# QUIDS: Query Intent Description for Exploratory Search via Dual Space Modeling

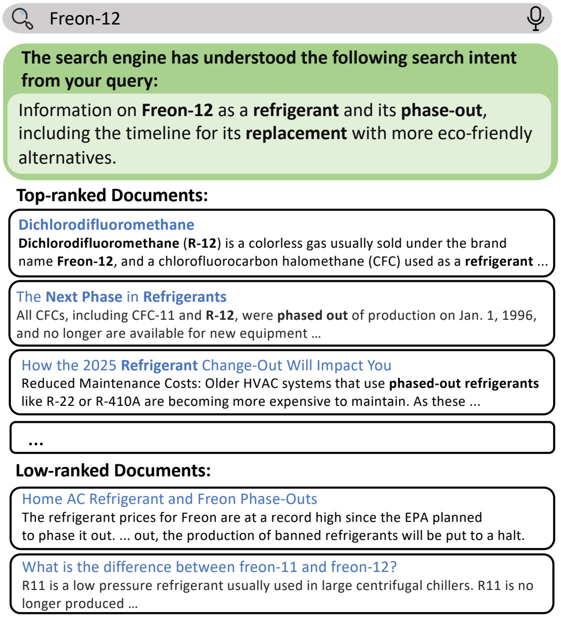
**Yumeng Wang**
Leiden University
Leiden, The Netherlands
y.wang@liacs.leidenuniv.nl

**Xiuying Chen**
MBZUAI
Abu Dhabi, United Arab Emirates
xiuying.chen@mbzuai.ac.ae

**Suzan Verberne**
Leiden University
Leiden, The Netherlands
s.verberne@liacs.leidenuniv.nl

## Abstract

In exploratory search, users often submit vague queries to investigate unfamiliar topics, but receive limited feedback about how the search engine understood their input. This leads to a self-reinforcing cycle of mismatched results and trial-and-error reformulation. To address this, we study the task of generating user-facing natural language query intent descriptions that surface what the system likely inferred the query to mean, based on post-retrieval evidence. We propose QUIDS, a method that leverages dual-space contrastive learning to isolate intent-relevant information while suppressing irrelevant content. QUIDS combines a dual-encoder representation space with a disentangling decoder that works together to produce concise and accurate intent descriptions. Enhanced by intent-driven hard negative sampling, the model significantly outperforms state-of-the-art baselines across ROUGE, BERTScore, and human/LLM evaluations. Our qualitative analysis confirms QUIDS' effectiveness in generating accurate intent descriptions for exploratory search. Our work contributes to improving the interaction between users and search engines by providing feedback to the user in exploratory search settings.[1]



Freon-12

**The search engine has understood the following search intent from your query:**

Information on **Freon-12** as a **refrigerant** and its **phase-out**, including the timeline for its **replacement** with more eco-friendly alternatives.

**Top-ranked Documents:**

**Dichlorodifluoromethane**
**Dichlorodifluoromethane** (**R-12**) is a colorless gas usually sold under the brand name **Freon-12**, and a chlorofluorocarbon halomethane (CFC) used as a **refrigerant** ...

The **Next Phase** in **Refrigerants**
All CFCs, including CFC-11 and **R-12**, were **phased out** of production on Jan. 1, 1996, and no longer are available for new equipment ...

How the 2025 **Refrigerant** Change-Out Will Impact You
Reduced Maintenance Costs: Older HVAC systems that use **phased-out refrigerants** like R-22 or R-410A are becoming more expensive to maintain. As these ...

...

**Low-ranked Documents:**

Home AC Refrigerant and Freon Phase-Outs
The refrigerant prices for Freon are at a record high since the EPA planned to phase it out. ... out, the production of banned refrigerants will be put to a halt.

What is the difference between freon-11 and freon-12?
R11 is a low pressure refrigerant usually used in large centrifugal chillers. R11 is no longer produced ...

Figure 1: A user-facing application of query intent generation in exploratory search. The system's inferred intent is generated by contrasting top-ranked (pseudo-relevant) and low-ranked (pseudo-irrelevant) documents. Key information contributing to the inferred intent is shown in bold.

## 1 Introduction

In exploratory search (Palagi et al., 2017), users often issue vague or underspecified queries to investigate unfamiliar topics through iterative refinement. This process gives rise to a persistent usability challenge, which we call the *dual-blind problem*: Users are uncertain about how to express their information needs; as a result they formulate ambiguous queries; the system silently infers the user's intent based on these ambiguous queries, without providing explicit feedback and retrieves mixed-quality results at best. This leads to a self-reinforcing cycle of the user receiving results that are misaligned with their actual information need. This cycle is difficult to break with traditional query understanding methods (Li et al., 2023; Lu et al., 2024; Wang et al., 2023) that operate in the *pre-retrieval* stage. Their goal is to optimize ranking effectiveness, not to provide feedback to the user. They offer little transparency about how the system arrived at its results, which is especially a problem when users are unsure of their intent.

In this work, we study the task of generating a natural-language query intent description (Zhang et al., 2020) that reflects what the system likely inferred the query to mean. The description is

[1] Our code is available at https://github.com/menauwy/QUIDS

*Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 33062–33077
November 4-9, 2025 ©2025 Association for Computational Linguistics

generated in the *post-retrieval* stage, incorporating system-inferred relevance of documents. These descriptions are not mechanistic explanations of the ranker, but instead serve as user-facing proxies of the system's inferred intent. By contrasting top-ranked (pseudo-relevant) and low-ranked documents, the intent descriptions provide feedback that helps users identify mismatches between their intended and inferred query meanings. This feedback supports more effective query refinement and improves the overall search experience.

Figure 1 illustrates an application scenario of query intent generation in exploratory search. Since the ground truth intent behind the query is unknown to the search system, it relies on the retrieved documents to infer the query intent. From the top-ranked documents (considered relevant by the search engine), key terms like 'Freon-12', 'refrigerant', and 'phased out' are captured and emphasized in the intent description. In contrast, topics such as 'costs' and 'prices', which appear in both high-ranked and low-ranked documents, are excluded from the final intent description. The resulting description provides diagnostic feedback on how the system understood the query, helping users assess whether the retrieved results align with their latent intent.

We introduced a novel dual-space modeling approach, QUIDS, for the query intent generation task. It models query intent through dual-space contrastive learning by performing contrastive learning in two complementary spaces, explicitly separating intent-relevant and irrelevant semantic information. Specifically, the method consists of: (i) a *representation space* via dual encoders and (ii) a novel *disentangling space* in the decoder. This dual-space design enables the model to subtract irrelevant semantics from relevant ones, generating concise and more accurate intent descriptions. Furthermore, we propose an intent-driven hard negative sampling strategy to expand the irrelevant representation space and improve contrastive learning during training.

Experiments on the Q2ID benchmark (Zhang et al., 2020), including TREC and SemEval datasets, show that our model significantly outperforms strong baselines, including the prior Q2ID-specific method, LLM-based, and Query-focused Summarization methods, both in automatic and human evaluations. Qualitative analysis confirms that QUIDS effectively filters out distracting or misleading content and generates concise intent descriptions. Our contributions are: (i) Our model generates high-quality intent descriptions, with performance significantly enhanced by incorporating hard negative data augmentation during training. (ii) We introduce contrastive learning in both the representation space and the disentangling space of transformer models, effectively capturing contrasting information from relevant and irrelevant documents. (iii) We perform a thorough evaluation of our model, providing us with insights into the model's strengths and weaknesses, as well as its potential application scenarios, especially for exploratory search.

