under the assumption that the teacher model is high-precision and provides stable responses.

From a security perspective, this assumption does not necessarily hold in edge deployment settings, where models are heavily quantized and interaction budgets are inherently constrained. Under such conditions, query structure becomes a critical factor governing both extraction efficiency and the amount of recoverable behavioral knowledge.

2.3 Edge-oriented and Quantized LLMs

Edge deployment has driven extensive research on model compression and quantization techniques, including GPTQ and QLoRA (Frantar et al., 2022; Dettmers et al., 2023), which reduce memory and computation costs while introducing quantization noise and reduced representational expressivity.

Quantization is often assumed to mitigate extraction risks due to reduced precision and compressed parameter representations. However, its impact on query-based knowledge extraction from edge-deployed models remains largely unclear. We address this gap by showing that structured queries can still efficiently recover substantial behavioral knowledge from quantized edge models.

3 Method

Figure 2 illustrates the threat framework considered in this work. Traditional model extraction studies typically assume high-performance server environments and interact with full-precision teacher models using a large number of queries. In contrast, our setting focuses on quantized LLMs deployed on resource-constrained edge devices, where both query budgets and computational resources are limited. Our goal is to study whether carefully designed structured queries can still recover substantial behavioral knowledge from such models.

To systematically analyze this risk, we instantiate a structured querying framework that organizes candidate instructions according to semantic regions and constructs representative probes for model interaction. The collected responses are then used to reconstruct model behavior through student modeling, providing a practical measure of how much knowledge can be extracted from the edge model under realistic deployment constraints.

3.1 Threat Model and Problem Setting

As illustrated in Figure 2, compared with server-side extraction settings, the attacker operates under a smaller query budget and a noisier model interface, while the model still retains substantial semantic knowledge.

Adversary capability. The attacker has black-box access to the model and can submit instruction queries and observe the generated responses. No access to model parameters, gradients, logits, or training data is assumed.

Query budget. The attacker is restricted to a limited number of interactions due to practical constraints including interface rate limits, inference latency, and edge-device resource overhead.

Adversary goal. The goal is to recover as much information as possible about the model's behav-

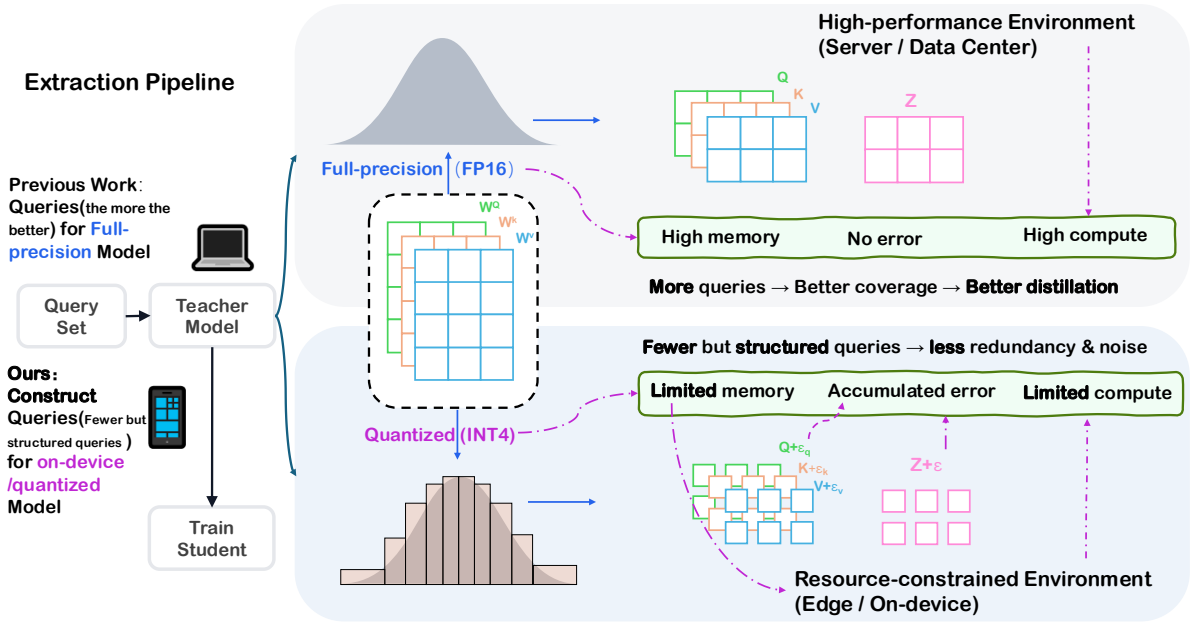

Figure 2: Threat framework for query-based knowledge extraction from quantized edge-deployed LLMs. Traditional extraction settings (top) assume full-precision teacher models in high-performance server environments, where abundant compute allows large-scale query probing. In contrast, edge-deployed LLMs (bottom) operate under quantization (e.g., INT4/INT8) and strict resource constraints, resulting in limited query budgets and noisy responses. Nevertheless, carefully structured queries can still recover substantial behavioral knowledge from the quantized model.

ior using a limited query budget. Following prior work on model extraction, we measure this leakage by training a student model to imitate the target responses. High-fidelity imitation indicates that substantial knowledge has been extracted through query interaction.

3.2 Quantized Models as Leakage Surfaces

Edge deployment commonly relies on aggressive weight quantization to reduce model size and computational cost. Although quantization reduces numerical precision and may degrade the reliability of naive extraction strategies, it does not remove the semantic knowledge encoded in model parameters. Instead, it transforms the extraction problem into a more constrained and noisier variant.

We model the behavior of a quantized teacher model as

$$T_\theta^{\text{quant}}(x) = T_\theta^{\text{full}}(x) + \epsilon(x), \tag{1}$$

where $\epsilon(x)$ denotes the output perturbation induced by quantization.

Importantly, quantization should not be interpreted as a security defense against query-based extraction. Although the perturbation term may reduce the effectiveness of naive random probing, the underlying behavioral and semantic knowledge of the model remains largely preserved. Therefore, an attacker who allocates queries more efficiently may still recover substantial information from the deployed model. This motivates a structured query design strategy that is robust to limited budgets and avoids excessive redundancy.

3.3 Semantic Redundancy under Limited Budgets

Under constrained query budgets, extraction performance depends not only on the number of interactions, but also on how efficiently queries explore the semantic capability space of the target model.

Natural instruction pools exhibit substantial semantic redundancy. Random query sampling repeatedly probes similar behaviors, resulting in inefficient semantic coverage under limited budgets.

Figure 8 shows that original queries exhibit substantially higher intra-cluster similarity, indicating redundant supervision signals and overlapping semantic regions.

Motivated by this observation, CLIQ organizes queries according to semantic structure and allocates the query budget across distinct capability regions.

3.4 CLIQ: Coordinated Linguistic Instruction Querying

We propose **CLIQ** (**C**oordinated **L**inguistic **I**nstruction **Q**uerying), a structured semantic probing framework designed to systematically analyze query-based behavioral extraction from quantized Edge LLMs under constrained interaction budgets.

Rather than relying on independently sampled instruction queries, CLIQ organizes the query space into semantically coherent regions and coordinates representative probes to improve behavioral coverage while reducing semantic redundancy.

The framework consists of three stages: (1) semantic query partitioning, (2) coordinated query construction, and (3) behavior reconstruction.

3.4.1 Semantic Query Partitioning

Let the initial query pool be

$$\mathcal{Q}_0 = \{q_1, q_2, \ldots, q_N\}.$$

Each query is mapped into a semantic embedding space using a sentence encoder:

$$e_i = E(q_i),$$

where $E(\cdot)$ denotes the embedding model.

We then perform unsupervised semantic partitioning over the embedding space to divide the query pool into K semantic regions:

$$\mathcal{Q}_0 = \bigcup_{k=1}^{K} \mathcal{Q}_k.$$

Each partition $\mathcal{Q}_k$ represents a semantically related subset of instruction behaviors.

