QoS-QoE Translation with Large Language Model

Yingjie Yu
yyu69@illinois.edu
University of Illinois
Urbana-Champaign
Urbana, Illinois, USA

Mingyuan Wu
mw34@illinois.edu
University of Illinois
Urbana-Champaign
Urbana, Illinois, USA

Ahmadreza Eslaminia
ae15@illinois.edu
University of Illinois
Urbana-Champaign
Urbana, Illinois, USA

Lingzhi Zhao
lz26@illinois.edu
University of Illinois
Urbana-Champaign
Urbana, Illinois, USA

Kaizhuo Yan
kaizhuo2@illinois.edu
University of Illinois
Urbana-Champaign
Urbana, Illinois, USA

Klara Nahrstedt
klara@illinois.edu
University of Illinois
Urbana-Champaign
Urbana, Illinois, USA

Abstract

QoS-QoE translation is a fundamental problem in multimedia systems because it characterizes how measurable system and network conditions affect user-perceived experience. Although many prior studies have examined this relationship, their findings are often developed for specific setups and remain scattered across papers, experimental settings, and reporting formats, limiting systematic reuse, cross-scenario generalization, and large-scale analysis. To address this gap, we first introduce *QoS-QoE Translation* dataset, a source-grounded dataset of structured QoS-QoE relationships from the multimedia literature, with a focus on video streaming related tasks. We construct the dataset through an automated pipeline that combines paper curation, QoS-QoE relationship extraction, and iterative data evaluation. Each record preserves the extracted relationship together with parameter definitions, supporting evidence, and contextual metadata. We further evaluate the capability of large language models (LLMs) on QoS-QoE translation, both before and after supervised fine-tuning on our dataset, and show strong performance on both continuous-value and discrete-label prediction in bidirectional translation, from QoS-QoE and QoE-QoS. Our dataset provides a foundation for benchmarking LLMs in QoS-QoE translation and for supporting future LLM-based reasoning for multimedia quality prediction and optimization. The complete dataset and code are publicly available at https://yyu6969.github.io/qos-qoe-translation-page/, for full reproducibility and open access.

CCS Concepts

• **Information systems** → **Multimedia databases**; • **Computing methodologies** → **Information extraction**; • **Networks** → **Network performance analysis**.

Keywords

Quality of Service, Quality of Experience, Large Language Model, Multimedia Databases, Multimedia Systems, Benchmark Dataset

1 Introduction

Quality of Service (QoS) and Quality of Experience (QoE) are two central concepts in multimedia systems. QoS describes measurable system, network, and service conditions such as bitrate, delay, jitter, packet loss, and rebuffering, while QoE reflects users' perceived quality of the delivered service [13]. Understanding the QoS-QoE relationship is important for multimedia applications because it supports system design, adaptive streaming, network optimization, and user-centered quality prediction [1, 4].

A large body of prior work has studied QoS-QoE relationships in multimedia applications, especially video streaming, by modeling how QoS factors map to perceived QoE, and in some cases how QoE targets guide adaptation decisions, using methods such as subjective experiments, heuristic rules, analytical modeling, and machine learning-based prediction [1, 5, 20, 33]. These studies have clarified how factors such as bitrate adaptation, stalling, startup delay, resolution changes, and network impairments affect perceived quality. However, many of these approaches are developed for particular setups and validated under specific conditions, which makes their findings and models difficult to generalize across scenarios. Applying them to new settings often requires substantial re-modeling, additional measurements, or new subjective studies.

These limitations motivate a more unified QoS-QoE translation capability that can support both forward translation from QoS to QoE and reverse translation from QoE targets to QoS conditions across diverse scenarios. Achieving this goal requires both strong models and high-quality data. LLM-based systems are a promising foundation because they have shown strong potential in multimedia-related tasks such as video understanding, audio processing, and multimodal agent-style decision making, while also supporting flexible reasoning and structured prediction [12, 15, 17, 26]. At the same time, constructing suitable source-grounded data is challenging because QoS-QoE relationships are scattered across the literature, reported in heterogeneous forms such as text, tables, figures, and equations, and often accompanied by incomplete or implicit contextual metadata.

To address this gap, we present *QoS-QoE Translation* dataset, a source-grounded dataset of structured QoS-QoE relationships from the literature, with a current focus on video streaming. Our goal is to transform prior studies into a reusable data resource for the multimedia community. Each entry preserves the extracted relationship together with supporting evidence and contextual metadata, enabling interpretability and reproducibility. Because reported QoS-QoE relationships rarely appear in a single uniform format and often must be recovered from multiple forms of source evidence together with their surrounding context, we construct the dataset through a pipeline that combines paper curation, QoS-QoE