## 2 Related Work

### 2.1 Query Understanding

Our work is related to traditional query understanding tasks such as classification (Broder, 2002; Verberne et al., 2013), clustering (Wen et al., 2002; Hong et al., 2016), and expansion (Wang et al., 2023; Mo et al., 2023; Jagerman et al., 2023). However, unlike these methods, which operate in the *pre-retrieval* stage to optimize retrieval effectiveness, Zhang et al. (2020) proposes the Query-to-Intent-Description (Q2ID) task in the *post-retrieval* stage that aims to generate search systems' inferred intent of a user query based on both relevant and irrelevant documents. Unlike their method, we directly model a query-aware irrelevant intent space via dual-space contrastive learning, and enhance the performance with hard negative data augmentation, leading to a more precise intent description.

### 2.2 Query-focused Summarization

In settings where annotated intent descriptions are available, a related task to Q2ID is query-focused summarization (QFS) (Vig et al., 2022; Pagnoni et al., 2023). QFS is a subtask of text summarization that aims to generate a summary of one or multiple documents, guided by a query. Traditional methods rely on unsupervised extraction, ranking text segments by similarity and query relevance (Wan and Xiao, 2009; Feigenblat et al., 2017). Recent QFS datasets (Kulkarni et al., 2020; Fabbri et al., 2022; Zhong et al., 2021) have enabled the rise of QA-driven approaches (Su et al., 2020, 2021). More advanced techniques model query relevance through evidence ranking (Xu and Lapata, 2021), latent query optimization (Xu and Lapata, 2022), or pipeline architectures like the coarse-to-fine model in (Xu and Lapata, 2020). To handle long documents, extract-then-abstract strategies (Vig et al., 2022) use sparse attention and segment scoring. Other innovations include question-driven pretraining (Pagnoni et al., 2023), contrastive learning (Sotudeh and Goharian, 2023), and joint token-utterance modeling with query-aware attention (Liu et al., 2023a).

We use QFS models as baselines for the Q2ID task, but there is a fundamental difference between QFS and Q2ID: QFS aims to compress the content of retrieved documents to help users consume information, whereas Q2ID aims to generate a description of what the system likely inferred about the query intent, based on retrieval results. We provide a comparison table with related tasks in Appendix A.

# 3 Methods

## 3.1 Pipeline Framework

We define the contrastive intent generation task as follows. Given a dataset $\mathcal{D} = \{(q, \mathcal{R}, \mathcal{I}, y)_j\}$ with $L$ samples, where $j \in \{0, 1, ..., L\}$: $q$ is a query, $\mathcal{R} = \{r_1, r_2, ..., r_{|\mathcal{R}|}\}$ is a collection of relevant documents for the query, $\mathcal{I} = \{i_1, i_2, ..., i_{|\mathcal{I}|}\}$ is a collection of irrelevant documents and $y$ is the human-annotated ground truth query intent. The modeling goal is to learn the distinctions between relevant and irrelevant inputs based on a query, while generating a system-inferred intent description that exclusively highlights the relevant aspects related to the query. To achieve this, our training pipeline consists of 2 steps: (1) *Intent-Driven Negative Augmentation (IDNA)* and (2) *Dual Space Modeling (DualSM)*.

## 3.2 Intent-Driven Negative Augmentation (IDNA)

The purpose of IDNA is to mine hard negative documents as irrelevant documents from the entire dataset $D$ based on the query, its relevant document collections, and the ground truth intent, i.e.,

$$IDNA(q, \mathcal{R}, y, \mathcal{D}) = \mathcal{I}'$$

where $\mathcal{I}' = \{i'_1, i'_2, ..., i'_h\}$ with $h$ the expected number of irrelevant documents. Inspired by Liu et al. (2022) on choosing in-context sample strategies for in-context learning, we design a method to choose intent-aware hard negative samples based on semantic similarity. Specifically, we use a Sentence Transformer model (Reimers and Gurevych, 2019) to represent both positive and negative samples from the training data in a vector space. Then we choose negative samples close to the positive ones in this space. The negative sample augmentation makes the task more challenging for the model, hence improving its discriminative capabilities. As shown in Figure 2 (a), we augment hard negative samples for each training query using Algorithm 1. For encoding, we use a Sentence Transformers model pre-trained on the MSMARCO Passage Ranking dataset (Nguyen et al., 2016). In practice, we set 0.8 to the similarity threshold with an analysis of its effect in subsection 5.4.3.

**Algorithm 1** Intent-Driven Negative Augmentation (IDNA)

```
Require: query $q$, relevant document collection $\mathcal{R}$, irrelevant
    document collection $\mathcal{I}$, whole dataset document corpus
    $\mathcal{D}$, target size $S$, threshold $\tau$
Ensure: augmented irrelevant documents $\mathcal{I}'$ for $q$
$h_{q;y} \leftarrow \text{Encode}(\text{Concatenate}(q; y))$
$\mathcal{R}^* \leftarrow$ Sort $\mathcal{R}$ descending by $\cos(h_{q;y}, \text{Encode}(r))$,
    $\forall r \in \mathcal{R}$
$h_{R^*} \leftarrow \text{Encode}(\text{Concatenate}(\mathcal{R}^*))$
$\mathcal{I}^* \leftarrow$ Sort $\mathcal{I}$ descending by $\cos(h_{R^*}, \text{Encode}(I))$,
    $\forall i \in \mathcal{I}$
$\mathcal{I}' \leftarrow$ top-$|\mathcal{I}^*|$ ranked docs
while $|\mathcal{I}'| < S$ do
    Sample a document $d$ from $\mathcal{D}$
    if $\cos(h_{R^*}, \text{Encode}(d)) > \tau$ then
        Add $d$ to $\mathcal{I}'$
    end if
end while
```

## 3.3 Dual Space Modeling (DualSM)

Dual Space Modeling aims to contrastively generate a descriptive intent for the query by modeling query-aware relevant and irrelevant intent spaces:

$$DualSM(q, \mathcal{R}, i') = \hat{y}$$

We use a Transformer-based encoder-decoder architecture, with the BART-large model (Lewis et al., 2020) and the T5-large model (Raffel et al., 2020a) as backbones, as illustrated in Figure 2(b). To capture the relationship between a query and its relevant and irrelevant documents, we implement a Siamese dual encoder architecture. Based on the encoder outputs, contrastive learning is performed in the representation space to differentiate between embeddings for relevant and irrelevant documents. Correspondingly, we design a contrastive decoder to model query-aware relevant and irrelevant intent spaces in a disentangled manner.