3.4.2 Coordinated Query Construction

For each semantic partition $\mathcal{Q}_k$, we compute the centroid representation

$$c_k = \frac{1}{|\mathcal{Q}_k|} \sum_{q \in \mathcal{Q}_k} e(q),$$

where $e(q)$ denotes the embedding of query q.

Representative seed queries nearest to the centroid are selected to characterize the semantic region. These examples are then used to construct cluster-conditioned prompts for query generation.

A strong instruction-tuned LLM generates semantically coordinated probing queries:

$$\tilde{\mathcal{Q}}_k = \mathrm{LLM}(c_k, \mathcal{E}_k),$$

Model	BERT-F1	R1	RL	BLEU
Qwen2.5-7B (FP16)	84.51	20.93	14.87	2.81
Qwen2.5-7B (GPTQ-INT4)	84.27	20.79	14.69	2.77
Qwen3-1.7B (FP16)	81.24	19.45	13.72	2.49
Qwen3-1.7B (GPTQ-INT4)	80.98	19.27	13.56	2.43

Table 1: Effect of GPTQ-based quantization on teacher model performance. Quantization introduces only marginal degradation across all metrics.All scores are reported in percentage (%).

where $\mathcal{E}_k$ denotes representative examples retrieved from semantic partition $\mathcal{Q}_k$.

The final coordinated query pool is formed as

$$\mathcal{Q} = \bigcup_{k=1}^{K} \tilde{\mathcal{Q}}_k.$$

3.4.3 Behavior Reconstruction

Given the constructed query set $\mathcal{Q}$, we interact with the target model and collect query-response pairs

$$\{(x, T_\theta(x))\}_{x \in \mathcal{Q}},$$

where T_θ denotes the target model.

A student model S_ϕ is trained to imitate the teacher responses by minimizing

$$\mathcal{L}_{\mathrm{KD}} = \frac{1}{|\mathcal{Q}|} \sum_{x \in \mathcal{Q}} \ell\big(S_\phi(x), T_\theta(x)\big),$$

where $\ell(\cdot, \cdot)$ denotes the imitation loss.

The resulting student performance is used to measure the extent of behavioral knowledge recoverable through black-box query interaction under constrained budgets.

4 Experiments

4.1 Experimental Setup

Teacher and Student Models. We use models from the Qwen family as both teachers and students. Specifically, Qwen2.5-7B (Qwen et al., 2025) serves as the teacher model, while edge-oriented students with up to 4B parameters are used for extraction experiments. To reflect realistic on-device deployment, all student models are evaluated under aggressive quantization settings, including INT8 and INT4 using GPTQ.

Unless otherwise stated, both teacher inference and student training use quantized weights, ensuring that the evaluation faithfully reflects practical edge deployment constraints.

Query Construction Strategies. We compare two query construction strategies under identical query budgets:

(1) **Original Queries (OQ):** instruction queries sampled from the original dataset. The training queries and evaluation queries are drawn from the same dataset distribution, but consist of disjoint instruction instances to avoid overlap between extraction and evaluation.

(2) **CLIQ (Ours):** coordinated linguistic instruction queries constructed through semantic partitioning and structured query coordination to improve behavioral coverage while reducing semantic redundancy.

In all settings, training queries are randomly sampled from the corresponding query pool during student training. Thus, the only difference lies in how the query pool is constructed.

Query Budget. Our main comparison uses a fixed query budget of 500 queries. We additionally analyze smaller budgets (100–400) and training dynamics under a fixed budget of 500 queries.

Evaluation Metrics. We report BERT-F1 (Zhang et al., 2020), BLEU (Papineni et al., 2002), ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-Lsum (Lin, 2004). All scores are averaged over held-out evaluation sets.

4.2 Teacher Quantization Sanity Check

Before evaluating extraction performance, we first examine the effect of quantization on teacher models. Table 1 compares full-precision and GPTQ-quantized versions of Qwen2.5-7B and Qwen3-1.7B.

These results confirm that quantized teachers remain sufficiently expressive for query interaction and supervision.

4.3 Main Extraction Results

Table 2 presents the unified evaluation results under a fixed 500-query budget and 500 training steps. All extraction experiments use the INT4-quantized Qwen2.5-7B model as the target teacher, while the students are INT4-quantized Qwen1.5 models with different parameter scales.

We first compare the teacher and base student models. Although GPTQ quantization introduces only marginal degradation, the INT4-quantized Qwen2.5-7B teacher remains highly competitive across all evaluation metrics, indicating that substantial semantic knowledge is preserved after compression. In contrast, the unaligned student models perform significantly worse than the teacher, particularly on ROUGE-based metrics, revealing a substantial capability gap prior to extraction-based alignment.

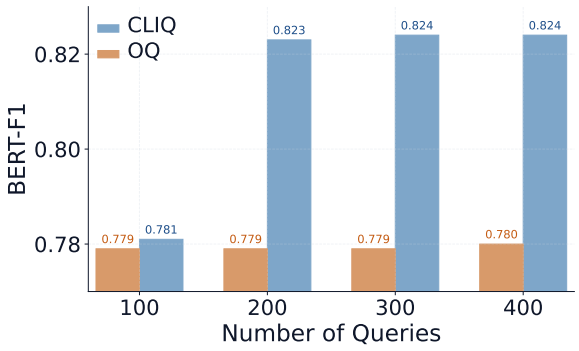

Figure 3: Effect of query budget on distillation performance. CLIQ exhibits rapid gains as the number of queries increases from 100 to 300, followed by diminishing returns, while Original Queries show minimal improvement.

To analyze extraction behavior under different model capacities, we evaluate two edge-oriented students with different parameter scales: a smaller 4B model and a larger 7B model. Across both settings, structured query supervision consistently achieves substantially stronger reconstruction performance than Original Queries (OQ).

On the smaller 4B student, CLIQ improves BERT-F1 from **77.92** to **82.44** and more than doubles the BLEU score compared with OQ. A similar trend is observed on the larger 7B student, where CLIQ achieves **83.04** BERT-F1 compared with **79.72** under OQ supervision.

Interestingly, the relative improvement of CLIQ becomes more pronounced on the smaller-capacity student model. This suggests that structured query supervision is particularly beneficial when model capacity is constrained and low-precision quantization further limits representational ability.

Overall, these results demonstrate that carefully structured queries can recover substantial behavioral knowledge from quantized edge models across different student model scales even under strict interaction budgets, highlighting a practical extraction risk for edge-deployed LLMs.

Model	Role	Teacher Quant.	Student Quant.	Query	BERT-F1	BLEU	R-1	R-2	R-L
Qwen2.5-7B	Teacher	FP16	–	–	83.69	2.94	22.08	9.36	16.85
		INT4	–	–	84.22	**3.15**	21.92	9.58	16.56
Qwen1.5-4B	Base	–	INT4	–	75.04	0.83	14.07	5.08	9.46
Qwen1.5-4B	Student	INT4	INT4	OQ	77.92	1.21	16.53	5.71	10.87
		INT4	INT4	CLIQ	**82.44**	**2.56**	**18.74**	**6.40**	**12.42**
Qwen1.5-7B	Base	–	INT4	–	78.02	1.04	17.03	5.51	10.87
Qwen1.5-7B	Student	INT4	INT4	OQ	79.72	1.48	17.88	6.04	11.77
		INT4	INT4	CLIQ	**83.04**	**2.72**	**19.01**	**6.38**	**12.55**

Table 2: Unified evaluation of query-based extraction across teacher, base, and aligned student models. Student results are reported under a fixed 500-query budget after 500 training steps. CLIQ substantially improves behavioral reconstruction over original query supervision (OQ), especially for INT4-quantized edge students. All scores are reported in percentage (%).

