CLIQ: On-Device Large Language Models Extraction

Anonymous ACL submission

Abstract

Large language models (LLMs) are increasingly deployed on edge devices under strict computation, memory, and quantization constraints. In such settings, extracting or distilling knowledge from heavily quantized on-device LLMs poses a fundamentally different challenge from conventional cloud-based distillation, due to limited query budgets and amplified quantization noise. We propose **CLIQ** (**Cl**ustered **I**nstruction **Q**uerying), a query-efficient distillation framework designed for extracting knowledge from quantized on-device LLMs. CLIQ explicitly models the semantic structure of the instruction space by clustering queries and generating a compact set of cluster-aware, representative instructions, thereby improving semantic coverage while reducing redundancy. Extensive experiments on quantized Qwen-family models under INT8 and INT4 settings show that, under identical query budgets, CLIQ consistently outperforms original query sampling across BERTScore, BLEU, and ROUGE metrics. Our results demonstrate that structured, semantically representative supervision is critical for effective distillation of edge-oriented language models.

1 Introduction

Large language models (LLMs) have demonstrated strong instruction-following capabilities across a wide range of tasks.Driven by demands for privacy preservation, low latency, and offline availability, LLMs are increasingly being deployed directly on edge devices such as mobile phones, personal assistants, and embedded systems.This rapid expansion of edge-side LLM deployment significantly broadens the attack surface of modern language models and raises new security concerns, including unauthorized model extraction and misuse. (Zheng et al., 2025; Li et al., 2024).

Despite their practical advantages, edge-deployed LLMs raise new and urgent *security concerns* (Miranda et al., 2025). Unlike cloud-hosted models, edge-side LLMs are directly accessible to adversaries through on-device inference APIs, making them particularly vulnerable to model extraction and misuse. Protecting the intellectual property and safety of these models therefore becomes a critical challenge as their real-world deployment scales. To meet strict constraints on computation, memory, and energy consumption, edge-side LLMs are typically *small in capacity and heavily quantized* (e.g., INT8 or INT4). While quantization enables efficient deployment, it substantially reduces model expressivity and amplifies sensitivity to noisy or redundant supervision signals (Dettmers et al., 2023; Nagel et al., 2021). These characteristics fundamentally differentiate edge-oriented LLMs from their full-precision cloud counterparts and render many existing attack and extraction techniques ineffective or inefficient.

Most prior model extraction and knowledge distillation approaches were originally developed for full-precision deep neural networks or cloud-based LLMs, where querying a powerful teacher model is relatively inexpensive (Hinton et al., 2015; Zhang et al., 2024). In such settings, increasing the number of queries is often sufficient to recover high-quality student models. However, this assumption breaks down in edge-oriented scenarios: each query requires costly on-device inference from a quantized teacher, and naive increases in query quantity not only become infeasible but may also degrade performance due to discretization noise.

In this work, we argue that *query construction is the central bottleneck* for model extraction and distillation under edge and quantization constraints. When both the teacher and the student are heavily quantized and the query budget is strictly limited, randomly sampled or heuristically designed instruction queries are often semantically redundant or overly complex. Such queries waste valuable query budgets, introduce noisy supervision, and destabi-

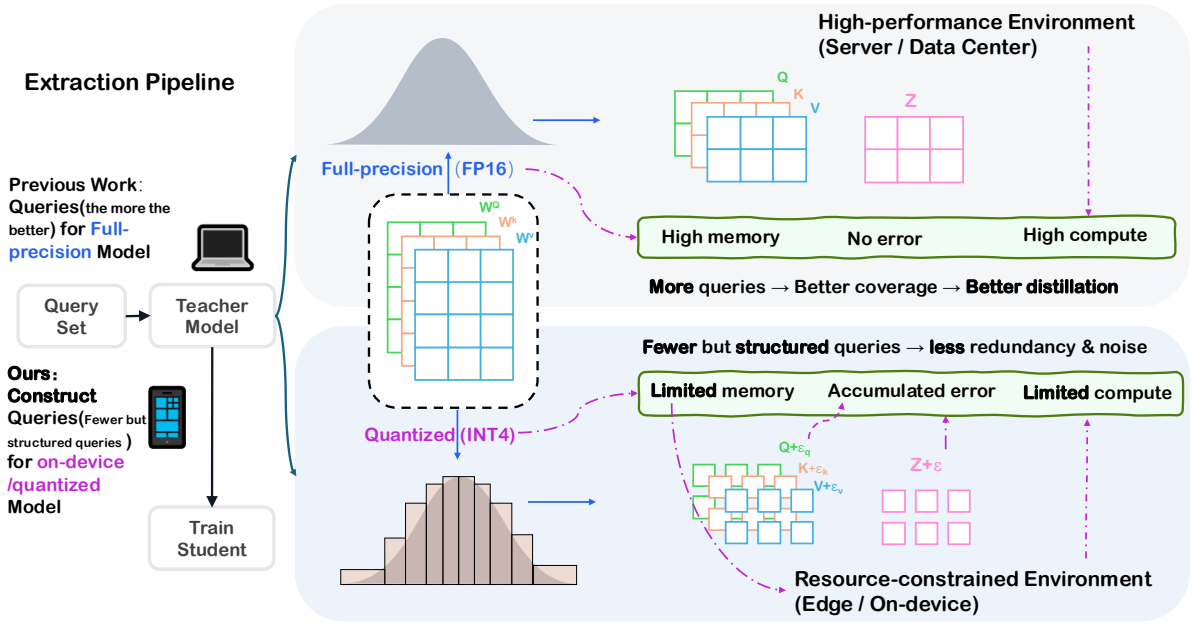

Figure 1: Motivation for cluster-aware query design under quantized, resource-constrained settings. In high-performance environments (server or data center), full-precision teachers can benefit from a large number of queries due to abundant memory and computation, leading to broader coverage and strong distillation signals. In contrast, under edge-oriented settings with quantized teachers (e.g., INT4), limited memory, accumulated quantization error, and restricted computation make distillation highly sensitive to redundant or poorly structured queries. This motivates the use of fewer but carefully designed, semantically representative queries to reduce noise and improve the effectiveness of student training under strict resource and query budgets.

lize quantized teachers (Kim et al., 2024; Yue et al., 2024; Liu et al., 2024). As a result, increasing the number of queries does not necessarily translate into better student performance.

Our key insight is that, in edge-oriented and quantization-constrained settings, the *structure* of the query pool fundamentally determines the learning dynamics of low-capacity student models. Rather than proposing new clustering algorithms or instruction generation techniques (Tang et al., 2019; Zhou and Chiam, 2023), we focus on how queries are *organized and allocated*. By explicitly modeling the semantic organization of instruction queries and enforcing balanced coverage across semantic clusters, structured query construction acts as an implicit regularizer that mitigates redundancy, reduces supervision noise, and better aligns with the limited representational capacity of quantized models.

Based on this perspective, we propose **CLIQ** (**Cl**ustered **I**nstruction **Q**uerying), a query-efficient distillation framework tailored for edge-oriented LLM extraction. CLIQ clusters candidate instructions in a semantic embedding space and constructs a compact set of cluster-aware, representative queries that maximize semantic coverage per query while avoiding redundancy. This design enables more robust and efficient supervision from quantized edge teachers and is fully compatible with common quantization techniques such as QLoRA (Dettmers et al., 2023).

Empirically, we demonstrate that under identical query budgets, CLIQ consistently outperforms random and heuristic query sampling strategies across multiple student architectures, quantization levels, and evaluation metrics. In our experiments, we restrict the query budget to 100–1000 instructions and keep both teacher and student models heavily quantized (INT8/INT4 GPTQ) (Frantar et al., 2022), faithfully reflecting the constraints of real-world edge deployment—even when evaluated on server hardware.

In summary, this work makes the following contributions:

(I) We formulate an edge-oriented model extraction problem in which the teacher is a quantized on-device LLM and query inference is strictly budgeted.

(II) We identify query construction, rather than

query quantity, as the dominant bottleneck for effective distillation under edge and quantization constraints.

(III) We propose **CLIQ**, a clustered instruction querying framework that constructs compact yet informative query sets via semantic clustering and cluster-aware generation.