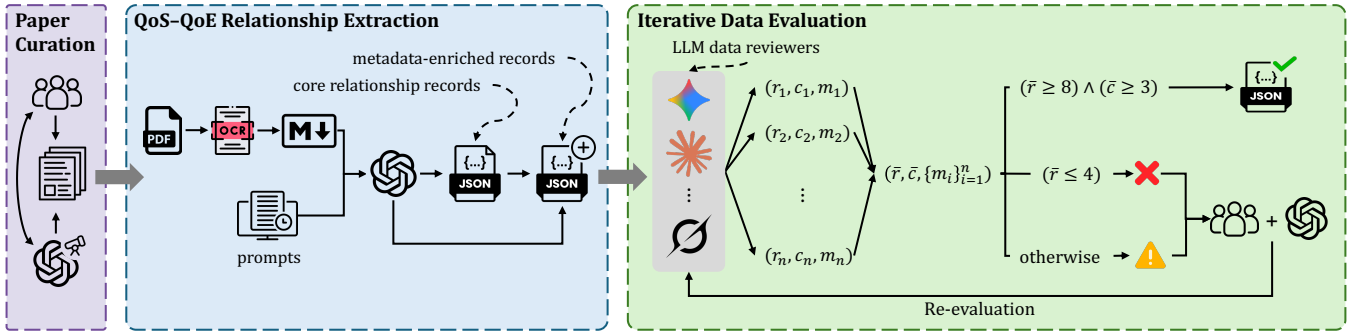

Figure 1: Overview of the QoS-QoE Translation dataset construction pipeline. The pipeline begins with paper curation, followed by QoS-QoE relationship extraction and iterative data evaluation.

relationship extraction, and iterative data evaluation, as shown in Figure 1. This design supports large-scale dataset construction while maintaining quality control and traceability. To assess the utility of *QoS-QoE Translation*, we perform supervised fine-tuning (SFT) of large language models (LLMs) on bidirectional QoS-QoE translation tasks and evaluate both continuous value and discrete label prediction. Results show strong performance gain, with the best fine-tuned model achieving 90.24% Accuracy for discrete label prediction and 8.49% MAPE (Mean Absolute Percentage Error) for continuous value prediction. These findings suggest that *QoS-QoE Translation* provides a strong foundation for training LLMs to reason about QoS-QoE relationships and opens up new opportunities for applying LLM in multimedia applications.

The main contributions of this work are three-fold: 1) We introduce *QoS-QoE Translation*, a source-grounded dataset of structured QoS-QoE relationships from the literature, with a current focus on video streaming. 2) We develop a reusable dataset construction pipeline for paper curation, relationship extraction, metadata enrichment, and iterative multi-reviewer quality evaluation. 3) We demonstrate that the dataset supports effective SFT of LLMs for bidirectional QoS-QoE translation and are the first to benchmark existing open-source LLMs in this domain.

2 Dataset Construction

Figure 1 overviews the *QoS-QoE Translation* construction pipeline, which includes paper curation, QoS-QoE relationship extraction, and iterative data evaluation. Together, these stages transform curated papers into structured records and improve their quality through iterative review. Although *QoS-QoE Translation* focuses on video streaming, the pipeline is reusable and can be adapted to other application domains that require extracting source-grounded relationships from the literature.

2.1 Paper Curation

We begin by constructing a curated corpus of research papers on QoS-QoE relationships in video streaming. To identify relevant and high-quality studies, we combine human screening with OpenAI deep research-assisted literature search [21]. We focus on papers published between 2017 and 2025, since advances in streaming systems, codecs, devices, and network configurations can change the practical meaning of reported QoS-QoE relationships over time. This restriction emphasizes recent and practically relevant evidence while reducing noise from older system settings. Using this human-AI curation process, we collect 505 papers related to QoS-QoE relationships in video streaming, which serve as the foundation for the downstream extraction pipeline and improve the relevance and reliability of the final dataset.

2.2 QoS-QoE Relationship Extraction

Starting from the curated paper corpus, we perform QoS-QoE relationship extraction from academic papers. Because the source papers are provided in PDF format, we first use MinerU [28], an OCR-based document parsing tool, to convert them into machine-readable markdown while preserving textual content and document structure for downstream processing. The converted content, together with carefully designed prompts, is then provided to an LLM for structured information extraction.

We use GPT-5.2 Thinking [22] as the core extraction model because its reasoning ability and long-context support make it well suited for source-grounded extraction from complex academic papers. During development, it provided a practical balance between extraction quality and cost for large-scale dataset construction.

This extraction stage produces two levels of outputs. The first is a set of core relationship records, which capture the fundamental QoS-QoE relationships extracted from source evidence such as equations, tables, and figures. The second is a set of metadata-enriched records, which augment the core relationship records with contextual metadata such as protocol, network type, device type, and scenario. By separating core relationship extraction from contextual metadata enrichment, the pipeline keeps the extracted relationships grounded in source evidence while still providing richer context for downstream analysis and reuse.

2.3 Iterative Data Evaluation

To improve data quality, we further introduce an iterative data evaluation stage. As shown in Figure 1, each metadata-enriched record is reviewed by multiple LLM-based data evaluators. In the current dataset construction setup, we instantiate this stage with three data reviewers: Gemini-2.5-flash-lite [9], Claude-haiku-4-5-20251001 [2], and Grok-4.20-0309-reasoning [31]. Each reviewer

Figure 2: An example JSON record from the QoS-QoE Translation, showing enriched metadata and core relationship.

provides a rating, a confidence score, and written feedback, denoted in the figure as a tuple of the form (r_i, c_i, m_i). The rating score r_i is assigned on a 0–10 scale, where 0, 2, 4, 6, 8, and 10 denote strong reject, reject, weak reject, weak accept, accept, and strong accept, respectively. The confidence score c_i is assigned on a 1–5 scale, where higher values indicate stronger reviewer confidence in the judgment. Reviewer comments are also required to describe the identified issues and suggest possible solutions.