### 3.3.1 Representation Space Modeling (RSM)

We design a dual cross-encoder architecture to distinguish relevant and irrelevant documents for

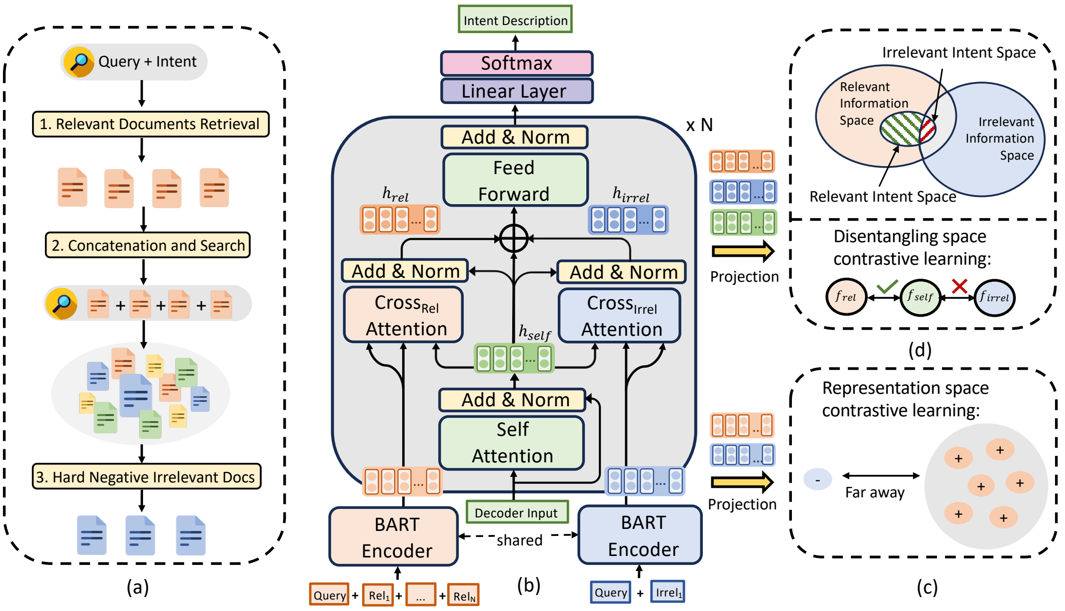


Figure 2: Overview of our proposed pipeline. From left to right, we show (a) Intent-Driven Negative Augmentation method, (b) Contrastive decoder structure with dual cross-attention layers, (c) and (d) Contrastive learning via dual space modeling.

a given query by jointly encoding each query-document pair. Relevant documents, which often share similar topics tied to the query's intent, are concatenated based on their ranking (Section 3.2, Step 1) and encoded together. In contrast, irrelevant documents can be irrelevant to a query in diverse ways, making it impractical to model a meaningful and comprehensive irrelevant feature representation space. Therefore, we focus on a single irrelevant document $i'$ at each training step, using a hard negative sample from the augmented irrelevant document collection $\mathcal{I}'$.

To model the feature space, we project the encoder's final hidden states through a linear layer. Document embeddings are obtained via average pooling over token representations. We optimize the representation space by pulling relevant embeddings closer and pushing the irrelevant one away (Figure 2(c)). The objective is to minimize:

$$\mathcal{L}_{rel} = \sum_{m=1}^{k} \sum_{n=m+1}^{k} d(e_m, e_n) \qquad (1)$$

where $e$ is the embedding of each relevant document, $k$ is the number of relevant documents, and $d$ is a distance function. We use cosine distance for $d$ in this work. For irrelevant feature representation space, we optimize the margin loss function:

$$\mathcal{L}_{irrel} = \sum_{m=1}^{k} max(t - d(e_m, \bar{e}), 0) \qquad (2)$$

where $\bar{e}$ is the embedding of the irrelevant document, and $t$ is a margin parameter, set to 1 in our case. We combine the relevant and irrelevant loss to obtain the encoder contrastive loss as follows:

$$\mathcal{L}_{encoder} = \mathcal{L}_{rel} + \mathcal{L}_{irrel} \qquad (3)$$

### 3.3.2 Disentangling Space Modeling (DSM)

In decoding, we aim to generate intent descriptions based on the encoded relevant and irrelevant document features. To achieve this, we design a contrastive decoder with an added cross-attention layer that attends to both sources. To further disentangle relevant from irrelevant information, we apply contrastive learning in a separate disentangling space. As shown in Figure 2(d), this helps the model focus on relevant intent while minimizing influence from irrelevant content, enabling more precise and nuanced intent generation.

Our decoder adopts a Transformer architecture, composed of N identical decoder layers. In the $l$-th decoder layer, at the $z$-th decoding step, we obtain hidden states $h^l_{self,z}$ by employing masked self-attention layers, to make sure the prediction of position $z$ depends only on the predictions before $z$. Based on $h^l_{self,z}$, we compute relevant document hidden states $h^l_{rel,z}$ by applying multi-head attention with cross-attention (MHAtt) to relevant encoder output:

$$h^l_{rel,z} = MHAtt(h^l_{self,z}, h_{R^*}) \qquad (4)$$

Similarly, we get the irrelevant document hidden states by attending to irrelevant encoder output:

$$h^l_{irrel,z} = MHAtt(h^l_{self,z}, h_{i'}) \tag{5}$$

From preliminary results, we found that a simple linear combination of $h^l_{self,z}$, $h^l_{rel,z}$, and $h^l_{irrel,z}$ works well to serve as the decoder hidden state to produce the distribution over the target vocabulary:

$$h^l_{combine,z} = h^l_{self,z} + h^l_{rel,z} - h^l_{irrel,z} \tag{6}$$

$$P^{vocab}_z = Softmax(W(h^N_{combine,z})) \tag{7}$$

where $W$ indicates a linear transformation. We optimize the model with the negative log likelihood (NLL) objective to predict the target words:

$$\mathcal{L}_{NLL} = -\sum_{z=1}^{|y|} log P^{vocab}_z(y_z) \tag{8}$$

Corresponding to the representation space contrastive learning, we perform another contrastive learning in the newly proposed disentangling space using hidden states from the last decoder layer. We apply an additional linear layer to $h^N_{self}$, $h^N_{rel}$, and $h^N_{irrel}$, projecting them into a new representation space. We then obtain the embeddings $f_c$, $f_r$, $f_{i'}$ by pooling these projected vector representations.

We follow the approach of SimCLR (Chen et al., 2020) and use from-batch negative samples $\mathcal{B}$ in the InfoNCE loss (He et al., 2020):

$$\mathcal{L}_{decoder} = -log \frac{exp(cos(f_c, f_r)/\tau)}{\sum_{i' \in \mathcal{B}} exp(cos(f_c, f_{i'})/\tau)} \tag{9}$$

where $\tau$ is the temperature and $cos(\cdot, \cdot)$ defines cosine similarity.

Finally, we combine the original NLL loss together with encoder and decoder loss to obtain the overall loss $\mathcal{L}$ to update all learnable parameters in an end-to-end learning setting:

$$\mathcal{L}_{NLL} = \lambda_0 \mathcal{L}_{NLL} + \lambda_1 \mathcal{L}_{Encoder} + \lambda_2 \mathcal{L}_{Decoder} \tag{10}$$

where the $\lambda$ parameters control the balance between the three losses, with their total sum equal to 1.

# 4 Experimental Settings

## 4.1 Data

We conduct experiments on the **Q2ID** dataset (Zhang et al., 2020), a benchmark for query-to-intent description derived from existing TREC and SemEval collections. Specifically, it comprises: **TREC**: Including the Dynamic Domain tracks (2015–2017) and the 2004 Robust Track, which focus on dynamic, exploratory search and consistency of retrieval technology. **SemEval**: Including the English SemEval-2015 and SemEval-2016 Task 3 tracks on Community Question Answering. Q2ID contains a total of 5,358 entries. Each entry is structured as a quadruple: <*query, relevant documents, irrelevant documents, intent description*>, where the intent descriptions are human-written narratives. The statistics and more details are provided in Appendix B.1.