(IV) We empirically show that **CLIQ** enables more efficient and robust distillation than random or heuristic query sampling methods, particularly under limited query budgets.

2 Related Work

2.1 Knowledge Distillation for Language Models

Knowledge distillation has been widely studied as a means of transferring knowledge from large teacher models to smaller student models. Early work focused on distillation for classification (Hinton et al., 2015) and sequence modeling (Kim and Rush, 2016), while more recent studies explore instruction-level (Honovich et al., 2023) and response-level distillation (Gu et al., 2024) for large language models. These approaches typically rely on large collections of queries or instructions to supervise the student, implicitly assuming that teacher inference is inexpensive and that increasing the number of queries reliably improves supervision quality.Recent studies also explore data-free or API-based distillation settings, where student models are trained using black-box teacher access without ground-truth annotations. However, these approaches generally assume either abundant query access or high-fidelity teacher responses, which may not hold in edge deployment scenarios where both computation and model precision are constrained.

2.2 Data-efficient and Query-efficient Distillation

Several recent works investigate data-efficient or query-efficient distillation strategies, including uncertainty-based sampling (Du et al., 2025) and synthetic data generation (Honovich et al., 2023; Wang et al., 2023). These methods aim to reduce annotation or inference costs by selecting informative queries or generating compact supervision sets.

However, existing query-efficient and active distillation methods typically share a common underlying assumption: the teacher model is high-precision, relatively noise-free, and capable of providing stable and informative responses across a wide range of queries. Under this assumption, query efficiency primarily concerns *which* queries to select, while the structure of the query pool plays a secondary role.

2.3 Edge-oriented and Quantized Language Models

Deploying language models on edge devices has motivated extensive research on model compression, quantization, and efficient fine-tuning techniques such as GPTQ and QLoRA (Frantar et al., 2022; Dettmers et al., 2023). These methods significantly reduce memory and computation requirements, but often introduce expressivity loss and increased sensitivity to noisy supervision.

3 Method

3.1 Overview

We study the problem of *query-efficient knowledge distillation from quantized large language models deployed on edge devices*. Edge-side models, such as on-device LLMs used in mobile or embedded systems, operate under *strict computational and memory constraints*, which often require aggressive weight quantization and reduced model capacity. While quantization significantly lowers storage and inference costs, it often reduces effective model expressivity, which can make naive query-based distillation inefficient and unstable (Dettmers et al., 2023; Frantar et al., 2022)

To address these challenges, we propose **CLIQ** (**Cl**ustered **I**nstruction **Q**uerying), a structured query construction framework for extracting high-quality supervision from quantized edge teachers under strict query budgets.As illustrated in Figure 2, CLIQ decomposes edge-oriented knowledge extraction into three stages: *(i) offline query analysis and construction*, *(ii) budgeted on-device querying of a quantized teacher*, and *(iii) student distillation for edge deployment*. Importantly, the expensive on-device inference is only performed on a compact, carefully constructed query set, while all computationally intensive query analysis and generation are conducted offline using a strong model.Specifically, CLIQ is designed to minimize the number of inference queries issued to the edge-side teacher, select semantically representative and diverse instruction queries through clustering, and

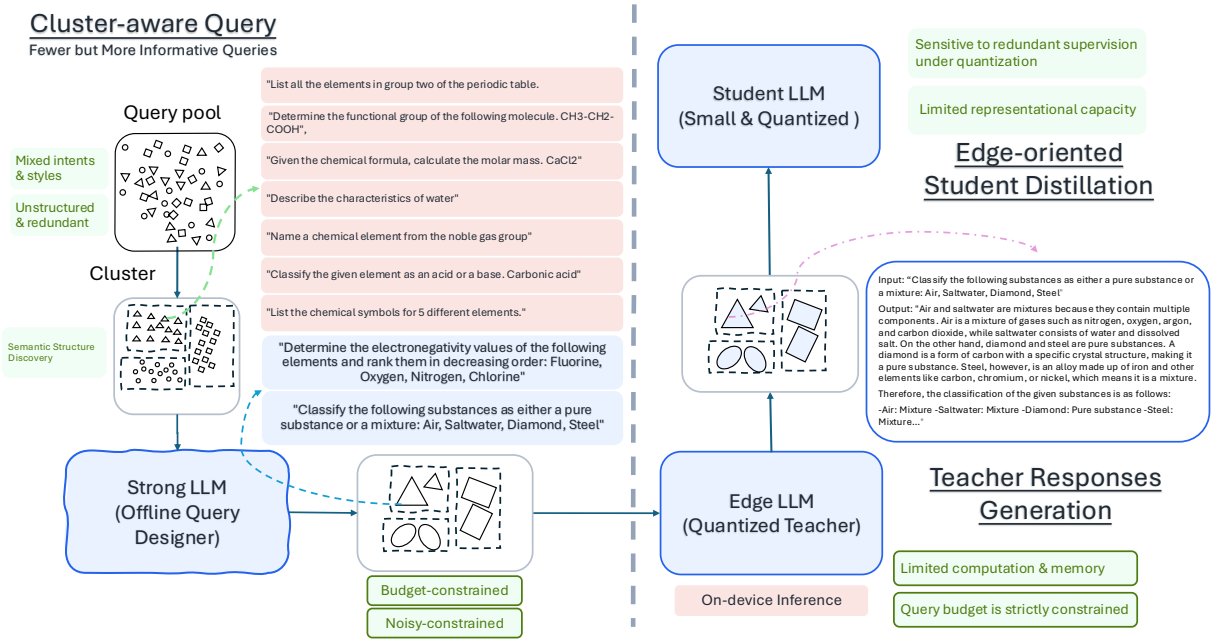

Figure 2: Overview of the CLIQ pipeline for cluster-aware query design and distillation under edge-oriented constraints. Given a large, unstructured query pool with mixed intents and styles, CLIQ first discovers latent semantic structure via clustering and selects a small set of representative, cluster-conditioned queries. These queries are designed offline by a strong LLM and issued to a quantized teacher model under strict query and computation budgets. The resulting teacher responses provide compact yet informative supervision for training a small, quantized student model. By explicitly controlling semantic coverage and redundancy at the query level, CLIQ mitigates the sensitivity of quantized models to noisy or repetitive supervision and enables effective on-device distillation with limited resources.

improve robustness to quantization-induced noise by avoiding redundant or overly complex instructions.This design explicitly shifts the focus from *query quantity* to *query structure*, which is critical for effective distillation in edge-oriented settings.

3.2 Problem Formulation

Let T_θ denote a quantized teacher model deployed on an edge device, parameterized by θ. Given a task distribution $\mathcal{D}$, our goal is to train a compact student model S_ϕ such that it approximates the behavior of T_θ under a limited query budget.

Formally, the student is trained by minimizing:

$$\min_{\phi} \mathbb{E}_{x \in \mathcal{Q}} \left[\mathcal{L}(S_\phi(x), T_\theta(x)) \right], \tag{1}$$

where $\mathcal{Q}$ denotes the set of instruction queries sent to the teacher, and $\mathcal{L}$ is a task-dependent loss function (e.g., cross-entropy or mean squared error).

CLIQ focuses on optimizing the construction of $\mathcal{Q}$ under these constraints. The construction of $\mathcal{Q}$ is subject to two constraints: (i) **Query budget:** $|\mathcal{Q}|$ must be small due to limited inference capacity on edge hardware.(ii) **Information:** Queries should maximize semantic coverage of the teacher's knowledge while minimizing redundancy.

3.3 Quantization-Induced Noise and Sensitivity

Quantization reduces the effective expressivity of large language models and introduces non-negligible quantization-induced errors. Prior studies have shown that quantization inevitably produces approximation errors, which can degrade model performance in downstream tasks (Nagel et al., 2021; Li et al., 2026). In the context of large language models, aggressive low-bit quantization has been observed to cause significant performance degradation on complex reasoning benchmarks (Li et al., 2025). Following this understanding, we model the behavior of a quantized teacher as:

$$T_\theta^{\text{quant}}(x) = T_\theta^{\text{full}}(x) + \epsilon(x), \tag{2}$$

where $\epsilon(x)$ denotes the quantization-induced perturbation on model outputs.

