Hierarchical User Intent Inference with Knowledge Graph Grounding

Tzu-Cheng Peng
National Taiwan University
Taiwan
d13725002@ntu.edu.tw

Chien Chin Chen
National Taiwan University
Taiwan
patonchen@ntu.edu.tw

Yung-Chun Chang
Taipei Medical University
Taiwan
changyc@tmu.edu.tw

Abstract

Understanding user intent in online reviews requires modeling not only explicit aspect ratings but also implicit motivations shaped by contextual factors. Existing large language models (LLMs) often lack structured grounding, fail to capture nuanced intent expression. We propose **HII-KG**, a two-stage Hierarchical Intent Inference framework that first predicts fine-grained aspect ratings and then generates natural language intent statements, guided by contextual subgraphs retrieved from a domain-specific knowledge graph (KG). We first employ parameter-efficient fine-tuning of LLaMA3.1-8B to predict aspect ratings in an instruction-based format. Moreover, we leverage Cypher-aware prompting to generate user intent from KG summaries. Experiments on a online hotel review dataset show that HII-KG consistently outperforms strong LLM and encoder-based baselines in both aspect classification (avg. F1 +4.5%) and intent generation (BLEU +3.3, ROUGE-L +2.9). The results demonstrate that structured KG integration can significantly enhance fluency, contextual relevance, and factual alignment in user intent inference.

1 Introduction

Understanding user intent from online reviews is crucial for enabling adaptive services in tourism industry. Unlike product reviews, hotel reviews often contain context-specific expectations shaped by a guest's purpose of visit, such as a family vacation, solo trip, or business travel. These subtle contextual signals influence how users evaluate service quality, room conditions, or overall satisfaction. However, most existing NLP systems for review understanding process textual input in isolation—ignoring these implicit cues—and rely solely on sentiment or aspect keyword extraction. (Ku et al., 2024a; Zhang and Niu, 2024; Chang et al., 2020)

While recent developments in LLMs have revolutionized review analysis by capturing nuanced semantics and generating abstractive summaries (Krugmann and Hartmann, 2024; Achiam et al., 2023), they often lack grounding in structured domain knowledge and user-aware contextualization. In parallel, KG integration has emerged as a powerful technique for augmenting LLMs with relational semantics and improving consistency across tasks such as recommendation, explanation, and classification (Wang et al., 2025).

In the domain of hotel review analysis, prior work has highlighted the strategic variation in user expressions and the importance of adapting to these nuances in response generation and customer modeling (Ku et al., 2024b; Chang et al., 2019). Despite this, little work has focused on jointly leveraging structured knowledge and contextual grounding to improve both aspect-level rating prediction and intent inference. To address this gap, we propose **HII-KG**, a *Hierarchical Intent Inference* framework with *Knowledge Graph* augmentation, which operates in two stages:

- **Stage 1:** Predicts six fine-grained aspect ratings (e.g., value, cleanliness, location) using a parameter-efficient fine-tuning of LLaMA-3.1-8B.
- **Stage 2:** Constructs user intent representations by leveraging Cypher-query–based subgraphs retrieved from the domain-specific knowledge graph.

Our results demonstrate that this hierarchical structure, combined with KG-grounded context, significantly improves aspect rating accuracy and enhances the fluency and consistency of generated intent summaries. We argue that even lightweight contextual signals, when properly aligned with KG, can lead to meaningful improvements in interpretability and downstream understanding of user needs.

Findings of the Association for Computational Linguistics: EACL 2026, pages 5371–5377

2 Related Work

Inferring user intent from textual reviews is inherently ambiguous and context-dependent. While LLMs have enabled direct inference through prompting and fine-tuning, they often lack structured grounding. This motivates integrating KGs with LLMs for enhanced reasoning and interpretability.Structured knowledge from KGs can guide LLM inference, reduce hallucination, and improve transparency. Pan et al. (2024) identify three paradigms of LLM-KG integration: KG-enhanced LLMs, LLM-augmented KGs, and synergistic frameworks. For instance, Chen et al. (2024) treat the LLM as an agent interacting with a KG, while Xu et al. (2024b) embed KG traversal into retrieval-augmented generation. These works show how KG structure can scaffold LLM reasoning, particularly when tasks require structured outputs. In review understanding, this helps anchor generative summaries to known entities. Zheng et al. (2025) enrich user–item graphs with intent-aware subgraphs for generative LLMs, showing applicability to recommendation and review domains. Ma et al. (2025) provide design patterns and evaluation concerns for LLM+KG pipelines.