## 4.2 Baselines

To reflect the shared focus on user queries and the extraction of relevant content, we compare our model with baselines from four categories: (i) *Pretrained Seq2Seq Models*: We fine-tune **T5-large** (Raffel et al., 2020b) and **BART-large** (Lewis et al., 2020) on the Q2ID dataset. BART also serves as the backbone of our QUIDS model. (ii) *Q2ID Baseline*: **CtrsGen** (Zhang et al., 2020) leverages contrastive generation using a bi-GRU encoder and contrast-weighted attention mechanism. (iii) *LLM Baseline*: We evaluate a instruction-tuned model **LLaMA3.1-8B-Instruct** (AI, 2024) and a reasoning model **OpenAI o3** (OpenAI, 2025) in zero-shot and two-shot settings. For the two-shot settings, examples are drawn from TREC and SemEval. (iv) *QFS Baselines*: We include extractive-abstractive models **RelReg**, **RelRegTT** (Vig et al., 2022), the segment-based model **SegEnc** (Vig et al., 2022), the question-driven **Socratic** (Pagnoni et al., 2023), and the contrast-enhanced **Qontsum** (Sotudeh and Goharian, 2023). Detailed descriptions, training configurations, and reproduction settings for baselines are in Appendix B.4 and B.5.

## 4.3 Implementation Details

Our method is implemented based on the BART-large model (Lewis et al., 2020) using the Huggingface Transformers library (Wolf et al., 2019). Cross$_{\text{Rel}}$ and Cross$_{\text{Irrel}}$ attention layers in the decoder are initialized with pre-trained BART weights. We optimize the weighted training loss using coefficients ($\lambda_0 = 0.2, \lambda_1 = 0.2, \lambda_2 = 0.6$) to balance multiple objectives (see Appendix B.2). The model is trained with the Adam optimizer, and the final checkpoint is selected based on average ROUGE-{1, 2, L} scores on the validation set. We provide additional training details in Appendix B.3.

### 4.4 Evaluation Metrics

We conduct three types of evaluations using different evaluator resources: automatic evaluation, LLM-based evaluation and human evaluation. For automatic evaluation, we report recall scores on ROUGE-{1, 2, L} following Zhang et al. (2020), along with BERTScore (Zhang et al., 2019), which assesses semantic and syntactic similarity beyond exact word matches. We also conduct a human evaluation study using 50 (Sotudeh and Goharian, 2023) randomly selected test samples. Five PhD students in Computer Science scored intent descriptions from our model and the best baseline, without knowing which model produced them. They rated both models on four customized qualitative criteria with scores ranging from 1 (worst) to 5 (best). Four criteria are: (1) **Fluency**: to what extent the generated query intent description reads naturally, understandably, and without noticeable errors or disruptions. (2) **Factual Alignment**: to what extent the generated query intent description is factually aligned with the ground truth intent. (3) **Inclusion score**: how well the generated query intent includes important details from the query and relevant documents. (4) **Exclusion score**: how well the generated query intent description excludes information present in the irrelevant documents that is not relevant to the query and relevant documents. Inspired by Liu et al. (2023b), we also adopt LLM-based evaluation (LLaMa3.1-8B-Instruct and GPT-4o) by prompting instruction-tuned models to assess generations across the same four qualitative metrics. Details on the prompt formats, and scoring computation are provided in Appendix F.

| **Models** | **RG-1** | **RG-2** | **RG-L** | **BS** |
|---|---|---|---|---|
| CtrsGen$^\dagger$ | 24.76 | 4.62 | 20.21 | - |
| T5-large | 28.87 | 13.91 | 23.85 | 61.64 |
| BART-large | 30.70 | 13.91 | 24.63 | 62.07 |
| LLaMa3.1 (0) | 29.28 | 7.42 | 20.90 | 57.26 |
| LLaMa3.1 (2) | 32.75 | 9.54 | 24.34 | 57.89 |
| o3 (0) | **34.91** | 6.84 | 23.93 | 59.19 |
| o3 (2) | 33.15 | 6.28 | 22.81 | 59.88 |
| RelReg | 26.67 | 12.83 | 21.99 | 59.24 |
| RelRegTT | 27.21 | 12.77 | 22.25 | 59.60 |
| SegEnc | 31.83 | 14.29 | 25.18 | 62.15 |
| + SOCRATIC Pret. | 31.38 | 13.88 | 24.91 | 62.26 |
| QONTSUM | 31.18 | 14.26 | 24.87 | 62.03 |
| QUIDS_T5 | 29.40 | 13.95 | 24.23 | 62.00 |
| QUIDS_BART | 34.47 | **14.86**$^*$ | **26.77**$^*$ | **63.55**$^*$ |

Table 1: Performance between our model and baselines in terms of automatic evaluation (%). $^\dagger$ indicates reported performance from previous work. '-' means the result is inaccessible. $^*$ indicates the model outperforms the best baseline significantly with paired t-test at $p$-value $<$ 0.05 level. Results are averaged over 5 random seeds. The best results are highlighted in bold, while the best baseline results are underlined.

| **Model** | **RG-1** | **RG-2** | **RG-L** | **BS** |
|---|---|---|---|---|
| QUIDS w/o IDNA | 33.48 | 14.20 | 25.95 | 63.17 |
| QUIDS w/o RSM | 34.57 | 14.39 | 26.38 | 63.62 |
| QUIDS w/o DSM | 33.45 | 13.46 | 25.88 | 63.33 |
| QUIDS | **35.95** | **14.80** | **27.21** | **64.33** |

Table 2: Ablation study of our QUIDS model with its variants under automatic evaluation (%).

## 5 Experimental Results

### 5.1 Overall Results

We compare model performance between QUIDS and baselines in Table 1. The results show that: (1) QUIDS outperforms all baselines except o3 on RG-1, including those with larger model sizes, indicating that the gains stem primarily from our dual-space modeling design rather than model size or pretraining knowledge; (2) Our approach is compatible with both T5 and BART architectures. Notably, BART-large outperforms T5-large despite having nearly half the model size; (3) The QFS models that we implemented for the query intent generation task outperform the Q2ID-specific baseline, CtrsGen. QUIDS further significantly outperforms the best QFS model, SegEnc; (4) The two-shot setting with the LLaMa3.1-8B-Instruct model significantly outperforms the zero-shot setting in ROUGE scores, while showing only minor improvements in BERTScore. This suggests that without fine-tuning, generated intents may be lexically similar but semantically misaligned; (5) While o3 shows strong reasoning and generates richer descriptions, it does not outperform our model on ROUGE and BERTScore. We observed that o3 often paraphrases intent with different wording, leading to lower ROUGE scores. Its longer outputs and inclusion of detailed justifications or summaries may dilute the concise intent signal. This suggests that strong reasoning alone may not align well with the goal of generating concise, system-inferred intent.

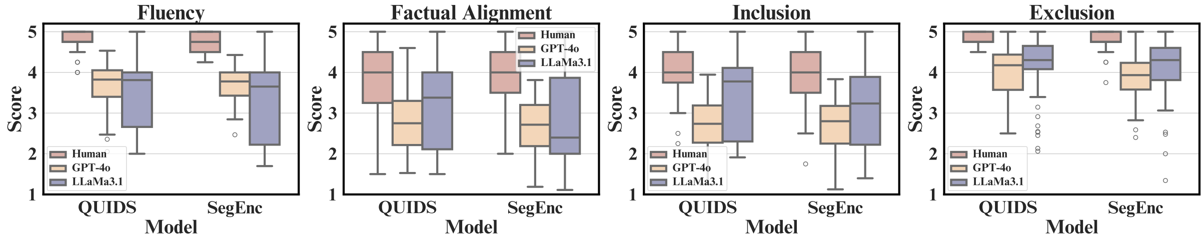


Figure 3: Distribution of human and LLM evaluation sores on four qualitative metrics.