Empirically, we observe that highly quantized models (e.g., 4-bit or 8-bit) exhibit increased output variance even for semantically similar queries.

While approaches such as QLoRA (Dettmers et al., 2023) can partially mitigate expressivity loss, quantized models remain sensitive to noisy or overly complex queries.

These observations motivate a key design principle of CLIQ:

> *Under a limited query budget, particularly for quantized edge-deployed teachers, prioritizing semantic representativeness over query quantity leads to more effective knowledge extraction than issuing a large number of unstructured queries.*

3.4 Query Construction via Clustered Instruction Querying

CLIQ adopts a **two-stage query construction strategy**: (i) semantic clustering of candidate queries, followed by (ii) cluster-aware high-quality query generation.

3.5 Semantic Query Clustering

Existing distillation approaches often treat instruction queries independently, ignoring latent semantic redundancy. In practice, many user instructions are highly similar, and naively sampling them leads to redundant teacher inferences and inefficient supervision.

Given an initial query pool $\mathcal{Q}_0 = \{q_1, \ldots, q_N\}$, we first encode each instruction query into a semantic embedding space using a sentence-level encoder, such as Sentence-BERT (Reimers and Gurevych, 2019) or LLM-based embeddings. We then apply unsupervised clustering to partition the query set into K clusters:

$$\mathcal{Q} = \bigcup_{k=1}^{K} \mathcal{Q}_k, \tag{3}$$

where each cluster $\mathcal{Q}_k$ groups semantically related instructions. Semantic clustering has been widely adopted to reduce redundancy and promote diversity in large-scale instruction collections (Honovich et al., 2023).

Semantic clustering enables CLIQ to reduce redundancy among instruction queries,ensure diverse semantic coverage,provide a structured basis for controlled query generation.

3.6 Cluster-aware Instruction Query Generation

After clustering, CLIQ generates a compact set of high-quality instruction queries from each cluster. For cluster $\mathcal{Q}_k$, we construct a cluster prototype (e.g., via centroid embeddings or representative queries) and design a cluster-aware prompt to guide query generation:

$$\tilde{\mathcal{Q}}_k = \mathrm{LLM}(f(\mathcal{Q}_k)).$$

The cluster-aware prompt is designed to generate queries that are semantically representative of each cluster, sufficiently diverse to reduce redundancy, and constrained in complexity to better match the capacity of low-precision, edge-oriented student models.

The final query set is given by $\mathcal{Q} = \bigcup_{k=1}^{K} \tilde{\mathcal{Q}}_k$, which is compact yet information-rich. Unlike generic instruction generation, CLIQ constrains reasoning depth and linguistic complexity, improving robustness under quantization-induced noise.

3.7 Student Training

Given the constructed query set $\mathcal{Q}$ and corresponding teacher outputs $\{T_\theta(x)\}$, the student model is trained by minimizing the distillation loss:

$$\mathcal{L}_{\mathrm{KD}} = \frac{1}{|\mathcal{Q}|} \sum_{x \in \mathcal{Q}} \ell(S_\phi(x), T_\theta(x)). \tag{4}$$

We employ standard distillation techniques, including temperature scaling and output smoothing, and use a small, quantized student architecture suitable for edge deployment.

3.8 Design Guideline for Cluster-aware Query Construction

CLIQ is motivated by a fundamental coverage–redundancy trade-off in query-efficient distillation. Under a fixed query budget, issuing queries independently from an imbalanced instruction distribution tends to over-sample dense semantic regions while under-covering sparse ones, leading to redundant supervision and early performance saturation, especially for quantized edge-oriented models.

From both theoretical intuition and empirical evidence, our framework suggests a simple and practical design guideline: *for a fixed query budget, effective distillation is achieved by allocating queries across a moderate number of semantic clusters while ensuring sufficient per-cluster diversity, rather than aggressively increasing either clustering granularity or total query count.*

This guideline directly informs the design choices in CLIQ. Moderate clustering granularity improves global semantic coverage, while a small

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 (%).

but sufficient number of queries per cluster captures dominant intra-cluster variations without introducing unnecessary redundancy. As demonstrated in Section 4 and Appendix Sections A.2.2 and A.2.3, this balance yields robust and sample-efficient distillation under aggressive quantization and strict edge deployment constraints.

3.9 Advantages of CLIQ

CLIQ provides several advantages for distilling edge-oriented and quantized language models. First, by explicitly modeling the semantic structure of the query space, CLIQ substantially reduces the number of teacher queries required to achieve strong distillation performance, making it highly query-efficient under strict inference budgets. Second, the cluster-aware construction process mitigates the impact of noisy or redundant supervision signals, which is particularly important for quantized models that are sensitive to supervision noise and precision loss. In addition, CLIQ is inherently compatible with edge deployment scenarios, as it enables effective knowledge transfer to compact student models with limited capacity and aggressive quantization. Finally, the proposed framework is model-agnostic and can be readily combined with quantized teachers or parameter-efficient adaptation techniques such as QLoRA, without relying on model-specific assumptions.

4 Experiments

4.1 Experimental Setup

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

Unless otherwise specified, both teacher inference and student training are performed using quantized weights, ensuring that our evaluation faithfully reflects edge deployment constraints.And our main comparison uses a fixed query budget of 1000 queries. We additionally report budget ablations (100–400) and training-step analyses under a fixed 500-query budget.

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

- **Original Queries (OQ):** Instruction queries sampled from the original dataset.
- **CLIQ (Ours):** Cluster-aware generated instruction queries constructed via semantic clustering and cluster-aware prompting.

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

For CLIQ, queries are generated using the Qwen API with explicit constraints on instruction length and reasoning complexity, ensuring compatibility with low-capacity, quantized student models.

Evaluation Metrics. We evaluate instruction-following performance using standard automatic metrics, including BERT-F1 (Zhang et al., 2020), BLEU (Papineni et al., 2002), ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-Lsum (Lin, 2004). All reported results are averaged over held-out validation or test sets.

4.2 Teacher Quantization Sanity Check

Before evaluating distillation performance, we first examine the impact of quantization on teacher models. Table 1 compares full-precision and GPTQ-quantized versions of Qwen2.5-7B and Qwen3-1.7B (Yang et al., 2025).

Across all evaluation metrics, GPTQ-based quantization introduces only marginal performance degradation. For example, Qwen2.5-7B-Instruct-GPTQ-Int4 achieves comparable BERT-F1 and ROUGE scores to its full-precision counterpart. These results confirm that quantized teacher models remain sufficiently expressive for both query generation and supervision, and justify the use of quantized teachers in resource-constrained settings.

Method	BERT-F1	BLEU	RLsum
Original Queries	77.97	1.05	13.37
CLIQ (Ours)	**84.35**	**2.77**	**17.50**

Table 2: Main distillation results on INT8-quantized 1.7B student models.All scores are reported in percentage (%).

Model	Query Type	BERT-F1 / ROUGE-L
4B, INT4	OQ	77.9 / 10.9
4B, INT4	**CLIQ(ours)**	**82.4 / 12.5**
7B, INT4	OQ	79.7 / 11.7
7B, INT4	**CLIQ(ours)**	**83.0 / 12.6**
1.7B, INT8	OQ	83.5 / 13.1
1.7B, INT8	**CLIQ(ours)**	**83.9 / 12.4**

Note: All results are obtained under identical query budgets (500 queries).OQ denotes original queries. CLIQ denotes our cluster-aware generated queries. Full results across different training steps are reported in Appendix. All scores are reported in percentage (%).

Table 3: Effect of Query Construction under Identical Query Budgets.

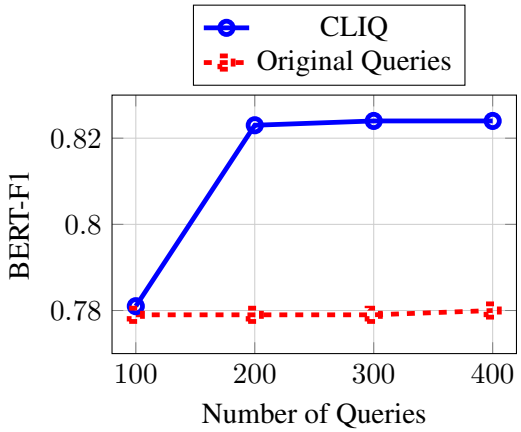

Figure 3: Effect of query budget on distillation performance. CLIQ exhibits rapid gains as the number of queries increases from 100 to 300, followed by diminishing returns, while Original Queries show minimal improvement.