Meanwhile, a complementary line of research has focused on LLM-only methods for intent and review understanding. Arora et al. (2024) benchmark multi-intent classification strategies. Bodonhelyi et al. (2024) emphasize clear intent representation via prompt reformulation. Rodriguez et al. (2024) introduce *IntentGPT* for few-shot novel intent discovery. In the tourism domain, Ouaddi et al. (2025) test LLM generalization across domain-specific labels. Lin et al. (2024) use LLM-generated candidates to augment zero-resource classifiers. While these models achieve strong results but often lack grounding in structured knowledge. Our proposed **HII-KG** bridges this gap by combining KG-based subgraph retrieval with LLM-driven generation. It inherits best practices from LLM pipelines (e.g., fluency, prompt design) while introducing graph-based priors for structured, domain-aware intent understanding.

3 HII-KG Framework

The proposed **HII-KG**, illustrated in Figure 1, is a two-stage Hierarchical Intent Inference framework that integrates structured aspect rating prediction with user intent inference grounded in domain-specific knowledge graphs. In this work, we focus on the domain of hotel reviews, where user intent is shaped by both explicit service aspects (e.g., value, cleanliness) and implicit motivations (e.g., family-oriented preferences, desire for a quiet environment). The framework is thus designed to capture the compositional nature of user needs by jointly modeling these auxiliary information.

3.1 Multi-Aspect Rating Prediction

The first stage formulates hotel review understanding as a multi-task classification problem. Given a raw review x, the model predicts six numerical aspect ratings $[r_1, r_2, ..., r_6] \in \{1, 2, 3, 4, 5\}^6$, where each r_i corresponds to an aspect from *Value*, *Location*, *Rooms*, *Service*, *Sleep Quality*, and *Cleanliness*. We adopt a parameter-efficient fine-tuning approach based on fine-tuned **LLaMA-3.1-8B**, using LoRA adapters to reduce computational cost, and formatting the input sequences as an instruction-style prompts.The model directly generates a structured rating sequence (e.g., `[4, 5, 3, 4, 5, 5]`) as free-form text, without using separate classification heads for each aspect. This is achieved by formatting the input in instruction style and training the model using standard causal language modeling loss (token-level cross-entropy) over the entire generated sequence. The model learns to decode the six aspect ratings in a consistent order, forming a compact representation of user satisfaction across multiple service dimensions.

3.2 KG Construction and Querying

Our **Hotel-KG** is constructed from 399,000 reviews across 745 hotels on TripAdvisor. Nodes include concepts such as `Facility`, `RoomType`, `TravelPurpose`, `Aspect`, and `Region`. Edges are built from review metadata and textual co-occurrence, enabling the encoding of meaningful semantic relationships such as:

- `(RoomType)` → `[:HAS_FACILITY]` → `(Facility)`
- `(TravelPurpose)` → `[:PREFERS]` → `(RoomType)`
- `(Aspect)` → `[:MENTIONED_IN]` → `(Review)`

Each review is preprocessed using an entity linking pipeline to detect and normalize spans corresponding to KG nodes. For instance, "indoor pool" maps to `Facility:pool`, while "suite" maps to `RoomType:suite`. Based on these anchor points, Cypher queries are generated to extract subgraphs reflecting travel-type, facility, and room-type preferences. The resulting subgraph is then linearized