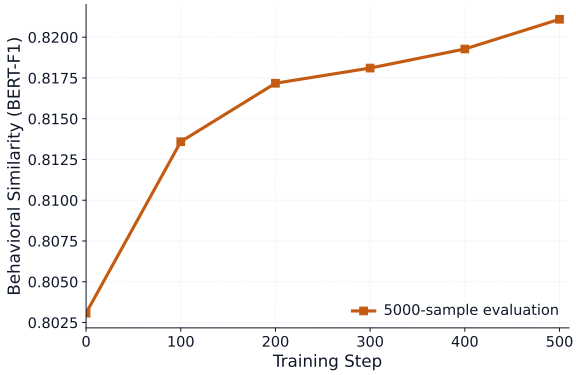

Figure 4: Progressive behavioral alignment between the CLIQ-trained student and the teacher model evaluated on Original Queries (OQ) using BERT-F1 under a fixed query budget.

Quant.	Query	Steps	BERT-F1	BLEU	ROUGE-L
INT4	Original	100	80.32	1.55	13.83
		300	80.44	1.49	14.12
		500	80.03	1.42	13.41
	CLIQ	100	78.20	1.08	12.45
		300	82.86	2.56	14.22
		500	**82.92**	**2.55**	**14.54**
INT8	Original	100	80.14	1.50	13.49
		300	80.17	1.41	13.51
		500	80.08	1.43	13.49
	CLIQ	100	78.18	1.07	12.53
		300	82.70	2.44	14.48
		500	82.74	2.47	14.51

Table 3: Effect of training steps under different quantization settings. CLIQ shows consistent improvements from 100 to 300 steps, while Original queries exhibit minimal gains.

4.4 Query Budget and Sample Efficiency

As shown in Figure 3, CLIQ exhibits rapid performance gains as the query budget increases from 100 to 300, followed by diminishing returns beyond this regime.

This trend suggests that structured semantic allocation efficiently captures dominant behavioral regions of the teacher model under limited interaction budgets. In contrast, Original Queries repeatedly probe redundant semantic regions and therefore exhibit substantially weaker sample efficiency.

4.5 Progressive Behavioral Reconstruction

Figure 4 illustrates how the CLIQ-trained student progressively approaches the target teacher behavior during optimization under a fixed query budget of 500 queries.

Unlike standard training-loss analysis, evaluation is performed on Original Queries (OQ), measuring generalization to the original instruction distribution rather than memorization of generated queries. The student exhibits rapid behavioral alignment during early optimization before gradually saturating at later training steps, suggesting that structured semantic allocation efficiently captures dominant behavioral regions of the teacher model under constrained interaction budgets.

Importantly, although training is performed using CLIQ-generated queries, teacher-matching performance on the original query distribution consistently improves throughout optimization, indicating that the learned behaviors generalize beyond the generated query set.

Table 3 further analyzes training dynamics under different quantization settings. Across both INT4 and INT8, CLIQ achieves rapid improvements from 100 to 300 training steps and consis-

Query Strategy	INT8	INT4
OQ	80.08	80.03
CLIQ (ours)	**82.74**	**82.92**

Note: BERT-F1 scores are reported. All results are obtained with a fixed query budget (500 queries) and 500 training steps. All scores are reported in percentage (%).

Table 4: Effect of quantization on different query construction strategies using a 1.8B edge-oriented student model.

tently outperforms OQ, which exhibits only limited gains throughout optimization.

Additional training dynamics across different student scales are provided in Appendix A.2.

4.6 Effect of Quantization

Finally, we evaluate extraction performance under different quantization levels.

Across both INT8 and INT4 settings, CLIQ consistently outperforms Original Queries, indicating that structured semantic allocation remains effective even under aggressive low-precision deployment.

Although stronger quantization slightly reduces overall reconstruction quality, the performance gap between CLIQ and OQ remains substantial. This suggests that quantization alone does not eliminate behavioral extractability under query-based interaction.

5 Analysis and Discussion

In this section, we analyze why structured semantic query coordination improves behavioral extraction from quantized Edge LLMs under constrained interaction budgets.

5.1 Why Does Semantic Coordination Help?

Edge-deployed LLMs are typically constrained by aggressive quantization and reduced model capacity, making them more sensitive to noisy and redundant supervision signals.

Naive instruction sampling often produces highly overlapping queries that repeatedly probe similar behaviors, resulting in inefficient use of limited interaction budgets. CLIQ mitigates this issue by organizing the instruction space into semantically coherent regions and coordinating representative probing queries across different semantic behaviors.

Importantly, the optimization procedure remains identical across all experimental settings. The observed improvements mainly originate from the semantic structure of the query pool itself rather than changes to the training algorithm. This effect becomes more pronounced under aggressive quantization settings such as INT4, where structured supervision helps reduce the impact of quantization-induced noise.

5.2 Behavioral Reconstruction under Limited Budgets

Our results show that substantial behavioral knowledge remains recoverable even under strict interaction budgets.

As shown in Figure 3, CLIQ achieves rapid improvements in low-budget regimes and consistently outperforms randomly sampled queries. This suggests that extraction efficiency depends more on semantic query organization than on query quantity alone.

More broadly, these findings indicate that quantization alone should not be viewed as an effective defense against query-based behavioral extraction. Although quantization perturbs model outputs, carefully coordinated semantic queries can still recover substantial behavioral knowledge from edge-deployed LLMs.

6 Conclusion

We investigate the security risks of query-based behavioral extraction from quantized edge-deployed large language models under constrained interaction budgets. Our results show that quantization alone does not effectively prevent behavioral knowledge leakage through black-box interaction. To systematically analyze this overlooked risk, we propose CLIQ (**C**oordinated **L**inguistic **I**nstruction **Q**uerying), a structured semantic probing framework that coordinates representative instruction queries across different semantic regions. Extensive experiments demonstrate that carefully coordinated semantic queries substantially improve behavioral reconstruction efficiency compared with naive query sampling, particularly under aggressive quantization and low-budget interaction settings. These findings suggest that semantic query coordination plays a critical role in model extractability and indicate that current edge deployment practices may provide weaker protection against model extraction than previously assumed.

7 Limitations

We investigate the security risks of query-based behavioral extraction from quantized edge-deployed large language models under constrained interaction budgets. However, several important limitations remain.

First, publicly available official edge-deployed quantized LLMs remain relatively limited and often depend on deployment-specific inference stacks or proprietary optimization pipelines. To ensure reproducibility and controlled analysis, our experiments focus on widely adopted open-source quantization settings. Evaluating extraction risks on newer commercial edge deployment systems remains important future work.

Second, our experiments primarily analyze instruction-following behaviors under standard black-box interaction settings. More complex deployment scenarios, including agentic systems, retrieval-augmented pipelines, and tool-augmented interactions, may exhibit different extraction characteristics and require further investigation.

Finally, while our results demonstrate that quantization alone does not effectively prevent behavioral knowledge leakage, we do not explore potential defense mechanisms against structured semantic probing. Developing robust extraction-resistant deployment strategies remains an important direction for future research.

8 Ethical Considerations

This work studies query-efficient knowledge extraction from edge-deployed and quantized LLMs, which raises potential concerns regarding model misuse and unauthorized behavioral transfer.

Model Extraction Risks. Our results show that quantization does not eliminate the risk of knowledge extraction under query-based interaction. Although CLIQ does not recover model parameters or exact output distributions, it improves the efficiency of behavioral imitation under limited query budgets. This highlights a practical security risk in edge deployment, rather than a vulnerability specific to our method.

Responsible Use and Compliance. All experiments are conducted using publicly available data and standard, rate-limited API access. We do not bypass safeguards or violate usage policies. We emphasize that applying such techniques should comply with model licenses, terms of service, and local regulations.

Broader Impact. Our goal is to better understand the security implications of deploying LLMs on resource-constrained devices. By identifying how structured queries affect extractability, this work can inform the design of more robust deployment strategies and defenses against misuse.