4.3 Main Results

Table 2 presents the main distillation results on quantized student models. Under identical query budgets, **CLIQ consistently outperforms Original Queries** across all evaluation metrics.

Notably, a 1.7B student model with INT8 quantization distilled using CLIQ achieves a **BERT-F1 score of 0.8435**, matching or even surpassing the performance of significantly larger teacher models (e.g., 7B). This result demonstrates that **carefully constructed, semantically representative queries** can effectively compensate for limited model capacity and aggressive quantization.

Improvements are consistently observed across BLEU and ROUGE metrics, indicating that the gains stem from improved instruction-following quality rather than metric-specific artifacts.

4.4 Comparison with Original Query Baseline

To isolate the effect of query construction, we further compare CLIQ with Original Queries under identical query budgets and training protocols. As shown in Table 3, students trained with CLIQ consistently outperform those trained with Original Queries.

The performance gap is particularly pronounced on ROUGE-1, ROUGE-2, and ROUGE-Lsum, suggesting that structured query construction yields more informative supervision signals. Although both settings employ random query sampling during training, CLIQ benefits from a more structured and semantically representative query pool.

4.5 Query Budget and Sample Efficiency

We analyze the effect of query budget by varying the number of generated queries from 100 to 400. As shown in Figure 3, student performance under CLIQ improves rapidly as the query budget increases from 100 to 300, after which the gains gradually saturate.

In contrast, models trained with Original Queries exhibit slower improvement and reach a performance plateau earlier. This trend is consistent across different teacher–student pairs and evaluation metrics, highlighting the **superior sample efficiency** of CLIQ under strict query budgets.

4.6 Effect of Quantization

We further investigate how different query construction strategies behave under varying quantization levels using a fixed 1.8B edge-oriented student model. Table 4 reports results under both INT8 and INT4 quantization.

Across both quantization settings, cluster-aware generated queries consistently outperform Original Queries by a large margin. Under INT4 quantization, CLIQ achieves a BERT-F1 score of 0.829, compared to 0.800 obtained by Original Queries.

More importantly, CLIQ demonstrates strong robustness to aggressive quantization. While Orig-

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.

inal Queries exhibit early saturation and limited sensitivity to increased training, the performance of CLIQ remains stable when moving from INT8 to INT4. This suggests that structured and semantically representative queries can effectively mitigate the noise amplification introduced by low-precision weights.

These results highlight that, for heavily quantized edge models, the quality and structure of supervision signals play a more critical role than the sheer number of queries or training iterations.

5 Analysis and Discussion

In this section, we analyze why cluster-aware query construction is particularly effective for distilling quantized, edge-oriented language models, and we discuss the limitations of our approach.

5.1 Why Does Query Clustering Help Quantized Students?

Quantized student models suffer from a significantly reduced effective capacity due to low-precision weights (INT8 / INT4) and smaller parameter counts. As a result, such models are more sensitive to noisy, redundant, or poorly structured supervision signals.

Query clustering explicitly models the latent structure of the query space by grouping semantically similar queries. This allows our method to reduce redundancy in the query pool and ensures that each generated query corresponds to a distinct semantic region. For quantized students, such structured supervision is crucial: instead of spreading their limited capacity across highly correlated instruction patterns, the model can focus on learning representative instruction behaviors. It is important to note that in all settings, training queries are randomly sampled; the observed differences therefore arise solely from the structure of the query pool. This effect is particularly evident under INT4 quantization, where we observe larger performance gaps between cluster-aware generated queries and original queries. These results suggest that clustering acts as an implicit regularizer by reducing redundancy and constraining the effective supervision space, enabling quantized models to allocate their limited representational capacity more efficiently.

5.2 Effect of Cluster-aware Query Generation

Beyond clustering, the cluster-aware query generation process plays a critical role in improving distillation quality. Unlike generic instruction generation, our prompts are explicitly designed to be aware of cluster-level semantics and to constrain query complexity, length, and reasoning depth.

Such constraints are especially important for edge-oriented models, which often struggle with long-chain reasoning or overly complex instructions. By generating queries that are both semantically representative and appropriately scoped, our method produces supervision signals that are better aligned with the inductive biases of small, quantized transformer models.

Empirically, this design choice is reflected in consistent improvements in ROUGE-based metrics, indicating better coverage of salient content rather than superficial fluency gains.

5.3 Sample Efficiency and Performance Saturation

Our experiments show that student performance improves rapidly as the number of cluster-aware generated queries increases from 100 to approximately 300, after which the gains gradually saturate. This trend suggests that cluster-aware query generation is highly sample-efficient: a relatively small number of carefully constructed queries is sufficient to cover most of the informative instruction patterns.In contrast, models trained on original query pools exhibit slower improvement and reach a performance plateau earlier.This indicates that increasing the number of queries alone is insufficient; the quality and structure of queries are critical factors for effective distillation under strict query budgets.

6 Limitations

Despite its effectiveness, our method has several limitations. First, the quality of query clustering depends on the underlying embedding model; poor embeddings may lead to suboptimal cluster assignments and reduced query diversity. Second, while

cluster-aware prompts constrain query complexity to better match the capabilities of edge-oriented models, they may underrepresent tasks that require long-chain reasoning or complex multi-step instructions. Finally, our approach assumes access to a reasonably consistent and capable teacher model via API. In scenarios where teacher responses are highly noisy or unstable, the quality of generated queries may degrade. We leave the exploration of adaptive clustering strategies and robustness to noisy teachers as promising directions for future work.

7 Ethical Considerations

This work studies query-efficient knowledge distillation for edge-oriented and heavily quantized language models. While our approach is motivated by practical deployment constraints, it is important to discuss potential ethical, legal, and societal implications.

7.1 Model Extraction and Misuse Risks

Our method is related to prior work on model distillation and behavioral transfer, which may raise concerns about unauthorized model extraction. We emphasize that CLIQ does *not* attempt to recover model parameters, internal representations, or exact output distributions of the teacher model. Instead, it focuses on improving task-level instruction-following behavior for significantly smaller and lower-capacity student models under strict query budgets.

All experiments assume standard, rate-limited API access and do not exploit vulnerabilities, bypass safeguards, or violate technical protections. The resulting student models remain substantially less capable than the teacher models, and therefore do not constitute a substitute for the original systems.

7.2 Legal and Compliance Considerations

Our experiments rely on publicly available instruction datasets and on teacher models accessed through research-permitted APIs. We do not claim universal legal permissibility across all providers or deployment contexts. Rather, our work is intended for research and development scenarios that comply with applicable licenses, terms of service, and local regulations.

We caution against applying the proposed method in ways that violate model usage agreements or applicable laws. We believe that clearly characterizing query-efficient distillation behavior can also help model providers better understand and mitigate potential misuse.

7.3 Societal Impact and Intended Use

The primary motivation of this work is to support efficient deployment of language models on edge devices, where constraints on memory, computation, and energy consumption are critical. Such settings include privacy-sensitive applications, offline or low-connectivity environments, and resource-constrained platforms.

By improving the sample efficiency and robustness of distillation for quantized student models, our approach may reduce the environmental and infrastructural costs associated with large-scale model deployment. At the same time, we acknowledge that any technique for behavioral transfer may be misused if applied irresponsibly. We therefore encourage responsible use aligned with ethical norms and legal requirements.

Overall, we believe that the benefits of enabling efficient, privacy-aware, and accessible on-device language models outweigh the potential risks, provided that the method is applied in a responsible and compliant manner.

References

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.

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.

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.

Taehyeon Kim, Joonkee Kim, Gihun Lee, and Se-Young Yun. 2024. Instructive decoding: Instruction-tuned large language models are self-refiner from noisy instructions. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

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.

Dongmin Li, Xiurui Xie, Dongyang Zhang, Athanasios V. Vasilakos, and Man-Fai Leung. 2026. Semq: Efficient non-uniform quantization with sensitivity-based error minimization for large language models. *Future Generation Computer Systems*, 175:108120.