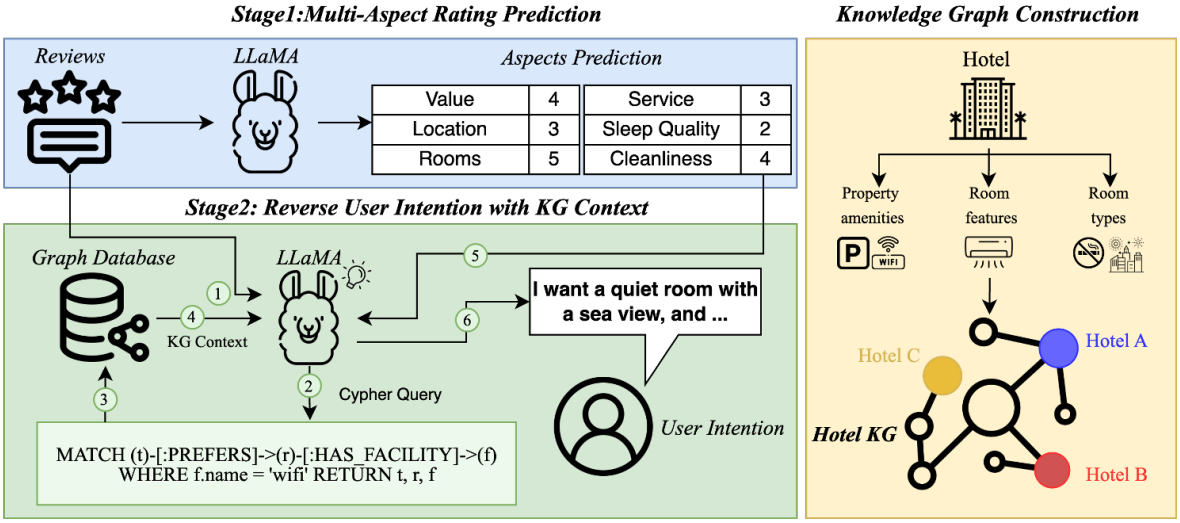

Figure 1: The HII-KG framework: a two-stage pipeline that first predicts structured aspect ratings, then retrieves KG context to guide intent generation. Cypher queries dynamically extract subgraphs relevant to the review content.

into a structured natural language template (e.g., "Families often prefer rooms with pool access."), providing factual and user-type-aware context to support downstream intent inference.

3.3 Intent Generation with KG Context

With the contextual subgraph retrieved and linearized, the second stage infers the reviewer's likely intent in natural language form. We employ the same backbone as in Stage 1, **Llama3.1-8B-Cypher** (Xu et al., 2024a), a Cypher-aware large language model pretrained on KG query generation tasks. Owing to its architectural bias toward structured reasoning, the model is particularly well-suited for incorporating KG-derived contextual inputs. Instead of parameter fine-tuning, we adopt a few-shot prompting strategy. For each instance, the input includes a KG-derived context string, followed by the raw review and an intent generation instruction. For example: given **KG Context:** *"Family travelers prefer suites with kitchen and pool access."*, **Review:** *"The kids loved the pool and the room was spacious..."*, and **Task:** *"Based on the context, reverse the user intention."* The model then generates intent expressions such as *"I am looking for a family-friendly place with a big room and a nice pool."*

During inference, the overall pipeline proceeds as follows: (1) aspect scores are predicted using Stage 1, (2) relevant KG subgraphs are retrieved based on the review content, and (3) intent statements are generated via few-shot prompting using both the review and KG context. This hierarchical flow allows the model to leverage explicit signals for interpretability and structured knowledge for contextual grounding.

4 Experiments

4.1 Aspect Rating Prediction

We first assess our model's ability to predict the six aspect scores using a curated dataset of 9,000 TripAdvisor hotel reviews annotated with the following aspects: *Value*, *Location*, *Rooms*, *Service*, *Sleep Quality*, and *Cleanliness*. Each review includes the raw text, travel type metadata, and complete aspect scores on a 1–5 scale. Reviews are filtered to exclude non-English content, short texts under 50 characters, and missing aspect labels. Label distribution statistics are shown in Appendix A. We compare HII-KG with strong baselines consisting of both encoder-only and encoder–decoder models: BERT, RoBERTa, BART, and T5.

As shown in Figure 2, HII-KG consistently achieves the highest F1-scores across all six aspects, outperforming both encoder-only (BERT, RoBERTa) and encoder–decoder (BART, T5) baselines. In particular, our model attains 0.87 on *Value* and 0.85 on *Sleep Quality*, two aspects with high variance in user ratings, indicating its ability to capture subtle user preferences and contextual cues. While BART and T5 show improved performance

over BERT and RoBERTa, especially on aspects with clearer surface signals such as *Location* (0.86) and *Rooms* (0.82), they lag behind HII-KG on precision-sensitive dimensions like *Service* (0.81 vs. 0.86) and *Cleanliness* (0.88 vs. 0.91). This suggests that simply scaling to larger encoder–decoder architectures is insufficient for aspects requiring deeper contextual grounding. BERT and RoBERTa, despite being strong sentence-level classifiers, remain limited when reviews contain mixed or implicit sentiment across aspects (e.g., positive facilities but negative pricing), leading to underperformance in *Value* and *Sleep Quality*. In contrast, HII-KG's integration of KG context provides structured context about traveler types, facilities, and room features, enabling more accurate and consistent aspect predictions across diverse review content.