References

Hongzhan Chen, Ruijun Chen, Yuqi Yi, Xiaojun Quan, Chenliang Li, Ming Yan, and Ji Zhang. 2024. Knowledge distillation of black-box large language models. *arXiv e-prints*, pages arXiv–2401.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. In *Advances in Neural Information Processing Systems*, volume 36, pages 10088–10115. Curran Associates, Inc.

Yingpeng Du, Zhu Sun, Ziyan Wang, Haoyan Chua, Jie Zhang, and Yew-Soon Ong. 2025. Active large language model-based knowledge distillation for session-based recommendation. In *Proceedings of the 39th Annual Conference on Artificial Intelligence (AAAI-25)*, pages 11607–11615. AAAI Press.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. 2022. GPTQ: accurate post-training quantization for generative pre-trained transformers. *CoRR*, abs/2210.17323.

Amir Gholami, Sehoon Kim, Zhen Dong, Zhewei Yao, Michael W. Mahoney, and Kurt Keutzer. 2021. A survey of quantization methods for efficient neural network inference. *CoRR*, abs/2103.13630.

Google. 2023. Introducing gemini: our largest and most capable ai model. Google Blog.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. 2024. Minillm: Knowledge distillation of large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, and 1 others. 2022. Training compute-optimal large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, pages 30016–30030.

Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. 2023. Unnatural instructions: Tuning language models with (almost) no human labor. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 14409–14428. Association for Computational Linguistics.

Yoon Kim and Alexander M. Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 1317–1327. The Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Leland McInnes and John Healy. 2018. UMAP: uniform manifold approximation and projection for dimension reduction. *CoRR*, abs/1802.03426.

Markus Nagel, Marios Fournarakis, Rana Ali Amjad, Yelysei Bondarenko, Mart van Baalen, and Tijmen Blankevoort. 2021. A white paper on neural network quantization. *CoRR*, abs/2106.08295.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

Qualcomm. 2024. Qualcomm brings the generative ai revolution to devices. Qualcomm Press Release.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Samsung Electronics. 2024. Galaxy ai. Samsung Official Website.

D. Sculley. 2010. Web-scale k-means clustering. In *Proceedings of the 19th International Conference on World Wide Web, WWW 2010, Raleigh, North Carolina, USA, April 26-30, 2010*, pages 1177–1178. ACM.

Youwei Shu, Shaomian Zheng, Dingnan Jin, Wenjie Qu, Ziyao Guo, Qing Cui, Jun Zhou, and Jiaheng Zhang. 2026. On representation redundancy in large-scale instruction tuning data selection. *arXiv preprint arXiv:2602.13773*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 13484–13508. Association for Computational Linguistics.

Jie Zhang, Chen Chen, and Lingjuan Lyu. 2023. IDEAL: query-efficient data-free learning from black-box models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. 2024. Tinyllama: An open-source small language model. *CoRR*, abs/2401.02385.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Yue Zheng, Yuhao Chen, Bin Qian, Xiufang Shi, Yuanchao Shu, and Jiming Chen. 2025. A review on edge large language models: Design, execution, and applications. *ACM Comput. Surv.*, 57(8):209:1–209:35.

A Appendix

A.1 CLIQ Pipeline Details

A.1.1 Semantic Query Representation and Partitioning

Query Pool Construction. Given a raw instruction dataset in JSON format, we construct a text query for each sample by concatenating the `instruction` field and the optional `input` field:

$$q = \texttt{instruction} \oplus \texttt{input}.$$

If `input` is empty, we use `instruction` only.

Sentence Embeddings. Each query is encoded using the sentence encoder `sentence-transformers/all-MiniLM-L6-v2`. We apply mean pooling with attention masking followed by ℓ_2 normalization:

$$\mathbf{e}(q) = \text{Normalize}\big(\text{MeanPool}(\text{LM}(q))\big).$$

The maximum sequence length is set to 512 with a batch size of 64.

Semantic Partitioning. All query embeddings are partitioned into semantic regions using MiniBatchKMeans with $K = 100$ and a fixed random seed of 42. MiniBatchKMeans (Sculley, 2010) is adopted for scalability and memory efficiency under large query pools.

Filtering Sparse Semantic Regions. Semantic regions whose size falls below a threshold of 5 are removed to avoid unstable semantic prototypes and excessively sparse supervision. The remaining regions are then re-indexed to ensure contiguous IDs.

A.1.2 Coordinated Linguistic Query Construction

Semantic Region-conditioned Prompting. For each semantic region, we construct a prompt containing up to M in-region examples (default: $M = 1000$) to characterize the local instruction semantics and linguistic patterns. In the main experiments, we generate $m = 10$ coordinated instruction queries per semantic region unless otherwise specified. For visualization-oriented analyses, we additionally report a more query-efficient setting with $m = 5$.

Teacher API and Decoding. We use a Qwen-family teacher accessed through an OpenAI-compatible API endpoint. The decoding temperature is set to 0.7, and the maximum token budget is set to 16384 to reduce truncation.

Component	Setting
Sentence encoder	`sentence-transformers/all-MiniLM-L6-v2`
Max length	512
Embedding batch size	64
Pooling	mean pooling w/ attention mask
Normalization	ℓ_2 normalization
Semantic partitioning	MiniBatchKMeans
Semantic regions K	100
Seed	42
MiniBatch size	$\min(1000, \lvert\mathcal{Q}\rvert/10)$
Region size filter	min size = 5
Teacher API model	`Qwen/Qwen3-30B-A3B-Instruct-2507`
Temperature	0.7
Max tokens	16384
Timeout	300s
Retries	5 (exponential backoff)
Examples per region M	1000
Generated queries per region m	5

Table 5: Hyperparameters for semantic partitioning and coordinated linguistic query construction.

Algorithm 1: CLIQ Pipeline

Input: Query pool $\mathcal{Q}$; semantic regions K; minimum region size s; examples per region M; generated queries per region m
Output: Generated query set $\widetilde{\mathcal{Q}}$
Encode each query $q \in \mathcal{Q}$ into $\mathbf{e}(q)$;
Partition embeddings into K semantic regions via MiniBatchKMeans;
Remove regions with size $< s$ and re-index remaining regions;
$\widetilde{\mathcal{Q}} \leftarrow \emptyset$;
foreach *semantic region* c **do**
  Sample up to M in-region queries;
  Construct a semantic region-conditioned prompt;
  Query the teacher to generate m coordinated queries;
  Parse and repair JSON outputs;
  $\widetilde{\mathcal{Q}} \leftarrow \widetilde{\mathcal{Q}} \cup$ generated queries;
return $\widetilde{\mathcal{Q}}$;

Robust Output Parsing. The teacher is instructed to output a JSON array of {`instruction`, `input`} pairs. We apply a robust parser that extracts JSON content from markdown blocks and repairs truncated outputs by retaining the last complete object and closing the array.

A.1.3 Implementation Details and Hyperparameters

Table 5 summarizes the key hyperparameters used in semantic partitioning and coordinated linguistic query construction. All settings are fixed across experiments unless otherwise stated.

A.1.4 Comprehensive Evaluation Results

Table 2 reports the comprehensive evaluation results under different model, quantization, and query settings. These results complement the main paper by presenting the full metric set across teacher

models, base students, and extraction-trained students.

Table 6 further reports additional unified evaluation results under different teacher quantization, student quantization, and semantic query organization settings. The results show that quantized teacher models remain highly extractable despite low-bit compression, while semantically coordinated querying consistently improves behavioral reconstruction quality compared with naive query sampling.

Overall, the results show that quantized teachers remain strong despite low-bit compression, while unaligned student models exhibit a substantial behavioral gap. Both Original Queries (OQ) and CLIQ improve student performance, but CLIQ consistently achieves stronger recovery across evaluation metrics. This confirms that the extraction risk persists even when the target model is quantized, and that semantic query organization plays a central role in determining extraction effectiveness.