Xiang Li, Zhenyan Lu, Dongqi Cai, Xiao Ma, and Mengwei Xu. 2024. Large language models on mobile devices: Measurements, analysis, and insights. In *Proceedings of the Workshop on Edge and Mobile Foundation Models*, EdgeFM '24, page 1–6, New York, NY, USA. Association for Computing Machinery.

Zhen Li, Yupeng Su, Runming Yang, and 1 others. 2025. Quantization meets reasoning: Exploring llm low-bit quantization degradation for mathematical reasoning. *arXiv preprint arXiv:2501.03035*.

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.

Xiaoyu Liu, Yun Zhang, Wei Li, Simiao Li, Xudong Huang, Hanting Chen, Yehui Tang, Jie Hu, Zhiwei Xiong, and Yunhe Wang. 2024. Multi-granularity semantic revision for large language model distillation. *CoRR*, abs/2407.10068.

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

Michele Miranda, Elena Sofia Ruzzetti, Andrea Santilli, Fabio Massimo Zanzotto, Sébastien Bratières, and Emanuele Rodolà. 2025. Preserving privacy in large language models: A survey on current threats and solutions. *Transactions on Machine Learning Research*.

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.

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.

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.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3980–3990. Association for Computational Linguistics.

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.

Raphael Tang, Yao Lu, Linqing Liu, Lili Mou, Olga Vechtomova, and Jimmy Lin. 2019. Distilling task-specific knowledge from BERT into simple neural networks. *CoRR*, abs/1903.12136.

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.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Yuanhao Yue, Chengyu Wang, Jun Huang, and Peng Wang. 2024. Distilling instruction-following abilities of large language models with task-aware curriculum planning. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 6030–6054. Association for Computational Linguistics.

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).

Tianxun Zhou and Keng-Hwee Chiam. 2023. Synthetic data generation method for data-free knowledge distillation in regression neural networks. *Expert Syst. Appl.*, 227:120327.

A Appendix

A.1 CLIQ Pipeline Details

A.1.1 Query Representation and Clustering

Query Pool Construction. Given a raw instruction dataset in JSON format, we construct the 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 a 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.

Clustering. All embeddings are clustered using MiniBatchKMeans with $K = 100$ clusters and a fixed random seed (42). MiniBatchKMeans (Sculley, 2010) is adopted for scalability and memory efficiency.

Filtering Small Clusters. Clusters whose size falls below a threshold (default: 5) are removed to avoid unstable prototypes and overly sparse supervision. Remaining clusters are re-indexed to ensure contiguous cluster IDs.

A.1.2 Cluster-aware Query Generation

Cluster-conditioned Prompting. For each cluster, we construct a prompt containing up to M in-cluster examples (default: $M = 1000$) to characterize the cluster semantics and instruction style. In our main experiments, we generate $m = 10$ queries per cluster unless otherwise stated, while for visualization we report a more query-efficient setting with $m = 5$.

Teacher API and Decoding. We use a Qwen-family teacher accessed via an OpenAI-compatible API endpoint. The decoding temperature is set to 0.7 with a large token budget (max tokens: 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
Clustering algorithm	MiniBatchKMeans
clusters K	100
Seed	42
MiniBatch size	$\min(1000, \lvert\mathcal{Q}\rvert/10)$
Cluster 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 cluster M	1000
Generated queries/cluster m	10

Table 5: Hyperparameters for clustering and cluster-aware query generation.

Algorithm 1: CLIQ Pipeline

Input: Query pool $\mathcal{Q}$; clusters K; min size s; examples M; queries m
Output: Generated query set $\widetilde{\mathcal{Q}}$
Encode each query $q \in \mathcal{Q}$ into $\mathbf{e}(q)$;
Cluster embeddings into K clusters via MiniBatchKMeans;
Remove clusters with size $< s$ and re-index;
$\widetilde{\mathcal{Q}} \leftarrow \emptyset$;
foreach *cluster* c **do**
  Sample up to M in-cluster queries;
  Construct cluster-conditioned prompt;
  Query teacher to generate m new 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 clustering and query generation. All settings are fixed across experiments unless otherwise stated.

A.2 Additional Quantitative Results

A.2.1 Full Model Evaluation

Table 6 reports the full evaluation results under a unified teacher–student setup, where **Qwen2.5-7B** serves as the teacher and **Qwen3-1.7B** is the student architecture. We report (i) the teacher under FP16 and Int4, (ii) base (unaligned) students under FP16 and Int8, and (iii) aligned students trained

with either original queries (OQ) or cluster-aware generated queries (CLIQ), including validation results ("valid").

Teacher vs. Base Student. The base student is substantially weaker than the teacher across all metrics (e.g., BERT-F1 and ROUGE family), which motivates alignment/distillation. Quantization of the base student (FP16 vs. INT8) changes performance only slightly, indicating that the base capability gap is primarily due to model size rather than precision.

Aligned Students with OQ vs. CLIQ. Alignment using queries consistently improves the student over its base version. Comparing query strategies, CLIQ achieves the strongest overall results among student settings, improving BERT-F1, BLEU, and ROUGE metrics over OQ. This pattern supports the central claim that *query quality and semantic coverage*, rather than merely the existence of alignment, is a key driver of the observed gains.

Validation Consistency. For both OQ and CLIQ, "valid" results are close to the corresponding test results, suggesting that the improvements are not due to overfitting to the test set but reflect stable generalization.

A.2.2 Training Dynamics Analysis

Fixed Query Budget. Figure 4 analyzes the training dynamics under a fixed query budget of 500 queries across three edge-oriented student models, reporting BERT-F1, BLEU, and ROUGE-L, with the 0-step base model included as initialization.

Across all models and metrics, CLIQ exhibits clear and consistent advantages over original queries (OQ). For BERT-F1, CLIQ leads to rapid performance gains within the first 200 training steps for all three students, followed by saturation around 200–300 steps, indicating diminishing returns once the structured supervision signal has been fully exploited. In contrast, OQ-trained models remain largely insensitive to additional training steps, showing near-flat trajectories across all models.

The same qualitative trend is observed for BLEU and ROUGE-L, where CLIQ yields substantial improvements over OQ, particularly for the smaller and more heavily quantized student models (A and B). Notably, under OQ supervision, BLEU and ROUGE-L remain almost unchanged throughout training, whereas CLIQ enables sustained improvements as optimization proceeds, highlighting its stronger and less redundant supervision signal.

Including the 0-step initialization further reveals that CLIQ consistently outperforms OQ beyond early training, rather than merely accelerating convergence. Together, these results demonstrate that under a fixed query budget, performance gains primarily arise from improved query quality and semantic coverage, rather than from increased optimization alone, and that CLIQ fundamentally alters the learning dynamics of quantized student models.

Different Quantization Settings. Table 7 further studies training steps under INT4 and INT8 quantization using the same 1.8B student and *randomly sampled* queries during training; the only difference is the construction of the query pool (Original vs. CLIQ). Two consistent patterns emerge across both quantization settings: (i) with Original queries, performance remains nearly flat from 100 to 500 steps, and can even slightly regress, and (ii) 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.2.3 Clustering Granularity and Query Allocation

Table 8 ablates two key design choices in CLIQ: the number of semantic clusters and the number of generated queries per cluster, evaluated on both Dolly and Alpaca.

Clustering Granularity. Across both datasets, moderate granularities (e.g., 25–50 clusters) tend to produce stable and competitive results. Overly fine-grained clustering (e.g., 100 clusters) exhibits higher variance and can lead to performance degradation under certain allocations, consistent with the intuition that excessive partitioning may fragment supervision into clusters that are too small or semantically narrow.