## 5.2 Ablation Study

We perform an ablation study based on the BART-Large model to evaluate the contribution of key components in our approach under three settings: (1) without the IDNA module (w/o IDNA), (2) without contrastive learning in the encoder (w/o RSM), and (3) without contrastive learning in the decoder (w/o DSM). Results are shown in Table 2. Excluding contrastive learning from the decoder leads to the largest performance drop, underscoring its role in modeling a discriminative intent space. Removing it from the encoder results in a smaller decline, suggesting that representation space modeling still contributes to relevance awareness.

## 5.3 Human and LLM Evaluation

We assess the quality of generated intents from QUIDS and the best baseline SegEnc using both human and LLM-based evaluations. Inter-annotator agreement, measured by weighted Cohen's $\kappa$, and LLM-human correlations, measured by Spearman and Kendall $\tau$, indicate fair to moderate consistency across metrics (Table 7, Appendix C). As shown in Table 3, QUIDS outperforms SegEnc on all metrics except Factual Alignment, where humans prefer SegEnc. Further analysis (subsection 5.4.1) shows this stems from the dominance of informational queries, on which SegEnc performs better. QUIDS, by contrast, performs better on exploratory queries. Figure 3 further illustrates score distributions, revealing three key insights: (1) Human scores are generally higher than LLM scores, especially for Fluency and Exclusion. The larger variability for fluency scores suggests humans may tolerate minor fluency issues. (2) Human evaluations show broader and lower score distributions for Factual Alignment and Inclusion, aligning more with LLaMa3.1 (Table 7). In contrast, they mirror GPT-4o's narrower distribution for Fluency and Exclusion, where correlation is higher. This suggests that evaluators differ in how they assess each metric. (3) Fluency and Factual Alignment show stronger alignment between LLM and human evaluations, likely due to being less context-dependent. In contrast, Inclusion and Exclusion scores exhibit weaker correlations, indicating inconsistencies in evaluating context-sensitive criteria.

| Method | Model | Fluen. | Align. | Inclu. | Exclu. |
|---|---|---|---|---|---|
| Human | SegEnc | 4.75 | **3.90** | 3.94 | 4.77 |
| | QUIDS | **4.80** | 3.80 | **4.06** | **4.80** |
| GPT-4o | SegEnc | 3.70 | 2.64 | 2.67 | 3.83 |
| | QUIDS | **3.71** | **2.79** | **2.69** | **4.00** |
| LLaMa3.1 | SegEnc | 3.25 | 2.91 | 3.17 | 4.07 |
| | QUIDS | **3.48** | **3.17** | **3.42** | **4.11** |

Table 3: Comparison of human evaluation and LLM evalaution in terms of Fluency, Factual Alignment, Inclusion score and Exclusion score.

| Intent | RG-1 | RG-2 | RG-L | BS |
|---|---|---|---|---|
| Informational | 35.69 | 14.51 | 26.82 | 63.88 |
| Exploratory | **41.55** | **23.24** | **38.28** | **76.65** |

Table 4: Comparison of automatic evaluation on our model for different intent types.

## 5.4 In-depth Analysis

### 5.4.1 Analysis of Intent Types

We classify the queries according to their underlying search intent into two categories: (1) **Informational Intent**: Natural language questions seeking detailed information or solutions, typically longer and contextual. Queries from the SemEval dataset fall under this category. (2) **Exploratory Intent**: Term-based queries aimed at broad exploration with minimal context or structure. Queries from the TREC datasets are categorized here. Our automatic and human evaluation by intent type indicates that QUIDS is indeed more successful for exploratory tasks than for informational intent tasks. Results are shown in Table 4 and Figure 4. See Appendix D for detailed analysis.

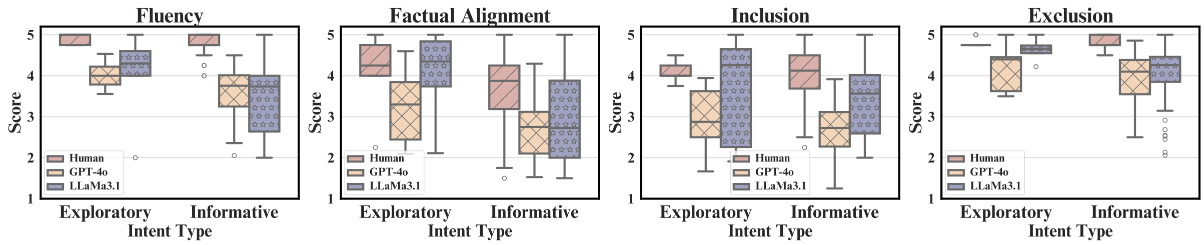
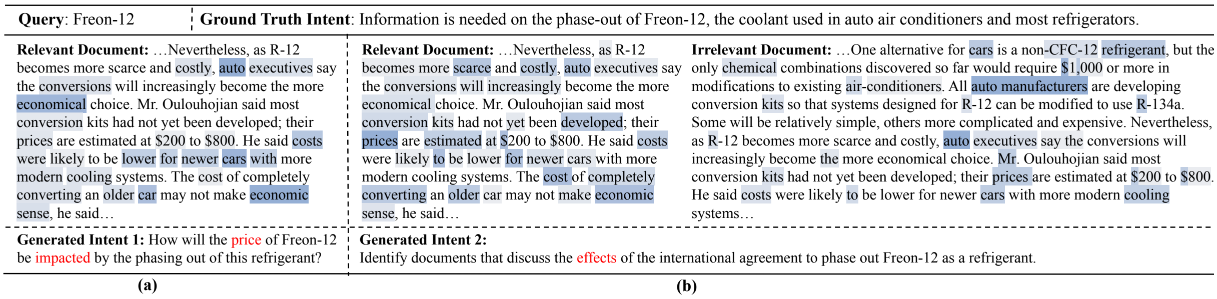


Figure 4: Boxplot of human and LLM evaluation scores on 4 metrics of our model on different intent types.



**Query**: Freon-12 | **Ground Truth Intent**: Information is needed on the phase-out of Freon-12, the coolant used in auto air conditioners and most refrigerators.

**Relevant Document:** …Nevertheless, as R-12 becomes more scarce and costly, auto executives say the conversions will increasingly become the more economical choice. Mr. Oulouhojian said most conversion kits had not yet been developed; their prices are estimated at $200 to $800. He said costs were likely to be lower for newer cars with more modern cooling systems. The cost of completely converting an older car may not make economic sense, he said…

**Generated Intent 1:** How will the price of Freon-12 be impacted by the phasing out of this refrigerant?

(a)

**Relevant Document:** …Nevertheless, as R-12 becomes more scarce and costly, auto executives say the conversions will increasingly become the more economical choice. Mr. Oulouhojian said most conversion kits had not yet been developed; their prices are estimated at $200 to $800. He said costs were likely to be lower for newer cars with more modern cooling systems. The cost of completely converting an older car may not make economic sense, he said…

**Irrelevant Document:** …One alternative for cars is a non-CFC-12 refrigerant, but the only chemical combinations discovered so far would require $1,000 or more in modifications to existing air-conditioners. All auto manufacturers are developing conversion kits so that systems designed for R-12 can be modified to use R-134a. Some will be relatively simple, others more complicated and expensive. Nevertheless, as R-12 becomes more scarce and costly, auto executives say the conversions will increasingly become the more economical choice. Mr. Oulouhojian said most conversion kits had not yet been developed; their prices are estimated at $200 to $800. He said costs were likely to be lower for newer cars with more modern cooling systems…

**Generated Intent 2:** Identify documents that discuss the effects of the international agreement to phase out Freon-12 as a refrigerant.