These reviewer outputs are aggregated into a decision using the average rating $\bar{r}$, the average confidence $\bar{c}$, and the collection of reviewer feedback messages. Based on our empirical inspection of reviewer outputs during dataset construction, we found the following thresholds to provide a reasonable balance between retaining high-quality records and filtering out unreliable extractions.

$$\text{Decision} = \begin{cases} \textbf{Accept}, & \text{if } \bar{r} \geq 8 \text{ and } \bar{c} \geq 3, \\ \textbf{Reject}, & \text{if } \bar{r} \leq 4, \\ \textbf{Conditional Accept}, & \text{otherwise.} \end{cases} \quad (1)$$

Records that satisfy the accept condition are retained as valid JSON entries, while records that satisfy the reject or conditional accept conditions are sent to a re-evaluation stage. In this stage, human guidance and an LLM are jointly used to revise the data before returning it to the evaluation loop. This iterative mechanism reduces unsupported, ambiguous, or low-quality extraction results and improves the consistency of the final dataset.

3 Dataset Overview and Analysis

QoS-QoE Translation contains 1026 source-grounded QoS-QoE relationship records extracted from 505 curated papers after extraction and iterative data evaluation. Figure 2 shows an example JSON record, Table 1 summarizes the field definitions, and Figure 3 summarizes the dataset composition in terms of metadata, temporal and venue coverage, and QoS/QoE parameter distributions. Each record contains two main components: **metadata**, which captures source information and contextual attributes, and **relationship**, which stores the extracted QoS-QoE relationship. This design preserves contextual information and source-grounded relationships in a unified machine-readable format for downstream analysis, benchmarking, and modeling.

Table 1: Field definitions of QoS-QoE Translation.

Field	Definition
id	Unique identifier for each dataset record.
year	Publication year of the source paper.
venue	Publication venue of the source paper.
domain	Application domain of the record.
protocol	Streaming or transport protocol used in the study.
network_type	Access network environment.
device_type	Client device used for content consumption.
video_type	Video content category or media format.
user_preference	User preference emphasized in the study.
scenario	Summary of the experimental or evaluation setting.
history_log	Temporally ordered QoS-QoE observations.
data_type	Evidence type, such as equation, table, or figure.
qos_parameter	QoS variables in the relationship.
qos_parameter_definition	Definitions of the QoS variables.
qoe_parameter	QoE variables in the relationship.
qoe_parameter_definition	Definitions of the QoE variables.
relationship	Extracted dependency between QoS and QoE.
description	Natural-language explanation of the relationship.
source	Source-grounded evidence trace.

3.1 Metadata Diversity and Coverage

Figures 3a–3d show that the dataset is concentrated in mainstream video streaming settings while still preserving cross-setting diversity. DASH accounts for 59.7% of the protocol distribution, followed by HTTP at 20.2%, while WebRTC, QUIC, RTP, and HLS together contribute a non-trivial share of interactive and transport-level settings. For network type, cellular 4G (33.2%), wired (23.0%), and Internet-based environments (19.3%) dominate, indicating that the dataset covers both mobile and fixed-network evaluations. Device type is largely split between desktop (45.2%) and mobile (41.0%), suggesting that the dataset mainly reflects common end-user viewing platforms. For video type, 2D video-on-demand is the largest category at 61.3%, followed by 2D live streaming at 27.8%, while short video and immersive formats such as 360 and VR video remain represented. Overall, these distributions show that the dataset is anchored in dominant real-world streaming scenarios, but still retains enough diversity to support cross-context analysis.

3.2 Temporal and Venue Coverage

Figure 3e shows that the dataset is concentrated in recent years, with 2024 contributing the largest share at 18.2%, while papers from 2017–2025 remain represented. Figure 3f further shows that the dataset is collected from a broad set of publication venues. ACM MM contributes the largest share at 26.0%, followed by arXiv (16.7%), IEEE Access (15.6%), IEEE INFOCOM (9.5%), ACM TOMM (8.2%), NSDI (7.6%), and so on. Together, these distributions show that *QoS-QoE Translation* is grounded in both multimedia and networking communities, while maintaining strong coverage of recent research and reducing noise from older system settings.