Queries per Cluster and Diminishing Returns. Increasing queries per cluster generally improves performance when moving from 4 to 8 or 10 queries, but the gains diminish and are not strictly monotonic across all settings. This indicates that CLIQ benefits primarily from improved semantic coverage and representative supervision rather than

Model	Role	Quant.	Query	BERT-F1	BLEU	R-1	R-2	R-L	R-Lsum
Qwen2.5-7B	Teacher	INT4	–	84.22	3.15	21.92	9.58	16.56	18.63
		FP16	–	83.69	2.94	22.08	9.36	16.85	18.76
Qwen3-1.7B	Base	INT8	–	81.40	1.73	12.23	4.36	8.70	10.03
		FP16	–	81.53	1.70	12.35	4.46	8.92	10.16
	Student	FP16→INT8	OQ	83.75	2.50	22.30	9.49	16.68	18.94
			OQ (valid)	83.78	2.46	22.24	9.57	16.83	18.96
		INT4→INT8	CLIQ	**84.35**	**3.04**	**23.42**	**10.60**	**17.60**	**19.87**
			CLIQ (valid)	84.23	2.92	23.07	10.26	17.27	19.53

Table 6: Comprehensive Evaluation Results under Different Model, Quantization, and Query Settings. All scores are reported in percentage (%).

Quantization	Query Pool	Steps	BERT-F1	BLEU	ROUGE-Lsum
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 7: Effect of Training Steps under Different Quantization Settings. All models use a 1.8B student with randomly sampled queries during training. Differences arise solely from the query pool construction. All scores are reported in percentage (%).

brute-force increases in query count.

Interaction Effects. We observe clear interactions between cluster count and per-cluster query allocation. Under finer clustering, insufficient queries per cluster can yield sparse and unstable supervision, while allocating more queries partially mitigates this issue. Overall, these results suggest that balanced configurations (moderate clusters with sufficient per-cluster queries) provide the best trade-off between semantic coverage and query efficiency.

Theoretical Intuition and Empirical Alignment. The observed trends in Table 8 and Figure 3 can be understood through a simple coverage–redundancy trade-off. Original query sampling corresponds to i.i.d. draws from a highly imbalanced mixture of latent semantic clusters, where dense clusters dominate early coverage while sparse clusters require substantially larger budgets to be reliably sampled. As a result, increasing the query budget under Original Queries often leads to redundant supervision and early performance saturation.

In contrast, CLIQ enforces a structured allocation by distributing a fixed number of queries across semantic clusters. Under a fixed budget $B = K \cdot m$, increasing the number of clusters K improves global semantic coverage, while the number of queries per cluster m controls intra-cluster diversity. This explains why moderate clustering granularities (e.g., $K = 25$ or 50) combined with a small but sufficient number of queries per cluster (e.g., $m = 8$ or 10) consistently achieve strong and stable performance in Table 8. When K is too large relative to the budget, insufficient per-cluster queries lead to sparse and unstable supervision, whereas increasing m beyond this range yields diminishing returns once dominant cluster modes are captured.

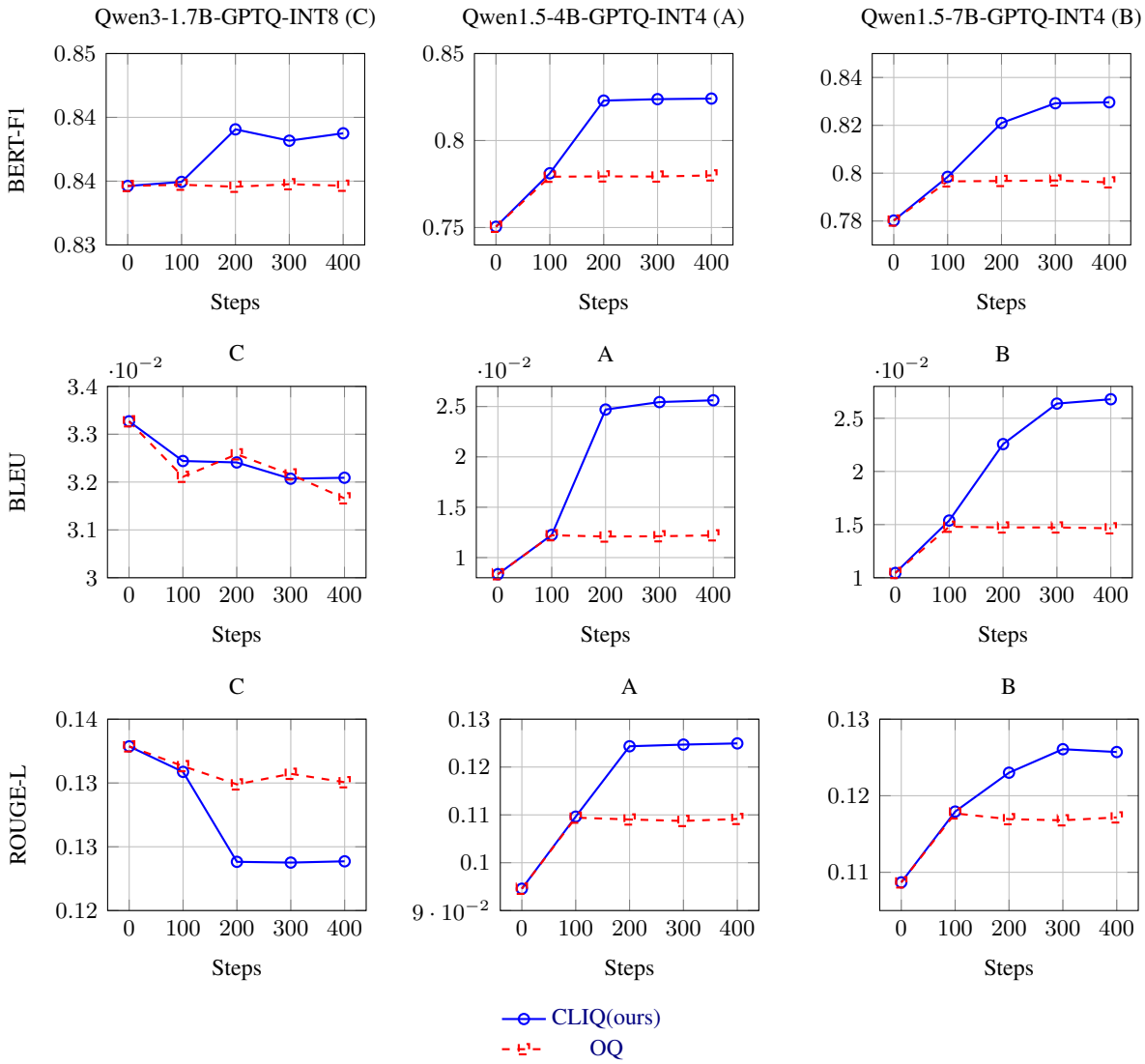

Figure 4: Effect of training steps under a fixed query budget (500 queries) across three student models. We report BERT-F1, BLEU, and ROUGE-L, and include the 0-step base model as initialization. CLIQ (ours) consistently outperforms OQ and exhibits diminishing returns after ∼200–300 steps.

A.3 Qualitative Analysis of Query Space

A.3.1 Cluster-level Coverage and Redundancy Analysis

To better understand the mechanisms behind the quantitative gains observed in Appendix A.2, we further analyze how original queries (OQ) and CLIQ-generated queries differ at the semantic cluster level.

Cluster Coverage under Fixed Query Budgets. Figure 6 reports the number of semantic clusters covered as a function of the query budget. While random sampling from the original query pool can eventually cover most clusters as the budget increases, CLIQ consistently achieves higher coverage under small to moderate budgets. This early-stage advantage provides more informative supervision signals during initial training, which helps explain the faster performance gains and earlier saturation observed for CLIQ-trained models (Section A.2.2).

Intra-cluster Redundancy. Figure 7 analyzes the average pairwise cosine similarity among queries within each cluster. Original queries exhibit substantially higher intra-cluster similarity, indicating strong redundancy in supervision. In contrast, CLIQ reduces redundancy by generating more diverse queries within each cluster. This observation explains why OQ-trained models are largely insensitive to additional training steps, whereas CLIQ continues to yield gains as optimization proceeds.

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 8: 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 (%).