A.1.5 Training Dynamics under a Fixed Query Budget

Figure 5 analyzes the training dynamics under a fixed budget of 500 queries across three edge-oriented student models. The figure reports BERT-F1, BLEU, and ROUGE-L, with the 0-step base model included as initialization.

Across all models and metrics, CLIQ shows clear and consistent advantages over Original Queries. For BERT-F1, CLIQ yields rapid gains within the first 200 training steps and then saturates around 200–300 steps, indicating diminishing returns after the structured supervision signal has been largely absorbed. In contrast, OQ-trained models remain nearly flat across training steps, suggesting that additional optimization alone cannot compensate for redundant and low-coverage queries.

The same trend appears for BLEU and ROUGE-L. Under OQ supervision, both metrics remain largely unchanged as training progresses. By contrast, CLIQ enables sustained improvement, especially for smaller and more heavily quantized students. This indicates that CLIQ does not merely accelerate convergence; rather, it provides a more informative query-response distribution that changes the student's learning dynamics under a fixed interaction budget.

Effect of Quantization Settings. Table 3 further studies training steps under INT4 and INT8 quantization using the same student model and randomly sampled training queries. The only difference lies in how the query pool is constructed: Original Queries versus CLIQ. Two consistent patterns emerge. First, with Original Queries, performance remains nearly flat from 100 to 500 steps and may even slightly regress. Second, with CLIQ, performance improves sharply from 100 to 300 steps and then saturates from 300 to 500 steps. This contrast suggests that CLIQ provides a higher-quality query distribution that the student can progressively absorb through optimization, whereas the Original pool contains substantial redundancy that limits the benefit of additional training.

A.1.6 Clustering Granularity and Query Allocation

Table 7 ablates two key design choices in CLIQ: the number of semantic clusters K and the number of generated queries per cluster m. The evaluation is conducted on both Dolly and Alpaca.

Clustering Granularity. Across both datasets, moderate granularities, such as 25–50 clusters, tend to produce stable and competitive results. Overly fine-grained clustering, such as $K = 100$, exhibits higher variance and can degrade performance under certain query allocations. This is consistent with the intuition that excessive partitioning may fragment the instruction space into overly narrow regions, making each cluster less informative under a fixed query budget.

Queries per Cluster. Increasing the number of generated queries per cluster generally improves performance when moving from 4 to 8 or 10 queries. However, the gains are not strictly monotonic and eventually diminish. This shows that CLIQ benefits primarily from improved semantic coverage and representative supervision, rather than from simply increasing the number of queries.

Coverage–Redundancy Trade-off. The trends in Table 7 can be understood as a coverage–redundancy trade-off. Original query sampling corresponds to i.i.d. draws from an imbalanced mixture of latent semantic clusters. Dense clusters are repeatedly sampled, while sparse clusters require much larger budgets to be reliably covered. As a result, increasing the query budget under Original Queries often produces redundant supervision

Model	Role	Teacher Quant.	Student Quant.	Query	BERT-F1	BLEU	R-1	R-2	R-L
Qwen2.5-7B	Teacher	FP16	–	–	83.69	2.94	22.08	9.36	16.85
		INT4	–	–	84.22	**3.15**	21.92	9.58	16.56
Qwen3-1.7B	Base	–	FP16	–	81.53	1.70	12.35	4.46	8.92
		–	INT8	–	81.40	1.73	12.23	4.36	8.70
	Student	FP16	INT8	OQ	83.75	2.50	22.30	9.49	16.68
		FP16	INT8	OQ (valid)	83.78	2.46	22.24	9.57	16.83
		INT4	INT8	CLIQ	**84.35**	3.04	**23.42**	**10.60**	**17.60**
		INT4	INT8	CLIQ (valid)	84.23	2.92	23.07	10.26	17.27

Table 6: Additional unified evaluation results under different teacher quantization, student quantization, and semantic query organization settings.

Dataset	Clusters	Queries	BERT-F1	BLEU	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum
Dolly	25	4	81.384	2.092	17.905	6.196	12.089	14.402
	25	8	82.272	2.325	18.325	6.263	12.397	14.686
	25	10	82.267	2.473	18.546	6.258	12.550	15.017
	50	4	82.558	2.643	18.335	6.377	12.258	14.825
	50	10	82.345	2.608	18.399	6.268	12.500	14.818
	100	8	78.707	1.476	16.895	5.895	11.298	13.499
	100	10	**83.042**	**2.972**	**18.831**	**6.556**	**12.667**	**15.138**
Alpaca	25	4	81.582	1.861	17.788	5.711	11.650	14.146
	25	8	81.451	2.010	17.749	5.873	11.563	14.039
	25	10	81.936	2.256	18.177	6.088	11.977	14.508
	50	4	80.914	1.885	17.701	6.060	11.669	14.165
	50	10	**82.518**	**2.460**	18.133	6.084	11.980	14.525
	100	4	82.056	2.236	**18.379**	**6.160**	**12.052**	**14.617**
	100	8	78.714	1.307	16.691	5.608	10.928	13.239

Table 7: Additional ablation results on clustering granularity and query budget on a 4B model quantized by int4. We vary the number of semantic clusters and the number of generated queries per cluster, and report instruction-following performance on Dolly and Alpaca. All scores are reported in percentage (%).

and early performance saturation.

In contrast, CLIQ distributes queries across semantic clusters. Under a fixed budget $B = K \cdot m$, increasing K improves global semantic coverage, while increasing m improves intra-cluster diversity. Balanced configurations, such as moderate K with sufficient m, provide the best trade-off between coverage and stability. When K is too large relative to the budget, each cluster receives too few queries, resulting in sparse and unstable supervision. When m becomes too large, additional queries provide diminishing returns once dominant cluster modes have already been captured.

A.2 Semantic Analysis of Query Space

The quantitative results above show that CLIQ improves extraction efficiency under strict query budgets. This section explains why this happens by analyzing the query space directly. We focus on four complementary properties: semantic coverage, cluster-level allocation, intra-cluster redundancy, and intra-cluster diversity.

A.2.1 Semantic Coverage under Limited Budgets

Figure 6 reports the number of semantic clusters covered as a function of the query budget. This directly measures how efficiently a query strategy probes different regions of the latent instruction space.

Compared with random sampling from the original query pool, CLIQ covers more semantic clusters under small and moderate budgets. This early-stage coverage advantage provides more informative supervision during initial training, which helps explain the faster improvement and earlier saturation observed in Appendix A.1.5. The result also supports the central claim that extraction efficiency is governed not only by the number of queries, but also by how those queries are organized.