(b)

Figure 5: Case study indicating the role of contrastive examples in the decoder stage. Token-level decoder cross-attention weights are shown for a generated intent token (red) are shown with (a) and without (b) an irrelevant document in the model input. Deeper color indicates a higher value.

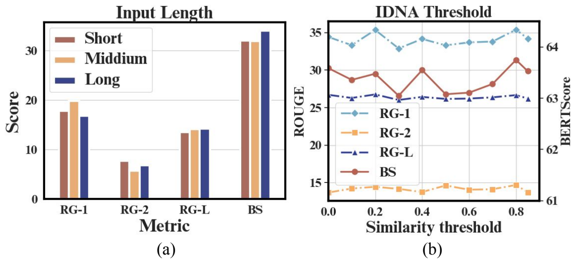


Figure 6: Robustness analysis on (a) input length and (b) IDNA threshold.

### 5.4.2 Analysis of Document Length

In real-world scenarios, relevant documents associated with a query can vary in length. Documents over 1024 tokens are truncated in our architecture. To assess robustness to input length, we group test documents into short (<512 tokens), medium (512–1024), and long (>1024). In the test dataset, 204 are short, 25 medium, and 29 long. Figure 6(a) shows that the overall differences in performance among all the input length categories are minimal across all metrics, suggesting that 1024-token inputs are sufficient for inferring the underlying query intent. This may explain why long-document QFS models offer limited gains in this task.

### 5.4.3 Sensitivity Analysis of IDNA threshold

Intent-driven negative augmentation (IDNA) selects irrelevant documents with high similarity to relevant ones as hard negatives for contrastive learning. We noticed that most document similarity scores fall between 0.6 and 0.9, making the method robust even with lower thresholds (e.g., <0.6). Performance improves as the threshold increases within this range, with 0.8 yielding the best results, which we use in our experiments.

## 5.5 Case Study

Figure 5 illustrates how the model uses cross-attention in the decoder stage to identify irrelevant semantics from a low-ranked document. When generating ‘impacted’ without an irrelevant document (Figure 5(a)), the model focuses on economic effects on cars, indicated by ‘price’ in Intent 1. With an irrelevant document in Figure 5(b), while similar economic attentions are observed across both relevant and irrelevant documents when generating the word ‘effects’ in Intent 2, the model successfully identifies tokens related to prices and cars in relevant documents as irrelevant. This demonstrates the model’s ability to filter out irrelevant content using contrastive learning in the decoder. We include another failure case study in Appendix E.

# 6 Conclusions

We introduced a novel dual-space modeling approach for the query intent generation task. Our approach implements contrastive learning in both encoding and decoding phases, combined with intent-driven hard negative augmentation during data preprocessing, to automatically generate detailed and precise intent descriptions, surfacing what the system likely inferred the query to mean. Experimental results show that our model can effectively filter out irrelevant information from the relevant intent space, leading to more accurate intent descriptions than all baselines, including models for Query-Focused Summarization. In future work, we plan to improve contextual understanding in distinguishing relevant from irrelevant information and extend our approach to conversational search by mining exploratory needs and explaining the understanding of query intents. Our long-term aim is to improve the transparency in the retrieval process, in particular for exploratory search needs.

## Acknowledgments

This publication is part of the project LESSEN with project number NWA.1389.20.183 of the research program NWA ORC 2020/21 which is (partly) financed by the Dutch Research Council (NWO).

## Limitations

**Training Efficiency Trade-offs.** Augmenting irrelevant documents enhances robustness but linearly increases training time. We therefore limited negative documents to three per query, which partially alleviates this trade-off but remains suboptimal. To fundamentally resolve this efficiency bottleneck, two promising directions are: (1) adaptive dynamic sampling that prioritizes high-impact negatives through real-time gradient analysis, and (2) curriculum-based augmentation progressively introducing harder negatives as training stabilizes.

**Dataset Imbalance.** Informational queries dominate the training data over exploratory ones. While our model shows promising performance in exploratory search scenarios, this bias limits deeper intent analysis. Future work should expand out experiments to more datasets, focusing on exploratory queries. One option would be to use LLM-generated synthetic data, specifically creating pseudo-documents that mimic multi-faceted exploratory intents. This approach maintains intent modeling consistency while enabling systematic investigation of query complexity, without requiring manual annotation efforts.

## References

Meta AI. 2024. Introducing llama 3.1: Our most capable models to date. https://ai.meta.com/blog/meta-llama-3-1/. Accessed: October 9, 2024.

Andrei Broder. 2002. A taxonomy of web search. In *ACM Sigir forum*, volume 36, pages 3–10. ACM New York, NY, USA.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Richard Csaky and Gábor Recski. 2020. The gutenberg dialogue dataset. *arXiv preprint arXiv:2004.12752*.

Alexander Fabbri, Xiaojian Wu, Srini Iyer, Haoran Li, and Mona Diab. 2022. AnswerSumm: A manually-curated dataset and pipeline for answer summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2508–2520, Seattle, United States. Association for Computational Linguistics.

Guy Feigenblat, Haggai Roitman, Odellia Boni, and David Konopnicki. 2017. Unsupervised query-focused multi-document summarization using the cross entropy method. In *Proceedings of the 40th International ACM SIGIR Conference on research and development in information retrieval*, pages 961–964.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738.

Yuan Hong, Jaideep Vaidya, Haibing Lu, and Wen Ming Liu. 2016. Accurate and efficient query clustering via top ranked search results. In *Web Intelligence*, volume 14, pages 119–138. IOS Press.

Rolf Jagerman, Honglei Zhuang, Zhen Qin, Xuanhui Wang, and Michael Bendersky. 2023. Query expansion by prompting large language models. *arXiv preprint arXiv:2305.03653*.

Sayali Kulkarni, Sheide Chammas, Wan Zhu, Fei Sha, and Eugene Ie. 2020. Aquamuse: Automatically generating datasets for query-based multi-document summarization. *arXiv preprint arXiv:2010.12694*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Juanhui Li, Wei Zeng, Suqi Cheng, Yao Ma, Jiliang Tang, Shuaiqiang Wang, and Dawei Yin. 2023. Graph enhanced bert for query understanding. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3315–3319.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.

Xingxian Liu, Bin Duan, Bo Xiao, and Yajing Xu. 2023a. Query-utterance attention with joint modeuing for query-focused meeting summarization. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Chang Lu, Liuqing Li, Donghyun Kim, Xinyue Wang, and Rao Shen. 2024. An effective, efficient, and stable framework for query clustering. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*, pages 5334–5340. IEEE.

Fengran Mo, Kelong Mao, Yutao Zhu, Yihong Wu, Kaiyu Huang, and Jian-Yun Nie. 2023. Convgqr: generative query reformulation for conversational search. *arXiv preprint arXiv:2305.15645*.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human-generated machine reading comprehension dataset.