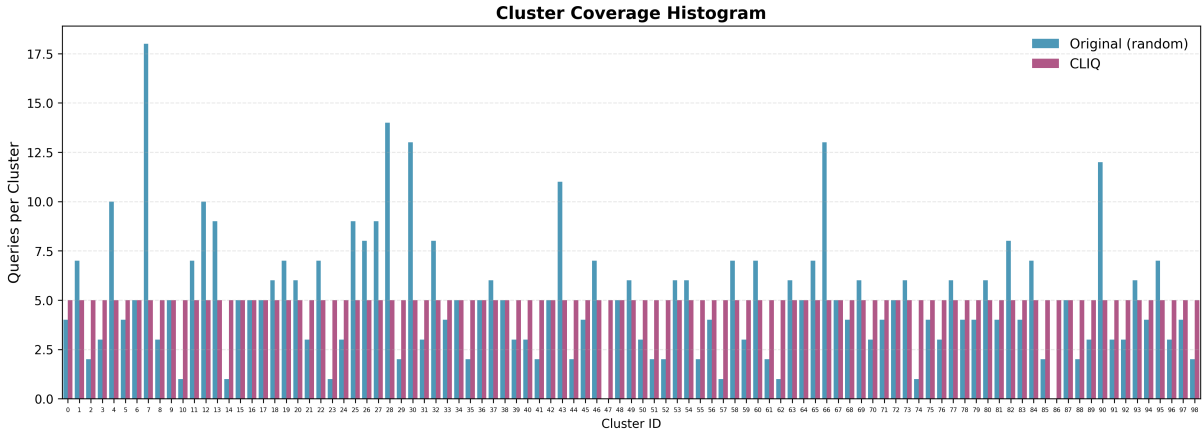

Figure 5: Histogram of query counts per cluster under a fixed query budget. Random sampling from original queries results in highly imbalanced cluster coverage, whereas CLIQ produces a near-uniform allocation across clusters, avoiding over-sampling of dense regions.

Distance to Cluster Centroids. Figure 8 shows the 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 suggests that CLIQ does not collapse queries to cluster prototypes, but instead preserves intra-cluster diversity while maintaining semantic coherence.

Cluster-level Query Allocation. Finally, Figure 5 visualizes the number of queries assigned to each cluster under a fixed query budget. Random sampling from original queries results in highly imbalanced cluster coverage, with certain dense clusters receiving disproportionate attention. In contrast, CLIQ produces a near-uniform allocation across clusters, avoiding over-sampling of dense regions and under-coverage of sparse ones.

Taken together, these analyses provide a unified explanation for the training dynamics and ablation results reported in Appendix A.2. CLIQ improves performance not by increasing the total number of queries or optimization steps, but by restructuring the query distribution to achieve balanced cluster-level coverage, reduced redundancy, and greater semantic diversity.

A.3.2 Global Query Coverage

Figure 9 visualizes the 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 embed-

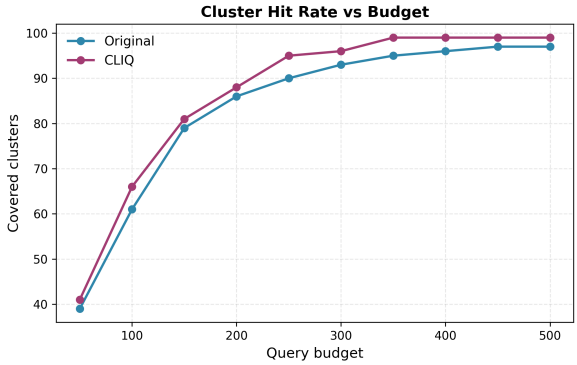

Figure 6: Cluster hit rate as a function of query budget. We measure the number of semantic clusters covered by at least one query under increasing query budgets. Compared with random sampling from the original query pool (OQ), CLIQ achieves substantially higher cluster coverage in the low-budget regime and reaches near-complete coverage with fewer queries. This indicates that CLIQ allocates queries more efficiently across semantic clusters, particularly when the query budget is limited.

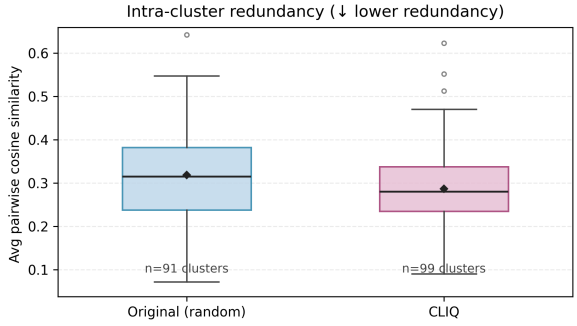

Figure 7: Intra-cluster redundancy measured by average pairwise cosine similarity within each semantic cluster. Original queries (OQ) exhibit significantly higher intra-cluster similarity, indicating strong redundancy and overlapping supervision signals. In contrast, CLIQ-generated queries reduce redundancy by promoting greater diversity within clusters, which explains the improved sensitivity of CLIQ-trained models to increased training steps and optimization.

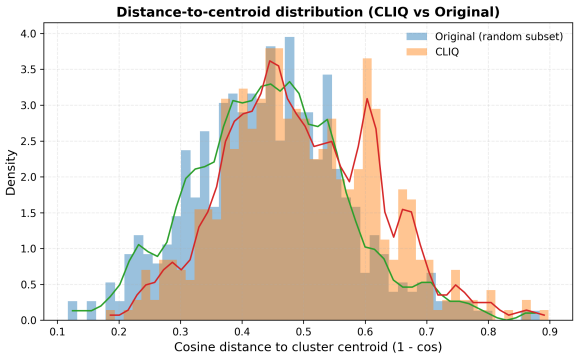

Figure 8: Distribution of cosine distances between queries and their corresponding cluster centroids. CLIQ-generated queries span a wider range of distances, covering both central and peripheral regions of each cluster. This indicates that CLIQ does not collapse queries to cluster prototypes, but preserves structured intra-cluster diversity while maintaining semantic coherence.

ding model for fair comparison.

Original Query Pool. The original query pool (OQ) forms highly concentrated and dense regions in the embedding space, with substantial overlap among queries. This indicates strong semantic redundancy, where many queries occupy similar regions and provide overlapping supervision signals. Such redundancy offers a qualitative explanation for the limited sensitivity of OQ-trained models to increased training steps and query budgets, as observed in the training-dynamics and ablation results (Section A.2.2).

CLIQ-generated Query Pools. In contrast, query pools constructed by CLIQ exhibit significantly more dispersed and uniform distributions. 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 is consistent with the quantitative gains reported in Table 8, where balanced CLIQ configurations achieve strong performance with relatively modest query budgets.

Effect of Clustering and Query Allocation. Varying the number of clusters and queries per cluster leads to clear and interpretable changes in the query space structure. Overly fine-grained clustering combined with insufficient queries per cluster produces sparse coverage and visible gaps in the embedding space. These gaps correspond to performance degradation in the quantitative ablation (Table 8), suggesting that excessive partitioning fragments supervision into clusters that are too small to

provide stable learning signals. Conversely, moderate clustering granularities with sufficient per-cluster queries yield more uniform coverage and stronger downstream performance.

Taken together, the UMAP visualization provides a qualitative explanation for the observed trade-offs between semantic coverage and query efficiency in CLIQ, and supports the interpretation that performance gains arise from structured query design rather than brute-force increases in query count.

A.3.3 Cluster-conditioned Projections

Figure 10 presents cluster-conditioned semantic projections comparing original queries (OQ) and CLIQ-generated queries (GQ). Each subfigure corresponds to one semantic cluster, where blue points denote original queries and orange points denote generated queries. Only $m = 5$ generated queries per cluster are shown, highlighting the extreme query-efficiency setting.

Despite the large imbalance between OQ and GQ, the generated queries consistently align with high-density semantic regions of their corresponding clusters. Rather than collapsing to global centroids, GQ samples often concentrate on cluster-internal modes, indicating that cluster-conditioned prompting captures salient semantic patterns specific to each cluster.

This observation complements the global UMAP analysis and the quantitative cluster–query ablation results. It provides qualitative evidence that CLIQ preserves semantic representativeness within clusters while substantially reducing the number of required queries, thereby explaining why CLIQ achieves strong performance under strict query budgets in the main experiments.

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 (e.g., ChatGPT) were employed to assist with language polishing, clarity improvement, and structural refinement of the text, as well as 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 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.