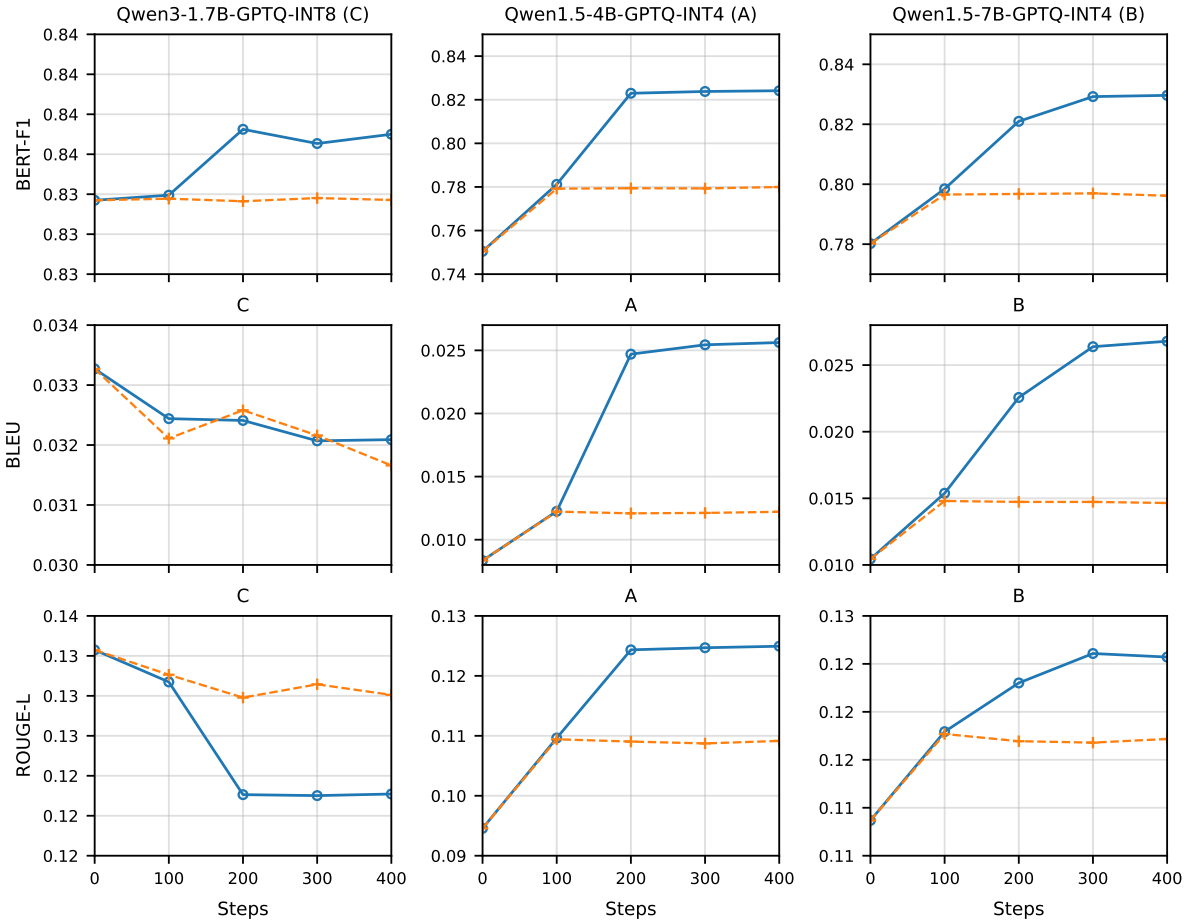

Figure 5: Effect of training steps under a fixed query budget (500 queries) across three student models. We report BERT-F1, BLEU, and ROUGE-L, and include the 0-step base model as initialization. CLIQ (ours) consistently outperforms OQ and exhibits diminishing returns after ∼200–300 steps.

A.2.2 Cluster-level Query Allocation

Figure 7 visualizes the number of queries assigned to each cluster under a fixed query budget. This analysis explains the source of the coverage gap observed in Figure 6.

Original Queries produce a highly imbalanced allocation: a small number of dense semantic regions receive many queries, while other clusters receive few or none. This leads to repeated probing of similar model capabilities and leaves parts of the instruction space under-explored. In contrast, CLIQ explicitly allocates queries across semantic clusters, yielding a more balanced distribution. This balanced allocation is the mechanism through which CLIQ improves coverage without increasing the total query budget.

A.2.3 Intra-cluster Redundancy

Coverage alone is not sufficient; queries within the same cluster should also avoid excessive duplication. Figure 8 analyzes intra-cluster redundancy using average pairwise cosine similarity among queries in each semantic cluster.

Original Queries exhibit substantially higher intra-cluster similarity, indicating that many queries provide overlapping supervision. This explains why OQ-trained models are largely insensitive to additional training steps: the student repeatedly observes similar behaviors and receives limited new information as optimization proceeds. By contrast, CLIQ reduces intra-cluster redundancy, providing more diverse supervision signals within the same semantic region. This helps explain why CLIQ-trained models continue to improve during early training and achieve stronger final extraction performance.

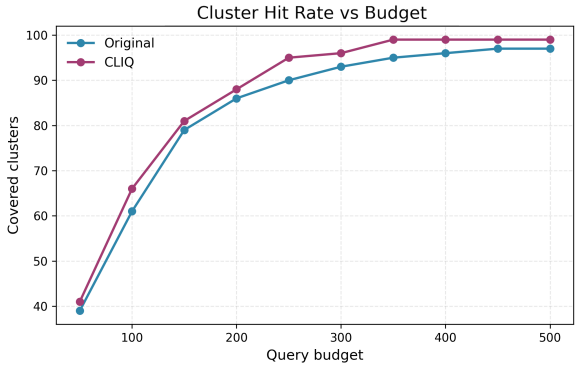

Figure 6: Cluster-level semantic coverage under different query budgets. Compared with Original Queries (OQ), CLIQ achieves broader semantic coverage under constrained interaction budgets, indicating that structured query allocation probes a larger portion of the latent capability space with fewer redundant interactions.

A.2.4 Global Semantic Distribution

Figure 10 visualizes the global semantic distribution of different query pools using UMAP (McInnes and Healy, 2018) projections of sentence embeddings. Each point corresponds to a single query, and all projections are computed using the same embedding model for fair comparison.

The original query pool forms highly concentrated regions in the embedding space, with substantial local overlap among queries. This confirms the redundancy observed in Figure 8 and provides a qualitative explanation for the weak training dynamics of OQ-trained students.

In contrast, CLIQ-generated query pools are more dispersed and uniform. Across different configurations, CLIQ expands semantic coverage while reducing local redundancy, enabling a smaller number of queries to span a broader instruction space. This qualitative pattern aligns with the quantitative ablation results in Table 7, where balanced CLIQ configurations achieve stronger performance under modest budgets.

The visualization also clarifies the effect of clustering granularity. Overly fine-grained clustering combined with insufficient queries per cluster can produce sparse coverage and visible gaps in the embedding space. These gaps correspond to performance degradation in Table 7, suggesting that excessive partitioning fragments supervision into clusters that are too small to provide stable learning signals. Moderate clustering granularities with sufficient per-cluster queries provide a more favorable balance between coverage and density.

A.2.5 Intra-cluster Diversity and Semantic Representativeness

Finally, we examine whether CLIQ simply generates cluster prototypes or preserves meaningful diversity within each semantic region. Figures 9 and 11 show that CLIQ-generated queries remain semantically coherent while covering both central and peripheral regions within clusters.

Figure 9 shows that CLIQ-generated queries span a wider range of distances from their corresponding cluster centroids. This means that CLIQ does not merely generate near-centroid prototypes. Instead, it preserves structured diversity within each cluster, covering both typical and less central instructions.

Figure 11 provides a cluster-conditioned view of this phenomenon. Although only $m = 5$ generated queries per cluster are visualized, the generated queries consistently align with high-density regions of their corresponding clusters. Rather than collapsing to global centroids, CLIQ-generated queries capture cluster-internal modes and remain representative of the original semantic region. This complements the global UMAP analysis and supports the interpretation that CLIQ improves extraction by preserving semantic representativeness while substantially reducing the number of required queries.

Taken together, the semantic analyses in this section provide a unified mechanistic explanation for the quantitative results. CLIQ improves extraction not by increasing the total number of interactions or training steps, but by restructuring the query distribution to achieve broader coverage, more balanced cluster allocation, lower redundancy, and stronger intra-cluster diversity.

A.3 Cluster-conditioned Query Examples

The following examples correspond to semantic regions visualized in Figure 10 and Figure 11.

A.4 Use of AI Assistants

AI-assisted tools were used in a limited and supportive manner during the preparation of this manuscript. Specifically, large language models such as ChatGPT were used to assist with language polishing, clarity improvement, structural refinement of the text, and minor LaTeX formatting suggestions.