OpenAI. 2025. Introducing openai o3 and o4-mini. `https://openai.com/index/introducing-o3-and-o4-mini/`. Accessed: 2025-09-14.

Artidoro Pagnoni, Alex Fabbri, Wojciech Kryscinski, and Chien-Sheng Wu. 2023. Socratic pretraining: Question-driven pretraining for controllable summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12737–12755, Toronto, Canada. Association for Computational Linguistics.

Emilie Palagi, Fabien Gandon, Alain Giboin, and Raphaël Troncy. 2017. A survey of definitions and models of exploratory search. In *Proceedings of the 2017 ACM workshop on exploratory search and interactive data analytics*, pages 3–8.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020a. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020b. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

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*. Association for Computational Linguistics.

Sajad Sotudeh and Nazli Goharian. 2023. Qontsum: On contrasting salient content for query-focused summarization. *arXiv preprint arXiv:2307.07586*.

Dan Su, Yan Xu, Tiezheng Yu, Farhad Bin Siddique, Elham Barezi, and Pascale Fung. 2020. CAiRE-COVID: A question answering and query-focused multi-document summarization system for COVID-19 scholarly information management. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online. Association for Computational Linguistics.

Dan Su, Tiezheng Yu, and Pascale Fung. 2021. Improve query focused abstractive summarization by incorporating answer relevance. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3124–3131.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Suzan Verberne, Maarten van der Heijden, Max Hinne, Maya Sappelli, Saskia Koldijk, Eduard Hoenkamp, and Wessel Kraaij. 2013. Reliability and validity of query intent assessments. *Journal of the American Society for Information Science and Technology*, 64(11):2224–2237.

Jesse Vig, Alexander Richard Fabbri, Wojciech Kryściński, Chien-Sheng Wu, and Wenhao Liu. 2022. Exploring neural models for query-focused summarization. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1455–1468.

Xiaojun Wan and Jianguo Xiao. 2009. Graph-based multi-modality learning for topic-focused multi-document summarization. In *Twenty-First International Joint Conference on Artificial Intelligence*. Citeseer.

Liang Wang, Nan Yang, and Furu Wei. 2023. Query2doc: Query expansion with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9414–9423.

Ji-Rong Wen, Jian-Yun Nie, and Hong-Jiang Zhang. 2002. Query clustering using user logs. *ACM Transactions on Information Systems*, 20(1):59–81.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and 1 others. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Yumo Xu and Mirella Lapata. 2020. Query focused multi-document summarization with distant supervision. *arXiv preprint arXiv:2004.03027*.

Yumo Xu and Mirella Lapata. 2021. Generating query focused summaries from query-free resources. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6096–6109.

Yumo Xu and Mirella Lapata. 2022. Document summarization with latent queries. *Transactions of the Association for Computational Linguistics*, 10:623–638.

Ruqing Zhang, Jiafeng Guo, Yixing Fan, Yanyan Lan, and Xueqi Cheng. 2020. Query understanding via intent description generation. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 1823–1832.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and 1 others. 2021. Qmsum: A new benchmark for query-based multi-domain meeting summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5905–5921.

# A Detailed Comparison with Related Work

A detailed comparison between our task—Query intent description generation —and related tasks such as Query Understanding (QU), Query-Focused Summarization (QFS), and Pseudo-Relevance Feedback (PRF) is provided in Table 6.

# B Experimental Setting Details

## B.1 Dataset Details

In constructing the Q2ID dataset, documents with multi-graded relevance labels were converted into binary labels, indicating whether each document is relevant to the query. The dataset is composed of:

- 510 entries from TREC tracks (Dynamic Domain 2015–2017, Robust 2004)
- 4,878 entries from SemEval-2015/2016 Task 3 on Community Question Answering

Each data point is formatted as a quadruple: <*query, relevant documents, irrelevant documents, intent description*>.

The average query length is 7.2 tokens, and the average intent description length is 45.5 tokens. We follow the original split of Q2ID: 5,000 queries for training, 100 for validation, and 258 for testing.

## B.2 Sensitivity Analysis of Loss Weights

We conduct a sensitivity analysis through a grid search over different combinations of the loss weights ($\lambda_0$, $\lambda_1$, $\lambda_2$). The results of several representative configurations are summarized in Table 5. We observed that configurations assigning relatively higher weight to $\lambda_2$ tend to yield better overall performance. Overly large $\lambda_2$ values (e.g., 0.8) degrade performance by reducing the contribution of the other loss components. Configurations that assign more weight to $\lambda_0$ or $\lambda_1$ alone lead to inferior performance, sometimes resulting in no valid generation, and are therefore omitted for clarity. Based on these observations, we adopt $(\lambda_0, \lambda_1, \lambda_2) = (0.2, 0.2, 0.6)$ as the default setting.

| $\lambda_0$ | $\lambda_1$ | $\lambda_2$ | RG-1 | RG-2 | RG-3 | BS |
|---|---|---|---|---|---|---|
| - | - | - | **34.53** | 14.42 | 26.85 | 63.33 |
| 0.33 | 0.33 | 0.33 | 33.36 | 14.64 | 26.22 | 63.45 |
| 0.20 | 0.20 | 0.60 | 34.51 | **15.36** | **27.13** | **63.52** |
| 0.30 | 0.30 | 0.40 | 34.17 | 14.31 | 26.30 | 63.41 |
| 0.10 | 0.10 | 0.80 | 32.37 | 14.59 | 25.81 | 62.87 |
| 0.20 | 0.40 | 0.40 | 33.20 | 14.56 | 26.10 | 63.22 |

Table 5: Sensitivity analysis of loss weights. '-' indicates that the $\lambda$ parameters were treated as learnable during training.

## B.3 Implementation

To balance efficiency and effectiveness during IDNA augmentation, we set the expected number

| Task | Query Understanding (Classification / Clustering / Expansion) | Query-Focused Summarization (QFS) | Pseudo-Relevance Feedback (PRF) | Our Task: Query Intent Generation (QIG) |
|---|---|---|---|---|
| **Goal** | Predict query intent classes, discover latent topics, or expand queries for better retrieval performance. | Summarize relevant documents to help users consume content. | Refine or reformulate queries to improve retrieval performance. | Generate a natural language description of the search system's inferred intent behind a query. |
| **Output Form** | Labels (e.g., informational/navigational), clusters, or expanded query terms. | Natural language summary (abstractive or extractive). | Modified query or reweighted terms. | Natural language explanation of inferred query intent. |
| **Use of Irrelevant Documents** | Not used. Focus is on query-only or top-ranked documents. | Rarely used; mainly uses pseudo-relevant documents. | Not used; PRF assumes top-ranked documents are relevant. | Explicitly contrasts relevant and irrelevant documents for intent disentanglement. |
| **Application Stage** | Pre-retrieval; typically before document scoring. | Post-retrieval summarization. | Interleaved or pre-retrieval (used for re-ranking or expansion). | Post-retrieval; supporting user query refinement and retrieval debugging |
| **User Utility** | Improves ranking accuracy and personalization; not visible to users. | Helps users consume content more efficiently. | Improves recall or relevance through backend query rewriting. | Helps users understand potential mismatches between their intended query meaning and the system's inferred intent. |

Table 6: Comparison between Query Understanding (QU) tasks, Query-Focused Summarization (QFS), Pseudo-Relevance Feedback (PRF), and our Query Intent Generation (QIG) task.

of irrelevant documents per query to three for training. This results in augmenting 1,984 queries with at least three irrelevant documents each. Training is conducted for 10 epochs using the Adam optimizer with a learning rate of 0.0001. During decoding, we set a maximum sequence length of 256 tokens and apply beam search with a beam size of 4. We also set a no-repeat n-gram size of 3 to reduce redundancy.