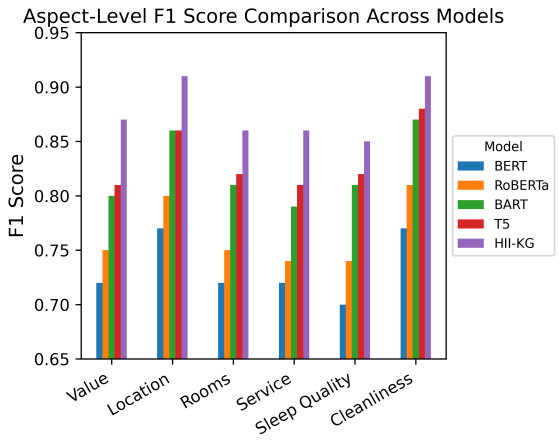

Figure 2: Comparison of aspect-level F1 scores across five models. HII-KG consistently outperforms baselines across all six dimensions.

4.2 Intent Generation Quality

To evaluate generated intent quality, we report BLEU-4, ROUGE-L, and BERTScore (Table 1). HII-KG outperforms GPT-3.5, Gemini 1.5 Flash, BART (fine-tuned), and LLaMA3.1-8B without KG, surpassing the latter by +3.3 BLEU and +2.9 ROUGE-L. The improvements stem from KG-grounded prompts; prepending subgraph-derived context in natural language guides the model toward fluent, specific, and factually consistent outputs. Notably, HII-KG uses only few-shot inference without additional fine-tuning. Prompt examples used during generation are provided in Appendix C.

Model	BLEU-4 ↑	ROUGE-L ↑	BERTScore ↑
GPT-3.5 (prompt only)	0.227	0.386	0.623
Gemini 1.5 Flash	0.248	0.421	0.639
BART (fine-tuned)	0.262	0.432	0.654
LLaMA3.1-8B (no KG)	0.289	0.445	0.669
HII-KG (Ours)	**0.322**	**0.471**	**0.684**

Table 1: Intent generation performance comparison using BLEU-4, ROUGE-L, and BERTScore. Our proposed method (HII-KG) outperforms strong LLM and fine-tuned baselines.

In addition to the observed performance improvements, the results highlight the impact of integrating external knowledge through a structured hotel knowledge graph. GPT-3.5 and Gemini 1.5 Flash, which rely entirely on prompt-based inference without access to external grounding, exhibit comparatively lower performance across all metrics. Although BART and LLaMA3.1-8B demonstrate better adaptability through pretraining or task-specific fine-tuning, their outputs remain less consistent when compared to HII-KG. The superior performance of HII-KG can be attributed to its use of KG contextual prompts, which provide domain-relevant information during inference. This enables the model to generate intent expressions that are not only fluent but also semantically aligned with the user's underlying needs. Furthermore, the improvement in BERTScore (+1.5 over LLaMA3.1-8B) confirms the benefit of incorporating KG context in enhancing factual consistency and relevance. These findings reinforce the value of KG-based grounding for intent inference, especially in low-resource or inference-only settings where full fine-tuning is infeasible.

5 Conclusion

We presented HII-KG, a hierarchical framework that combines structured aspect rating prediction with knowledge-graph–augmented intent generation for hotel review understanding. By leveraging the same Cypher-aware backbone across both stages, our approach captures explicit user signals and implicit motivations in a unified pipeline. Experiments on a TripAdvisor dataset show that HII-KG outperforms strong encoder-based and generative baselines on both aspect-level classification and intent generation metrics. These results demonstrate that even lightweight contextual cues, when aligned with graph-structured priors, shows the significant improvements in interpretability, factual consistency, and downstream user-need inference.

Limitations

Despite strong performance, our work has several limitations. The Hotel-KG is derived from English TripAdvisor reviews with manually curated types, limiting multilingual or domain transfer. Intent evaluation relies mainly on automatic metrics with limited human review. Our approach depends on high-capacity models such as LLaMA3.1-8B-Cypher, which may pose deployment challenges. Finally, while interpretability is improved, biases in both user data and pretrained models remain.

Acknowledgments

This work was supported in part by the National Science and Technology Council, R.O.C. under Grant "NSTC 114-2410-H-002-214-MY2, NSTC 114-2410-H-002-215-MY2, NSTC 114-2410-H-038-034-MY3, NSTC 113-2627-M-A49-002".

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Gaurav Arora, Shreya Jain, and Srujana Merugu. 2024. Intent detection in the age of llms. *arXiv preprint arXiv:2410.01627*.