The AI tools were not used to generate experimental results, derive theoretical claims, design algorithms, conduct data analysis, or produce figures or tables. All technical content, experimental

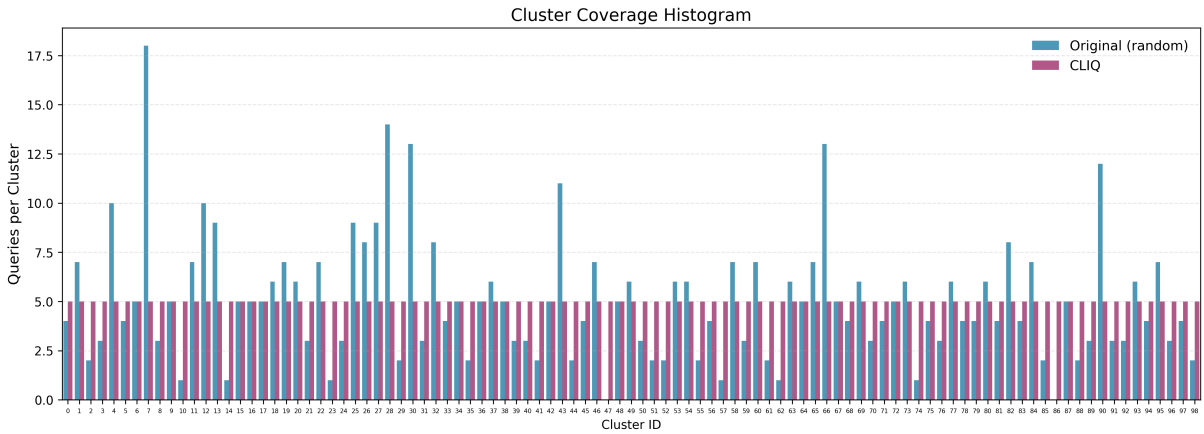

Figure 7: Histogram of query counts per cluster under a fixed query budget. Random sampling from Original Queries results in highly imbalanced cluster coverage, whereas CLIQ produces a near-uniform allocation across clusters, avoiding over-sampling of dense regions.

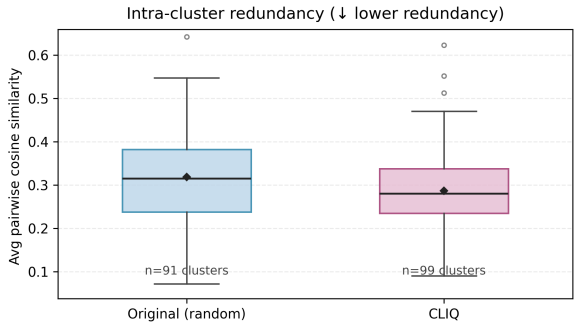

Figure 8: Intra-cluster redundancy measured by average pairwise cosine similarity within each semantic cluster. Original Queries exhibit higher intra-cluster similarity, indicating strong redundancy and overlapping supervision signals. CLIQ reduces redundancy by promoting greater diversity within clusters.

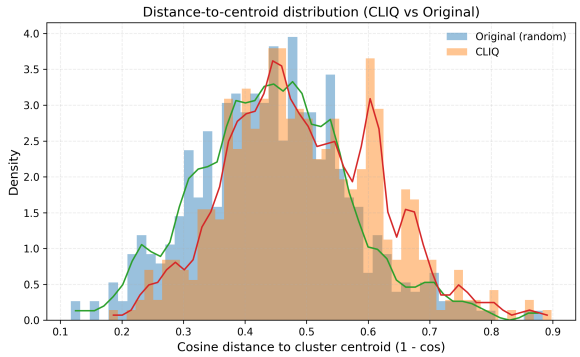

Figure 9: Distribution of cosine distances between queries and their corresponding cluster centroids. CLIQ-generated queries span a wider range of distances, covering both central and peripheral regions of each cluster. This indicates that CLIQ does not collapse queries to cluster prototypes, but preserves structured intra-cluster diversity while maintaining semantic coherence.

design, analysis, and conclusions were conceived, verified, and finalized by the authors.

The authors take full responsibility for the correctness, originality, and integrity of the content presented in this paper.

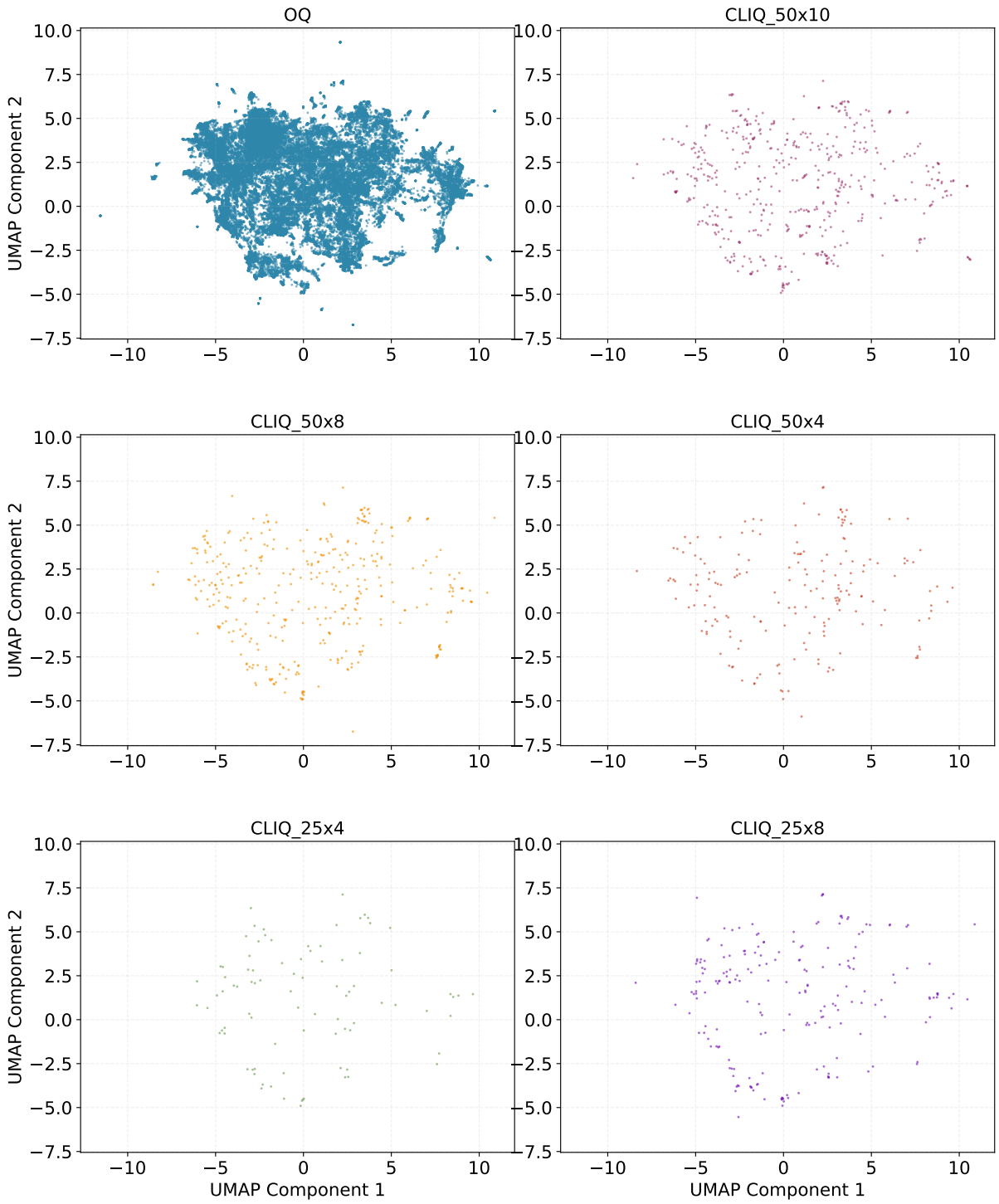

Figure 10: UMAP visualization of query coverage under different CLIQ configurations. The original query pool exhibits dense and redundant regions, while CLIQ-generated queries maintain broader and more uniform semantic coverage under fixed query budgets. Different combinations of cluster numbers and queries per cluster illustrate how CLIQ controls coverage granularity.