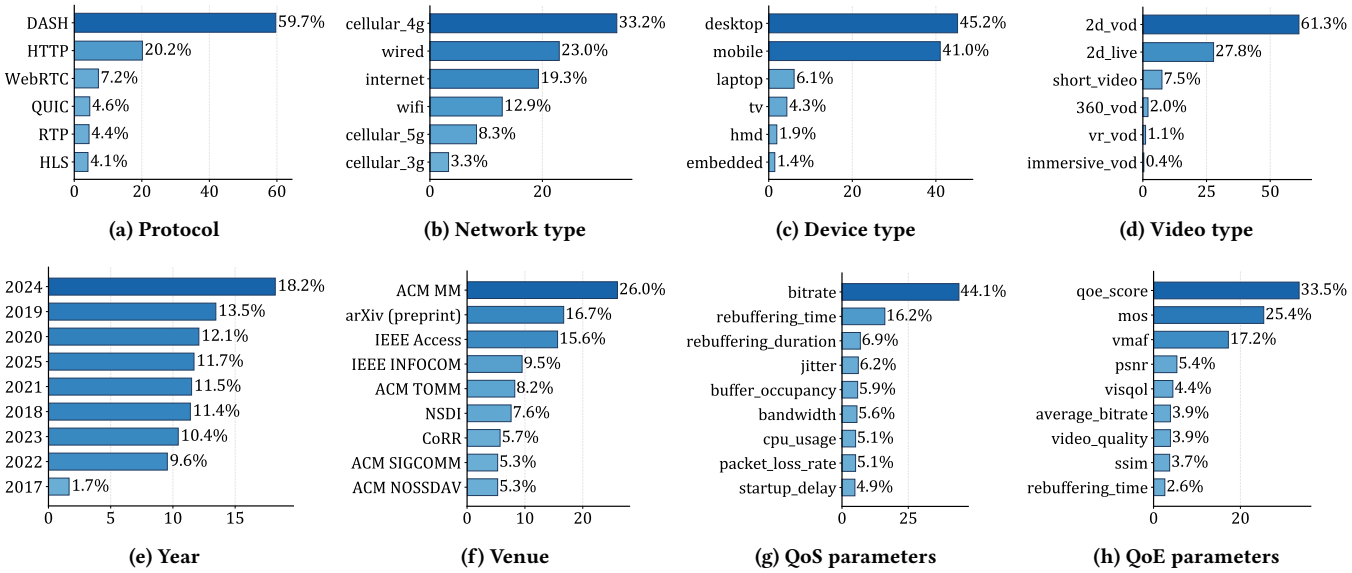

Figure 3: Dataset analysis of QoS-QoE Translation. The first row shows key metadata distributions, including protocol, network type, device type, and video type. The second row shows distributions over year, venue, and QoS/QoE parameters.

3.3 QoS and QoE Parameter Coverage

Figures 3g and 3h show that the dataset covers the most commonly studied QoS and QoE parameters in the literature. On the QoS side, bitrate is the most frequent parameter at 44.1%, followed by rebuffering time at 16.2%, while rebuffering duration, jitter, buffer occupancy, bandwidth, CPU usage, packet loss rate, and start up delay each appear in around 5–7% of records. This pattern shows that the dataset emphasizes system-level factors that are central to adaptive streaming and quality degradation. On the QoE side, QoE score (33.5%) and MOS (25.4%) are the most frequent targets, followed by VMAF at 17.2%, while PSNR, VISQOL, average bitrate, video quality level, SSIM, and rebuffering time appear less frequently. These results indicate that the dataset covers both subjective QoE indicators and objective quality metrics, making it suitable for translation across heterogeneous QoE formulations. Notably, some variables, such as rebuffering time, appear on both the QoS and QoE sides, highlighting that the same concept may be treated differently across studies depending on how the authors define system conditions and user experience outcomes.

3.4 Dataset Availability and Licensing

QoS-QoE Translation is made publicly available on our project website[1] to encourage research. The released dataset consists of processed structured JSON records derived from the literature and is distributed under the Creative Commons Attribution 4.0 International (CC BY 4.0) license.

4 Experiments

The QoS-QoE translation task evaluates whether a model can predict user-experience outcomes from system-level service conditions, or predict system-level service conditions from user-experience observations. The experimental results are summarized in Tables 2 and 3. The evaluated models are Qwen3-8B, Qwen3-32B, Qwen3.5-35B-A3B, Llama-3.1-8B-Instruct, and Llama-3.3-70B-Instruct [10, 23, 32]. Table 2 reports the overall performance before and after SFT, and Table 3 presents results by translation direction after SFT.

Task Formulation. We formulate both directions, QoS→QoE and QoE→QoS, as structured prediction tasks derived from the source-grounded relationship records in our dataset. In the QoS→QoE direction, the model predicts user-experience parameters from system-level conditions. In the QoE→QoS direction, the model predicts system-level conditions from user-experience observations. In both cases, the model receives a structured input instance with contextual information and predicts the queried target field in JSON.

Supervised Fine-tuning. We use the Tinker framework [27] for SFT on our QoS-QoE translation tasks. Each example is represented as a multi-turn chat-style JSON instance. The input includes an instruction, task identifier, contextual metadata, scenario description, parameter mapping, source-grounded evidence, history log, and a query, while the target output contains only the predicted JSON field for the queried task.

To construct the SFT corpus, we transform the 1026 source-grounded relationship records into 8107 chat-style instances through holdout-based history reconstruction. For each instance, one target time point from the history log is held out as the query/output pair, and the remaining time points are retained as input context. We use 7205 instances for training and 902 for testing. Unless otherwise noted, all evaluated models use the same split and training configuration. We use the default SFT configuration in Tinker, with a maximum sequence length of 32,768 tokens, batch size 128, learning rate 2×10^{-4}, a linear learning-rate schedule, and 1 training epoch.