Anna Bodonhelyi, Efe Bozkir, Shuo Yang, Enkelejda Kasneci, and Gjergji Kasneci. 2024. User intent recognition and satisfaction with large language models: A user study with chatgpt. *arXiv preprint arXiv:2402.02136*.

Yung-Chun Chang, Chih-Hao Ku, and Chien-Hung Chen. 2020. Using deep learning and visual analytics to explore hotel reviews and responses. *Tourism Management*, 80:104129.

Yung-Chun Chang, Chih-Hao Ku, and Chun-Hung Chen. 2019. Social media analytics: Extracting and visualizing hilton hotel ratings and reviews from tripadvisor. *International Journal of Information Management*, 48:263–279.

Liyi Chen, Panrong Tong, Zhongming Jin, Ying Sun, Jieping Ye, and Hui Xiong. 2024. Plan-on-graph: Self-correcting adaptive planning of large language model on knowledge graphs. *Advances in Neural Information Processing Systems*, 37:37665–37691.

Jan Ole Krugmann and Jochen Hartmann. 2024. Sentiment analysis in the age of generative ai. *Customer Needs and Solutions*, 11(1):3.

Chih-Hao Ku, Yung-Chun Chang, and Yichuan Wang. 2024a. How to strategically respond to online hotel reviews: A strategy-aware deep learning approach. *Information & Management*, 61(5):103970.

Chih-Hao Ku, Yung-Chun Chang, and Yichuan Wang. 2024b. How to strategically respond to online hotel reviews: A strategy-aware deep learning approach. *Information & Management*, 61(2).

I-Fan Lin, Faegheh Hasibi, and Suzan Verberne. 2024. Generate then refine: data augmentation for zero-shot intent detection. *arXiv preprint arXiv:2410.01953*.

Chuangtao Ma, Yongrui Chen, Tianxing Wu, Arijit Khan, and Haofen Wang. 2025. Large language models meet knowledge graphs for question answering: Synthesis and opportunities. *arXiv preprint arXiv:2505.20099*.

Charaf Ouaddi, Lamya Benaddi, Lahbib Naimi, Mohamed Rahouti, Abdeslam Jakimi, Rachid Saadane, and 1 others. 2025. Assessing the effectiveness of large language models for intent detection in tourism chatbots: A comparative analysis and performance evaluation. *Scientific African*, 28:e02649.

Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2024. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*, 36(7):3580–3599.

Juan A Rodriguez, Nicholas Botzer, David Vazquez, Christopher Pal, Marco Pedersoli, and Issam Laradji. 2024. Intentgpt: Few-shot intent discovery with large language models. *arXiv preprint arXiv:2411.10670*.

Shijie Wang, Wenqi Fan, Yue Feng, Lin Shanru, Xinyu Ma, Shuaiqiang Wang, and Dawei Yin. 2025. Knowledge graph retrieval-augmented generation for LLM-based recommendation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 27152–27168, Vienna, Austria. Association for Computational Linguistics.

Xinyang Xu, Zeyu Wang, Xinyi Li, Hang Dong, Qingqing Zheng, Yijia Liu, Wayne Xin Zhao, and Ji-Rong Wen. 2024a. Text2cypher: Training llms to generate executable cypher queries over knowledge graphs. *arXiv preprint arXiv:2412.10064*.

Zhentao Xu, Mark Jerome Cruz, Matthew Guevara, Tie Wang, Manasi Deshpande, Xiaofeng Wang, and Zheng Li. 2024b. Retrieval-augmented generation with knowledge graphs for customer service question answering. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, pages 2905–2909.

Dong Zhang and Baozhuang Niu. 2024. Leveraging online reviews for hotel demand forecasting: A deep learning approach. *Information Processing & Management*, 61(1):103527.

Wenqing Zheng, Noah Fatsi, Daniel Barcklow, Dmitri Kalaev, Steven Yao, Owen Reinert, C Bayan Bruss, and Daniele Rosa. 2025. Explain what you mean: Intent augmented knowledge graph recommender built with an llm. *arXiv preprint arXiv:2505.10900*.

A Dataset

The dataset is sampled from a cleaned subset of TripAdvisor hotel reviews. Reviews are filtered to exclude:

- Non-English content,
- Reviews with fewer than 50 characters,
- Incomplete aspect ratings.

Each review is associated with:

- Raw text,
- Six aspect scores (Value, Location, Rooms, Service, Sleep Quality, Cleanliness),
- Reviewer metadata (e.g., traveler type, country),
- Linked KG subgraph (used in Stage 2).

The aspect score distribution is provided in the following Table 2.