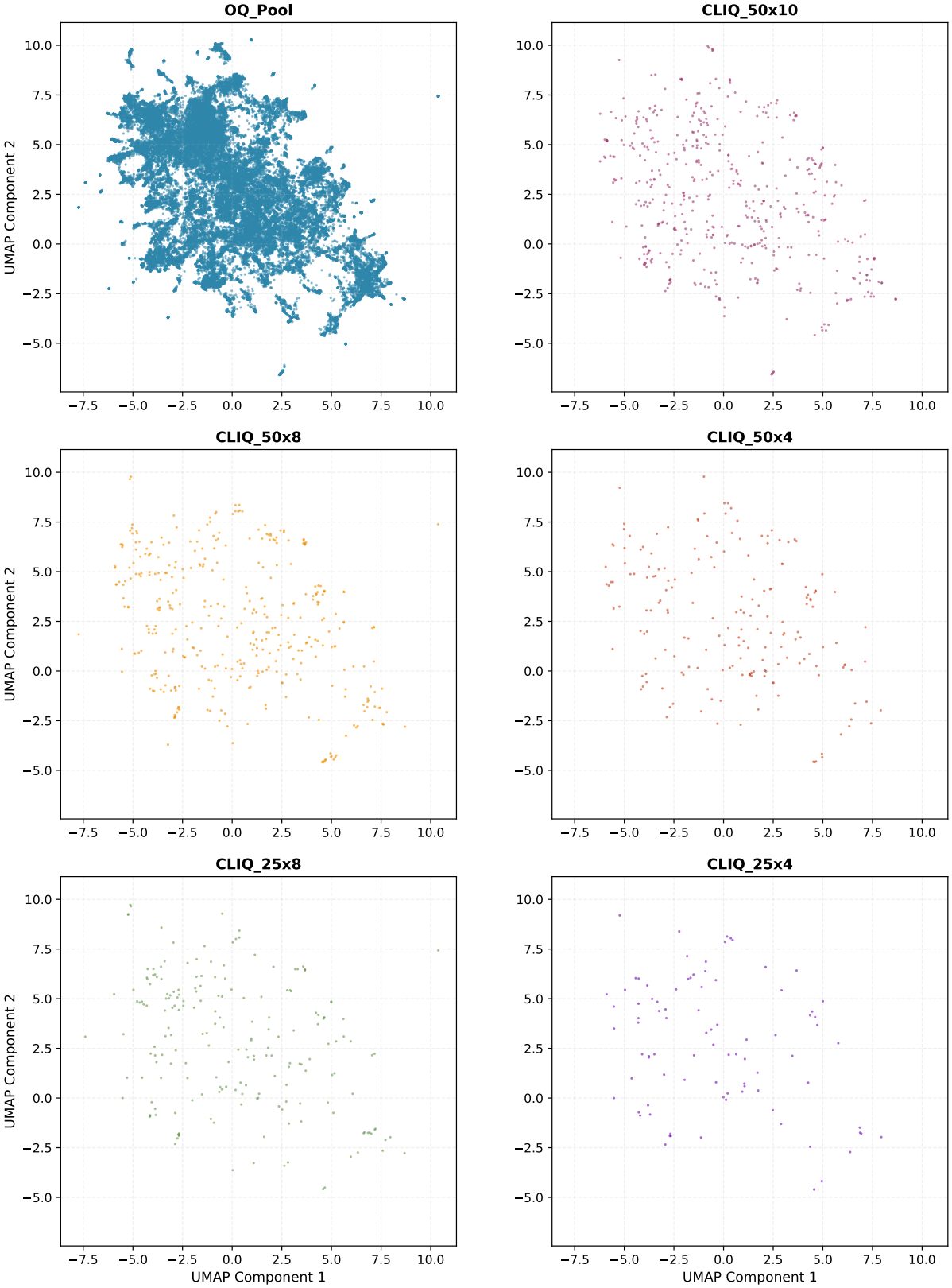

Figure 9: UMAP visualization of query coverage under different CLIQ configurations. The original query pool (OQ) exhibits dense and redundant regions, while CLIQ-generated queries maintain broader and more uniform semantic coverage under fixed query budgets. Different combinations of cluster numbers and queries per cluster illustrate how CLIQ controls coverage granularity.

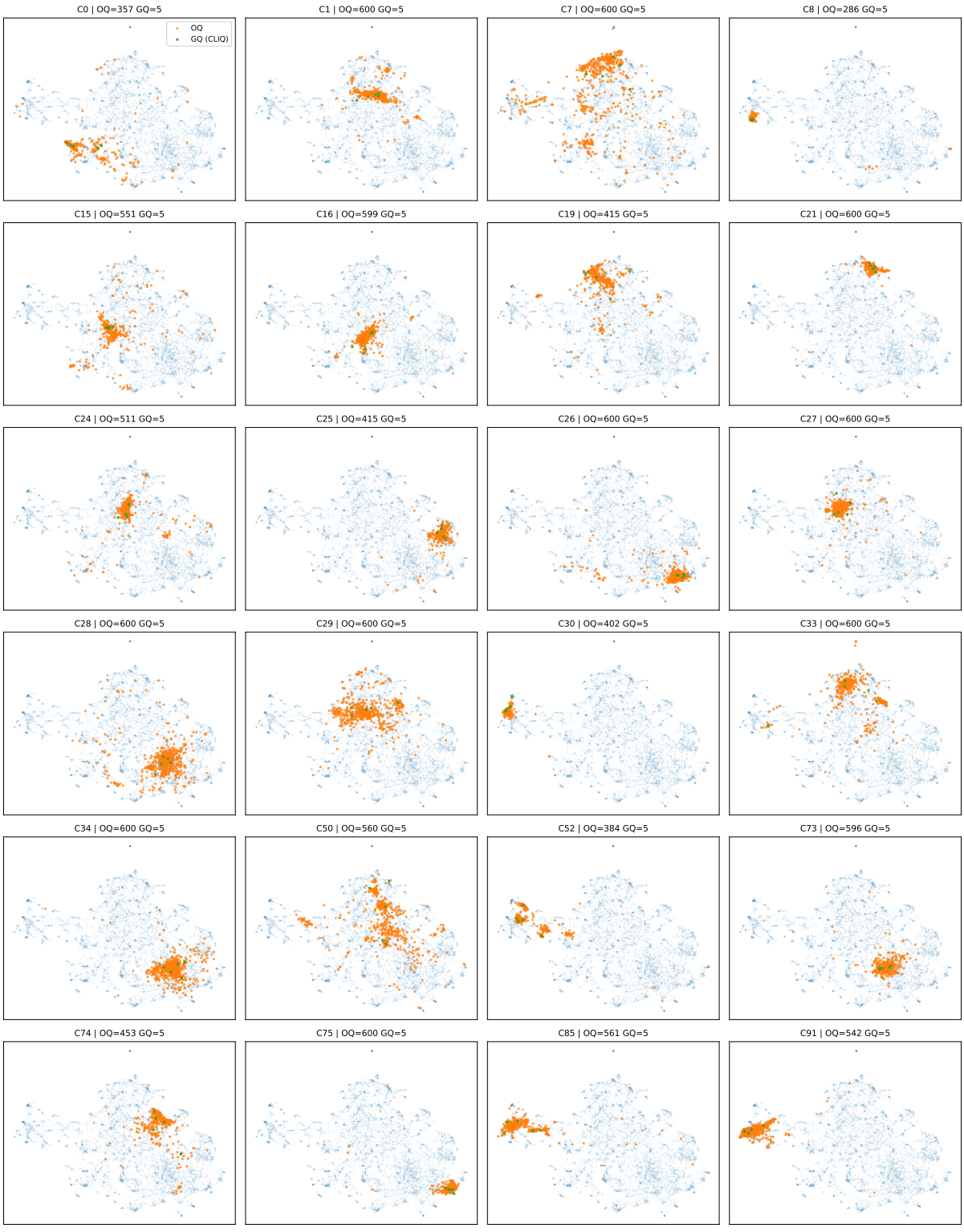

Figure 10: Cluster-conditioned semantic projections of original queries (OQ) and CLIQ-generated queries (GQ). Each subfigure corresponds to one semantic cluster c. Blue points denote original queries in the cluster, while orange points denote generated queries. Only $m = 5$ generated queries per cluster are visualized to illustrate the query efficiency of CLIQ.