Metrics. We report four evaluation metrics in our QoS-QoE translation tasks. For continuous value prediction, we use MAPE and

[1]Dataset website: https://yyu6969.github.io/qos-qoe-translation-page/

Table 2: Overall model evaluation results before and after SFT. All values are percentages (%).

Model	Before SFT				After SFT			
	MAPE↓	Accuracy@δ↑	Accuracy↑	Macro-F1↑	MAPE↓	Accuracy@δ↑	Accuracy↑	Macro-F1↑
Qwen3-8B	20.23	64.20	67.48	55.63	11.79	78.13	80.49	72.29
Qwen3-32B	16.46	67.95	68.29	56.69	9.41	80.24	84.55	70.12
Qwen3.5-35B-A3B	14.43	**72.94**	69.92	**61.33**	**8.49**	**83.41**	**90.24**	**84.63**
Llama-3.1-8B-Instruct	26.72	55.90	50.41	35.62	11.34	79.49	83.74	70.79
Llama-3.3-70B-Instruct	**13.76**	72.70	**70.73**	56.29	9.49	81.90	**90.24**	81.93

Table 3: Overall model evaluation results after SFT for QoS→QoE and QoE→QoS translation. All values are percentages (%).

Task	Model	MAPE (%)↓	Accuracy@δ (%)↑	Accuracy (%)↑	Macro-F1 (%)↑
QoS → QoE	Qwen3-8B	9.11	78.55	75.00	73.43
	Qwen3-32B	7.41	82.61	77.50	62.18
	Qwen3.5-35B-A3B	7.05	**83.77**	77.50	66.53
	Llama-3.1-8B-Instruct	8.79	81.74	75.00	61.31
	Llama-3.3-70B-Instruct	**6.88**	**83.77**	**85.00**	**77.78**
QoE → QoS	Qwen3-8B	13.47	77.67	83.13	71.43
	Qwen3-32B	10.66	77.67	87.95	73.45
	Qwen3.5-35B-A3B	**9.40**	**83.02**	**96.39**	**91.81**
	Llama-3.1-8B-Instruct	12.94	77.04	87.95	73.79
	Llama-3.3-70B-Instruct	11.12	79.87	92.77	82.74

Accuracy@δ. MAPE measures the average percentage error between predicted and ground-truth values, where lower is better. Accuracy@δ measures the fraction of predictions that fall within a predefined tolerance of the ground truth. Because different QoS and QoE parameters have different acceptable error ranges, we use a parameter-specific δ rather than a single shared threshold. Depending on the parameter, δ is defined as either an absolute or relative tolerance. The full δ configuration is provided in the supplementary material. For discrete label prediction, we report Accuracy, which measures exact label matches, and Macro-F1, which averages F1 equally across classes and is less sensitive to class imbalance.

4.1 Overall Performance

Table 2 summarizes the overall performance of each model before and after SFT. After fine-tuning, all evaluated models perform reasonably well on both continuous value and discrete label prediction, suggesting that *QoS-QoE Translation* provides a useful supervision signal. Within the same model family, larger models generally outperform smaller ones. For example, Qwen3-32B improves over Qwen3-8B after SFT, reducing MAPE from 11.79% to 9.41% and increasing discrete label accuracy from 80.49% to 84.55%. Llama-3.3-70B-Instruct also consistently outperforms Llama-3.1-8B-Instruct across all reported post-SFT metrics.

Across model families, Qwen3.5-35B-A3B achieves the strongest overall post-SFT performance. It obtains the best MAPE of 8.49%, the highest Accuracy@δ of 83.41%, ties for the highest discrete label accuracy at 90.24%, and achieves the best Macro-F1 of 84.63%. Llama-3.3-70B-Instruct is also highly competitive, reaching 9.49% MAPE, 81.90% Accuracy@δ, tying for the best discrete label accuracy at 90.24%, and obtaining the second-best Macro-F1 of 81.93%.

Comparing performance before and after SFT, all evaluated models show clear gains on both continuous value and discrete label prediction, indicating that *QoS-QoE Translation* provides effective supervision for bidirectional QoS-QoE translation. For continuous value prediction, Qwen3-8B reduces MAPE from 20.23% to 11.79% and improves Accuracy@δ from 64.20% to 78.13%, while Qwen3.5-35B-A3B improves from 14.43% to 8.49% in MAPE and from 72.94% to 83.41% in Accuracy@δ. Llama-3.1-8B-Instruct shows the largest gain, with MAPE decreasing from 26.72% to 11.34% and Accuracy@δ increasing from 55.90% to 79.49%.

The gains are also substantial for discrete label prediction. Qwen3.5-35B-A3B improves from 69.92% to 90.24% in accuracy and from 61.33% to 84.63% in Macro-F1, while Llama-3.3-70B-Instruct improves from 70.73% to 90.24% and from 56.29% to 81.93%, respectively. Llama-3.1-8B-Instruct again shows the largest improvement, with accuracy increasing from 50.41% to 83.74% and Macro-F1 increasing from 35.62% to 70.79%. Overall, these results show that *QoS-QoE Translation* provides an effective supervision signal, leading to strong post-SFT performance and consistent improvements over the corresponding pre-trained baselines.