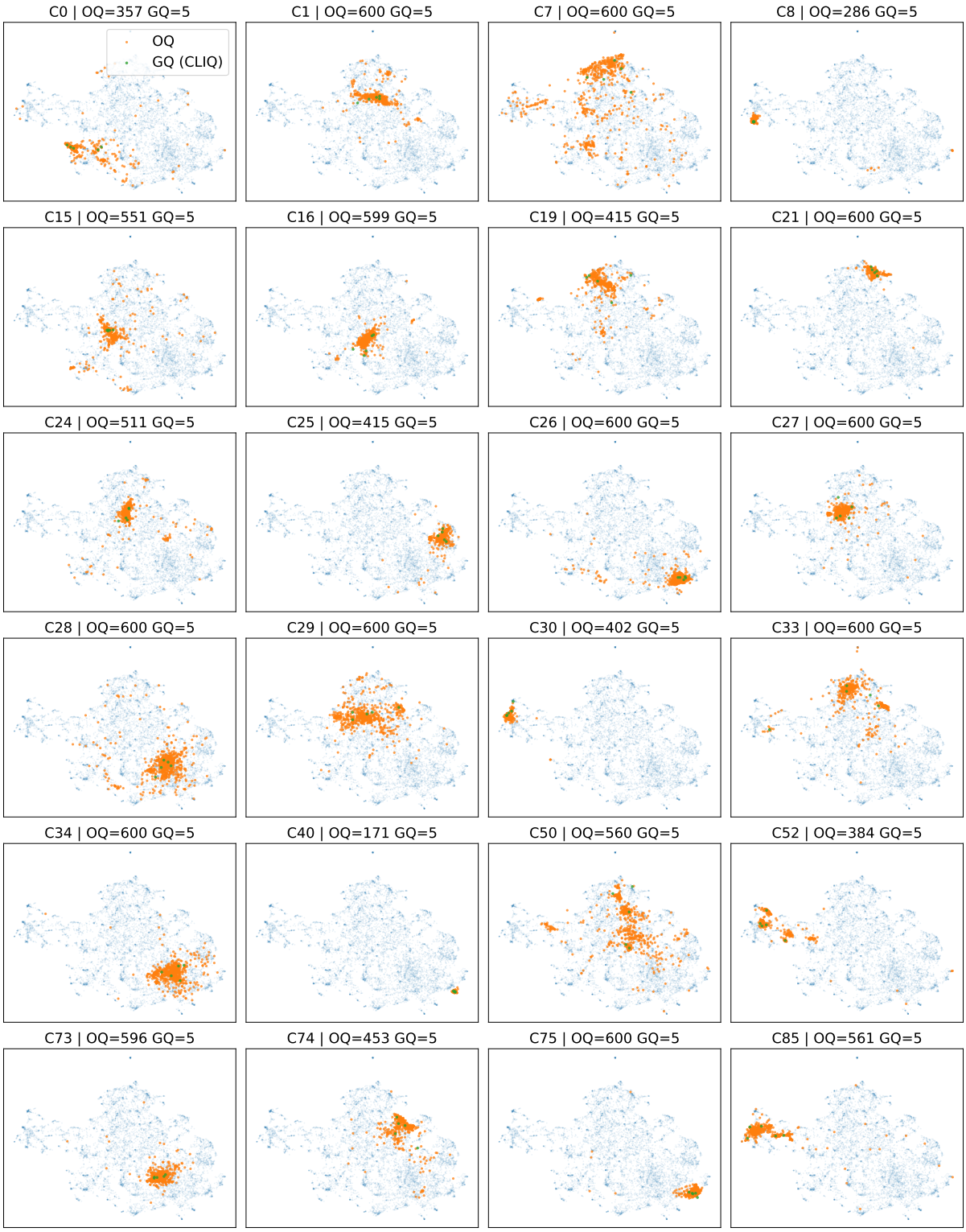

Figure 11: Cluster-conditioned semantic projections of Original Queries (OQ) and CLIQ-generated queries (GQ). Each subfigure corresponds to one semantic cluster c. Blue points denote original queries in the cluster, while orange points denote generated queries. Only $m = 5$ generated queries per cluster are visualized to illustrate the query efficiency of CLIQ.

Concept Explanation Cluster (Cluster 7)

Cluster Theme. Definition-oriented and explanatory instructions involving scientific concepts, algorithms, and general knowledge.

Original Queries	CLIQ-generated Queries
What does DNA stand for?	**Explain the concept of a zero-knowledge proof in cryptography using a real-world analogy.** *A scenario where one person proves to another that they know a secret without revealing the secret itself*
Explain the concept of a bubble sort algorithm to a non-technical audience.	**Describe how a transformer architecture works in natural language processing, focusing on attention mechanisms.**
Explain the concept of a linear regression model using simple language.	**Explain the difference between supervised and unsupervised learning with one practical example for each.**
Explain what an API is.	**What is the purpose of a firewall in network security, and how does it protect against cyber threats?**
Explain the concept of entropy in thermodynamics.	**Summarize the concept of edge computing and its advantages over traditional cloud computing in IoT applications.**

Mathematical Reasoning Cluster (Cluster 30)

Cluster Theme. Mathematical reasoning tasks involving algebraic manipulation, derivatives, and symbolic computation.

Original Queries	CLIQ-generated Queries
Convert the given equation into an algebraic expression. $3x + 5y = 9$	**Find the vertex of the parabola defined by the quadratic equation** $y = 2x^2 - 8x + 5$**.** $y = 2x^2 - 8x + 5$
Find the solution to the quadratic equation. $3x^2 + 7x - 4 = 0$	**Solve the system of equations using substitution method:** $3x + y = 10$ **and** $2x - 3y = -5$**.** $3x + y = 10,\ 2x - 3y = -5$
Compute the following expression. $2^3 + 3^2$	**Evaluate the limit of the function** $f(x) = (x^2 - 9)/(x - 3)$ **as** x **approaches** 3**.** $f(x) = (x^2 - 9)/(x - 3),\ x \to 3$
Find the slope of the following line. $y = 3x - 2$	**Determine the area under the curve** $y = x^2$ **from** $x = 0$ **to** $x = 3$ **using integration.** $y = x^2,\ x = 0$ *to* $x = 3$
Compute the derivative of $2x^2 + 5x$**.**	**Find the magnitude and direction of the vector** $v = 4i - 3j$**.** $v = 4i - 3j$

Haiku Writing Cluster (Cluster 40)

Cluster Theme. Short-form creative writing tasks centered on haiku generation and nature-oriented imagery.

Original Queries	CLIQ-generated Queries
Generate a haiku poem.	**Compose a haiku that captures the essence of a solitary traveler at dawn.** *Traveler walking through misty hills at sunrise*
Compose a haiku about cats.	**Create a haiku using only sensory details from the natural world, focusing on sound and touch.** *Forest at twilight, wind rustling through dry leaves*
Design a haiku about snow.	**Write a haiku that metaphorically represents the passage of time using seasonal imagery.** *Cherry blossoms fall like time slipping through fingers*
Write a Haiku of 3 lines, with the first containing 5 syllables and the remaining lines containing 7 syllables.	**Generate a haiku about a moment of unexpected beauty in an ordinary setting.** *A cracked sidewalk sprouting green through concrete*
Write a haiku with the following phrase: "frog in the pond".	**Craft a haiku that reflects on inner peace using elements of water and stillness.** *Still pond after rain, ripple fades into silence*

Figure 12: Representative original and CLIQ-generated queries from different semantic clusters. CLIQ preserves cluster-level semantics while improving semantic diversity and coverage.