Aspect	1	2	3	4	5
Value	45	205	1120	3120	4510
Location	30	150	980	2900	4940
Rooms	60	240	1320	2980	4400
Service	75	310	1085	3050	4480
Sleep Quality	50	190	1250	3105	4405
Cleanliness	25	130	1050	2985	4810

Table 2: Aspect score distribution (scale 1–5) across 10,000 reviews. Most scores are skewed toward positive ratings (4–5).

B Data Annotation

To obtain intent supervision for Stage 2, we manually annotated 1,000 hotel reviews with concise intent statements. Each annotation captures the reviewer's likely booking motivation, rewritten in first-person (e.g., *"I'm looking for a family-friendly hotel with a spacious suite and a pool."*). Two annotators labeled each sample, and disagreements were resolved by a third reviewer. Inter-annotator agreement (Cohen's κ) was 0.84, indicating high consistency.

C Prompt of HII-KG

We use few-shot prompting in both stages of our hierarchical framework. Below, we provide examples of the exact input format used during inference.

Stage 1: Aspect Rating Prediction

The following prompt is used to instruct the model to assign six aspect ratings (Value, Location, Rooms, Service, Sleep Quality, Cleanliness) based on the raw review and associated travel type:

> **Review:** do not go there unless you have a great deal. Terrible noisy location, outrageous parking fees. Rooms overlooking highways. Helpful staff. No easy access to subway.
> **Travel Type:** Family
> **Labels:** [1, 1, 5, 3, 2, 5]
>
> **Review:** Billing Issues Not Easily Resolved. Stay was not up to Hilton standards.
> **Travel Type:** Business
> **Labels:** [3, 2, 3, 2, 1, 3]
>
> **Review:** Out dated. Location is great, but rooms are old and tired.
> **Travel Type:** Solo
> **Labels:** [3, 5, 4, 3, 2, 3]
>
> **Review:** Great atmosphere and nice hotel. Staff very friendly and helpful. Room comfortable and quiet.
> **Travel Type:** Business
> **Labels:** [4, 4, 4, 4, 5, 5]

Query Review:

> **Review:** `{context}`
> **Travel Type:** `{travel_type}`
> **Task:** Assign ratings (1–5) for Value, Location, Rooms, Service, Sleep Quality, Cleanliness. (Higher is greater)
>
> IMPORTANT: Only output in this exact format:
> `Labels: [v, l, r, s, sq, c]`
> Do not output explanations. Do not output labels one by one. Do not include any text other than the label bracket.

Stage 2: Intent Generation with KG Context

Before generation, we extract structured context from a knowledge graph using Cypher queries tailored to the review content. The extraction process consists of three steps:

1. Entity Linking from Review We identify keywords and noun phrases in the review that can be mapped to known KG nodes. For example:

> "The kids loved the pool and the suite was spacious" ⇒ `Facility:pool`, `RoomType:suite`, `TravelType:Family`

2. Cypher Query Construction Using matched entities, we dynamically build Cypher queries to extract relevant subgraphs. For instance:

```
MATCH
    (t:TravelType)-[:PREFERS_ROOMTYPE]->(r:RoomT
ype)-[:HAS_FACILITY]->(f:Facility)
WHERE t.name = "Family" AND f.name = "pool"
RETURN t.name, r.name, f.name
```

3. KG Context Linearization The retrieved subgraph is converted into a natural sentence template. For example:

> *"Family travelers often prefer suites with pool access."*

This KG context is prepended to the review as input to the generation model.

Few-Shot Prompt Format (Stage 2)

We use 3-shot prompting with the following template:

> **KG Context:** Family travelers often prefer suites with pool access and kitchens.
> **Review:** The suite had a small kitchen and my kids really enjoyed the indoor pool. Clean and quiet at night.
> **Task:** Reverse the user intention based on the context and review.
> **Intent:** I am looking for a family-friendly hotel with a suite, kitchen, and pool.
>
> **KG Context:** Business travelers tend to value clean rooms, quiet environments, and fast check-in.
> **Review:** Very smooth check-in process, clean and quiet room. Good for one-night stay before meetings.
> **Task:** Reverse the user intention based on the context and review.
> **Intent:** I need a clean, quiet hotel with quick check-in for a business trip.

Query Prompt:

> **KG Context:** `{kg_summary}`
> **Review:** `{review_text}`
> **Task:** Reverse the user intention based on the context and review.

The model generates an intent statement such as:

> *I'm looking for a family-friendly resort with a large room and a kids' pool.*