4.2 Bidirectional Analysis

Table 3 reports post-SFT results for the two translation directions separately. Overall, QoS→QoE appears slightly easier for continuous value prediction, with the best model reaching a MAPE of 6.88%, compared with 9.40% for QoE→QoS. In the QoS→QoE setting, Llama-3.3-70B-Instruct performs best overall, achieving the lowest MAPE of 6.88%, tying for the highest Accuracy@δ of 83.77%, and obtaining the best discrete label results with 85.00% accuracy and 77.78% Macro-F1. In the QoE→QoS setting, Qwen3.5-35B-A3B performs best overall, with the lowest MAPE of 9.40%, the highest

Accuracy@δ of 83.02%, the highest categorical accuracy of 96.39%, and the highest Macro-F1 of 91.81%.

The two directions exhibit different patterns. QoE→QoS is generally more challenging for continuous value prediction, as reflected by higher MAPE values across models, but it yields better discrete label results, especially for Qwen3.5-35B-A3B. One possible explanation is that multiple QoS configurations can correspond to similar QoE outcomes, which makes reverse numeric translation more ambiguous. At the same time, the strong discrete label performance in QoE→QoS suggests that although precise numeric recovery is harder, coarse-grained reverse translation remains highly learnable.

5 Related Work

5.1 QoS-QoE Modeling and Analysis

QoS-QoE relationships have been widely studied in multimedia systems [1, 13]. Prior studies have examined how factors such as bitrate, delay, packet loss, startup latency, stalling, and adaptation behavior affect user-perceived quality across multimedia applications, especially video streaming, using methods including subjective experiments, correlation analysis, analytical modeling, and machine learning-based prediction [5]. Representative studies further showed that video quality impairments and stalling events can strongly affect user engagement and viewer behavior [7, 14], and developed predictive models for Internet video QoE using system- and network-level features [3]. Survey papers have further summarized a broad range of QoS-QoE modeling approaches and challenges in HTTP adaptive streaming [1, 5, 24]. However, these works mainly focus on understanding or predicting QoE from experimental observations, rather than transforming the published literature itself into a structured and source-grounded dataset for systematic reuse.

5.2 QoE Datasets and Benchmark Resources

A separate line of work has produced datasets and benchmark resources for QoE research. Many of these datasets are derived from controlled subjective studies or system-level measurements and provide annotated samples for evaluating QoE prediction models [8, 16, 34]. They have been valuable for benchmarking and model development, particularly in adaptive video streaming and related applications [8, 34]. However, such resources are typically tied to specific experimental settings, user studies, or measurement campaigns. In contrast, our goal is not to build another experiment-specific QoE benchmark, but to curate a literature-grounded dataset of reported QoS-QoE relationships together with supporting evidence, parameter definitions, and contextual metadata.

5.3 LLM-based Scientific Document Extraction

Recent LLMs have shown strong capabilities in information extraction and structured generation from scientific text and documents [6, 25]. Multimodal and layout-aware document models extend these abilities to tables, forms, and rich documents, making them promising tools for document understanding and literature mining [29]. Iterative refinement and LLM-based evaluation frameworks have also been explored to improve quality through feedback and aggregation [11, 19]. Our work builds on these advances, but differs in objective: rather than using LLMs for generic scientific document extraction, we use them to construct a source-grounded QoS-QoE dataset from the literature and combine extraction with iterative multi-reviewer evaluation to improve reliability and traceability.

6 Discussion

Although recent large language models have shown strong potential in multimedia applications, important limitations still remain, especially for complex reasoning tasks. In particular, current models are not yet consistently reliable when multi-step inference, temporal understanding, and long-context multimodal reasoning are required [12, 18, 30]. These challenges are especially relevant in multimedia system settings, where useful decisions often depend not only on recognizing visual or audio content, but also on integrating information across modalities and reasoning jointly about system conditions, user experience, and their interactions.

7 Potential Applications and Impact

QoS-QoE Translation can support several practical applications in video streaming systems and intelligent network management. First, it can serve as a supervision source for training models that predict user-perceived quality from measurable system-level signals such as bitrate, delay, packet loss, or rebuffering. Such models can help service providers estimate QoE in real time without relying only on expensive user studies. Second, the dataset can support reverse prediction from QoE targets to QoS conditions, which is useful for resource planning and adaptive system control, where an operator may want to identify what network or application conditions are needed to achieve a desired level of user experience.

Beyond direct prediction, the dataset also provides a structured knowledge base for retrieval and reasoning. Because each record is source-grounded and linked to equations, tables, or figures in the literature, it can be used in retrieval-augmented systems that answer QoS-QoE questions with evidence and support trustworthy decision making.

More broadly, *QoS-QoE Translation* provides a foundation for LLM-based AI agents for QoS-QoE translation. Such agents could parse user goals, retrieve relevant source-grounded relationships, compare evidence across studies, and generate structured predictions or recommendations for network optimization. They could assist with streaming configuration, quality diagnosis, QoE-aware adaptation, and automatic report generation. Future work will focus on extending the dataset to more diverse and complex scenarios and improving the evaluation framework with more human assessment and finer-grained source analysis.

8 Conclusion

In this paper, we present *QoS-QoE Translation*, a source-grounded dataset of QoS-QoE relationships in video streaming. We construct the dataset through a pipeline that combines paper curation, source-grounded relationship extraction, metadata enrichment, and iterative data evaluation, producing machine-readable JSON records with contextual metadata and explicit source traceability. Experiments with supervised fine-tuned large language models show strong performance on both numeric and categorical QoS-QoE translation tasks in both forward and reverse directions, suggesting that *QoS-QoE Translation* is a useful benchmark and training resource for source-grounded QoS-QoE modeling.

References

[1] Mohammed Alreshoodi and John Woods. 2013. Survey on QoE/QoS Correlation Models for Multimedia Services. *International Journal of Distributed and Parallel Systems* 4, 3 (2013), 53–72. https://doi.org/10.5121/ijdps.2013.4305

[2] Anthropic. 2025. Introducing Claude Haiku 4.5. https://www.anthropic.com/news/claude-haiku-4-5 Official model announcement. Accessed: 2026-04-01.

[3] Athula Balachandran, Vyas Sekar, Aditya Akella, Srinivasan Seshan, Ion Stoica, and Hui Zhang. 2013. Developing a Predictive Model of Quality of Experience for Internet Video. *ACM SIGCOMM Computer Communication Review* 43, 4 (2013), 339–350. https://doi.org/10.1145/2486001.2486025

[4] Alcardo Alex Barakabitze, Nabajeet Barman, Arslan Ahmad, Saman Zadtootaghaj, Lingfen Sun, Maria G. Martini, and Luigi Atzori. 2020. QoE Management of Multimedia Streaming Services in Future Networks: A Tutorial and Survey. *IEEE Communications Surveys & Tutorials* 22, 1 (2020), 526–565. https://doi.org/10.1109/COMST.2019.2958784

[5] Nabajeet Barman and Maria G. Martini. 2019. QoE Modeling for HTTP Adaptive Video Streaming: A Survey and Open Challenges. *IEEE Access* 7 (2019), 30831–30859. https://doi.org/10.1109/ACCESS.2019.2901778

[6] John Dagdelen, Alexander Dunn, Sanghoon Lee, Nicholas Walker, Andrew S. Rosen, Gerbrand Ceder, Kristin A. Persson, and Anubhav Jain. 2024. Structured information extraction from scientific text with large language models. *Nature Communications* 15, 1 (2024), 1418. https://doi.org/10.1038/s41467-024-45563-x

[7] Florin Dobrian, Vyas Sekar, Asad Awan, Ion Stoica, Dilip Joseph, Aditya Ganjam, Jibin Zhan, and Hui Zhang. 2011. Understanding the Impact of Video Quality on User Engagement. *ACM SIGCOMM Computer Communication Review* 41, 4 (2011), 362–373. https://doi.org/10.1145/2018436.2018478

[8] Zhengfang Duanmu, Abdul Rehman, and Zhou Wang. 2018. A Quality-of-Experience Database for Adaptive Video Streaming. *IEEE Transactions on Broadcasting* 64, 2 (June 2018), 474–487. https://doi.org/10.1109/TBC.2018.2822870

[9] Google. 2025. Gemini 2.5 Flash-Lite. https://ai.google.dev/gemini-api/docs/models/gemini-2.5-flash-lite Official model documentation. Accessed: 2026-04-01.

[10] Aaron Grattafiori, Abhimanyu Dubey, et al. 2024. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783* (2024). arXiv:2407.21783 https://arxiv.org/abs/2407.21783

[11] Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2024. A Survey on LLM-as-a-Judge. *arXiv preprint arXiv:2411.15594* (2024). https://arxiv.org/abs/2411.15594

[12] Jiaxing Huang and Jingyi Zhang. 2024. A Survey on Evaluation of Multimodal Large Language Models. *arXiv preprint arXiv:2408.15769* (2024). arXiv:2408.15769 [cs.CV] https://arxiv.org/abs/2408.15769

[13] ITU-T. 2023. *Roadmap for QoS and QoE in the ITU-T Study Group 12 Context.* Technical Report GSTR-RQ. International Telecommunication Union. https://www.itu.int/dms_pub/itu-t/opb/tut/T-TUT-QOS-2023-2-PDF-E.pdf

[14] S. Shunmuga Krishnan and Ramesh K. Sitaraman. 2012. Video Stream Quality Impacts Viewer Behavior: Inferring Causality Using Quasi-Experimental Designs. In *Proceedings of the 2012 Internet Measurement Conference (IMC '12)*. Association for Computing Machinery, New York, NY, USA, 211–224. https://doi.org/10.1145/2398776.2398799

[15] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. 2024. MVBench: A Comprehensive Multi-modal Video Understanding Benchmark. *arXiv preprint arXiv:2311.17005* (2024). https://arxiv.org/abs/2311.17005

[16] Yanan Li, Guangqing Deng, Changming Bai, Jingyu Yang, Gang Wang, Hao Zhang, Jin Bai, Haitao Yuan, Mengwei Xu, and Shangguang Wang. 2023. Demystifying the QoS and QoE of Edge-hosted Video Streaming Applications in the Wild with SNESet. *Proceedings of the ACM on Management of Data* 1, 4, Article 236 (2023), 29 pages. https://doi.org/10.1145/3626723

[17] Xiao Liu, Tianjie Zhang, Yu Gu, Iat Long Iong, Yifan Xu, Xixuan Song, Shudan Zhang, Hanyu Lai, Xinyi Liu, Hanlin Zhao, et al. 2024. VisualAgentBench: Towards Large Multimodal Models as Visual Foundation Agents. *arXiv preprint arXiv:2408.06327* (2024). arXiv:2408.06327 [cs.CV] https://arxiv.org/abs/2408.06327

[18] Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. 2024. TempCompass: Do Video LLMs Really Understand Videos?. In *Findings of the Association for Computational Linguistics: ACL 2024*. Association for Computational Linguistics, Bangkok, Thailand, 8731–8772. https://doi.org/10.18653/v1/2024.findings-acl.517

[19] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-Refine: Iterative Refinement with Self-Feedback. *arXiv preprint arXiv:2303.17651* (2023). https://arxiv.org/abs/2303.17651

[20] Hongzi Mao, Ravi Netravali, and Mohammad Alizadeh. 2017. Neural Adaptive Video Streaming with Pensieve. In *Proceedings of the Conference of the ACM Special Interest Group on Data Communication (SIGCOMM '17)*. Association for Computing Machinery, New York, NY, USA, 197–210. https://doi.org/10.1145/3098822.3098843

[21] OpenAI. 2025. Introducing deep research. https://openai.com/index/introducing-deep-research/ Accessed: 2026-03-26.

[22] OpenAI. 2025. Introducing GPT-5.2. https://openai.com/index/introducing-gpt-5-2/ Accessed: 2026-03-27.

[23] Qwen Team. [n. d.]. Qwen3.5-35B-A3B. https://huggingface.co/Qwen/Qwen3.5-35B-A3B Official model card. Accessed: 2026-03-27.

[24] Michael Seufert, Sebastian Egger-Lampl, Martin Slanina, Thomas Zinner, Tobias Hoßfeld, and Phuoc Tran-Gia. 2015. A Survey on Quality of Experience of HTTP Adaptive Streaming. *IEEE Communications Surveys & Tutorials* 17, 1 (2015), 469–492. https://doi.org/10.1109/COMST.2014.2360940

[25] Mahsa Shamsabadi, Jennifer D'Souza, and Sören Auer. 2024. Large Language Models for Scientific Information Extraction: An Empirical Study for Virology. *arXiv preprint arXiv:2401.10040* (2024). https://doi.org/10.48550/arXiv.2401.10040 arXiv:2401.10040

[26] Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2023. SALMONN: Towards Generic Hearing Abilities for Large Language Models. *arXiv preprint arXiv:2310.13289* (2023). arXiv:2310.13289 [cs.SD] https://arxiv.org/abs/2310.13289

[27] Thinking Machines Lab. 2025. Tinker. https://thinkingmachines.ai/tinker/ Accessed: 2026-04-02.

[28] Bin Wang, Chao Xu, Xiaomeng Zhao, Linke Ouyang, Fan Wu, Zhiyuan Zhao, et al. 2024. MinerU: An Open-Source Solution for Precise Document Content Extraction. *arXiv preprint arXiv:2409.18839* (2024). arXiv:2409.18839 [cs.CV] https://arxiv.org/abs/2409.18839

[29] Dongsheng Wang, Natraj Raman, Mathieu Sibue, Zhiqiang Ma, Petr Babkin, Simerjot Kaur, Yulong Pei, Armineh Nourbakhsh, and Xiaomo Liu. 2024. DocLLM: A Layout-Aware Generative Language Model for Multimodal Document Understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Bangkok, Thailand, 8529–8548. https://doi.org/10.18653/v1/2024.acl-long.463

[30] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. 2024. LongVideoBench: A Benchmark for Long-context Interleaved Video-Language Understanding. *arXiv preprint arXiv:2407.15754* (2024). https://doi.org/10.48550/arXiv.2407.15754 arXiv:2407.15754 [cs.CV]

[31] xAI. [n. d.]. Grok 4.20 0309 Reasoning. https://docs.x.ai/developers/models/grok-4.20-0309-reasoning Official model documentation. Accessed: 2026-04-01.

[32] 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, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388* (2025). https://doi.org/10.48550/arXiv.2505.09388 arXiv:2505.09388 [cs.CL]

[33] Xiaoqi Yin, Abhishek Jindal, Vyas Sekar, and Bruno Sinopoli. 2015. A Control-Theoretic Approach for Dynamic Adaptive Video Streaming over HTTP. In *Proceedings of the 2015 ACM Conference on Special Interest Group on Data Communication (SIGCOMM '15)*. Association for Computing Machinery, New York, NY, USA, 325–338. https://doi.org/10.1145/2785956.2787486

[34] Zehao Zhu, Wei Sun, Jun Jia, Wei Wu, Sibin Deng, Kai Li, Ying Chen, Xiongkuo Min, Jia Wang, and Guangtao Zhai. 2024. Subjective and Objective Quality-of-Experience Evaluation Study for Live Video Streaming. *arXiv preprint arXiv:2409.17596* (2024). https://doi.org/10.48550/arXiv.2409.17596 arXiv:2409.17596