### B.4 Baselines

*Pretrained sequence-to-sequence model baselines*: **T5** (Raffel et al., 2020b): a Transformer-based encoder-decoder (Vaswani et al., 2017) model trained on a diverse and extensive dataset. We use a pretrained T5-large model that we finetuned on the original Q2ID training dataset. **BART** (Lewis et al., 2020): also a transformer-based encoder-decoder model, trained by corrupting documents and then optimizing a reconstruction loss. The BART model serves as the backbone of our QUIDS model.

*Query-to-intent description (Q2ID) baseline*: **CtrsGen** (Zhang et al., 2020): a Q2ID model using a bi-directional GRU as encoder architecture. During decoding, it computes contrast scores by considering irrelevant documents to adjust sentence-level attention weights in the relevant documents.

*Large Language Model (LLM) baseline*: **LLaMa3.1-8B-Instruct** (AI, 2024): We evaluate the LLaMa3.1-8B instruction-tuned text-only model under both zero-shot and two-shot settings and conduct five experimental runs for each setting. For two-shot setting, we randomly using two different examples per run – one sourced from TREC and the other from SemEval. **OpenAI o3** (OpenAI, 2025): We evaluate the reasoning-focused OpenAI o3 model. While the model internally generates reasoning tokens, we focus our analysis on the visible output tokens. The same experimental settings and number of runs are used for comparison.

*Query Focused Summarization (QFS) baselines*: **RelReg** (Vig et al., 2022) and **RelRegTT** (Vig et al., 2022): two-step approaches for QFS consisting of an score-and-rank extractor and an abstractor. The extractor is trained to predict ROUGE relevance scores and then the ranked results based on ROUGE are passed to the abstractor. **SegEnc** (Vig et al., 2022): an end-to-end approach tailored for handling longer input texts. SegEnc splits a long input into fixed-length overlapping segments and encodes them separately. The encoding sequences are concatenated so that the decoder can attend to all encoded segments jointly. **Socratic** (Pagnoni et al., 2023): an unsupervised, question-driven pretraining approach designed to tailor generic language models for controllable summarization tasks. **Qontsum** (Sotudeh and Goharian, 2023): an abstractive summarizer that applied Generative Information Retrieval (GIR) techniques. It builds on SegEnc by adding a segment scorer and contrastive learning modules. We train the QFS baselines on the Q2ID dataset using the original code provided by the authors, except for Qontsum, which we independently reproduced.

### B.5 Implementation Details for Baselines

RelReg and RelRegTT share the same abstractor, a BART-large model, which also serves as the backbone model for SegEnc, Socratic, and Qontsum. For RelReg and RelRegTT, we use an input segment length of 1024, whereas SegEnc-based models utilize an input segment length of 512, with a total input length of 4096. For Socratic training, we use the checkpoint pretrained on Books3 (Csaky and Recski, 2020) from the Huggingface Model Hub[2] and fine-tune it on Q2ID dataset using SegEnc mechanism. We reproduce the work of Qontsum with the segment length of 512 tokens, temperature of 0.6 and ($\lambda_0 = 0.6, \lambda_1 = 0.2, \lambda_2 = 0.2$) in joint learning. For all models that divide input text into segments, we apply a 50% overlap between each segment and its adjacent one.

## C Correlation with Human Evaluation

We assess the correlation between human and LLM evaluators across four qualitative evaluation criteria, presenting the Spearman ($\rho$) and Kendall-Tau ($\tau$) correlations for the best SOTA model SegEnc and our QUIDS model in Table 7. Overall, our model demonstrates significantly higher human correspondence across all metrics compared to SegEnc, with the exception of the Exclusion score. Correlation performance varies by metric; for Fluency and Factual Alignment—criteria requiring less contextual information—there is a relatively higher degree of agreement with human evaluations. In contrast, the Inclusion and Exclusion scores, which depend on diverse and contextual sources, show lower correlation, suggesting that humans and LLM evaluators adopt different evaluation strategies for more complex criteria. Additionally, we observe that different LLM evaluators exhibit human-like evaluation behaviors across various metrics: LLaMa3.1 shows greater human correspondence in Factual Alignment and Inclusion scores, whereas GPT-4o aligns more closely with human evaluations in Fluency and Exclusion scores.

## D Evaluation on Intent Types

In Table 3, we observe a human preference over the SegEnc model on metric Factual Alignment, which measures how well the generated query intent description is factually aligned with the ground truth intent. We guess it is due to the model performance difference on different sub-datasets, or on different intent types. And hence we further analyse the evaluations on different intent types.

**Automatic Evaluation** In the 258 test samples, there are 20 queries with exploratory intents and 238 with informational intents. As shown in Table 4, queries with exploratory intents substantially outperform those with informational intents, achieving 60% higher ROUGE-2 scores and 20% higher BERTScores. This indicates that our model is better suited for exploratory queries. This finding contrasts with the results of (Zhang et al., 2020), where the CtrsGen model performed slightly better on the informational SemEval queries than on the exploratory TREC queries. A potential explanation is that the backbone language model used in our approach more freely generates text than the GRU model used in (Zhang et al., 2020), particularly when reconstructing complex scenarios for informational intents. This is an aspect that makes our approach more suitable for exploratory search rather than informative search.

**Human and LLM Evaluation** Table 8 presents human and LLM evaluations on our model regarding two intent types. In general, exploratory intents consistently outperform informational intents across all metrics. This finding, derived from 50 test samples, aligns with automatic evaluation results on the full test dataset (Table 4). Figure 4 shows the evaluation score distribution by intent type on our model, compared to the overall model performance in Figure 3. The informational intent distribution closely mirrors the overall performance, suggesting that informational queries dominate the dataset and largely influence performance. However, exploratory queries, despite being less frequent, demonstrate superior perfor-

[2] https://huggingface.co/Salesforce/socratic-books-30M

| Correlation | Model | Fluency |  | Alignment |  | Inclusion |  | Exclusion |  | Average |  |
|---|---|---|---|---|---|---|---|---|---|---|---|
|  |  | $\rho$ | $\tau$ | $\rho$ | $\tau$ | $\rho$ | $\tau$ | $\rho$ | $\tau$ | $\rho$ | $\tau$ |
| Corr (Human, GPT-4o) | SegEnc | 0.320 | 0.252 | 0.387 | 0.295 | 0.158 | 0.121 | **0.375** | **0.309** | 0.310 | 0.244 |
|  | QUIDS | **0.476** | **0.375** | 0.495 | 0.386 | 0.265 | 0.199 | 0.361 | 0.286 | **0.399** | **0.312** |
| Corr (Human, LLaMa3.1) | SegEnc | 0.224 | 0.173 | 0.434 | 0.329 | 0.153 | 0.116 | 0.343 | 0.272 | 0.289 | 0.223 |
|  | QUIDS | 0.373 | 0.294 | **0.557** | **0.424** | **0.365** | **0.262** | 0.158 | 0.126 | **0.363** | **0.276** |

Table 7: Spearman ($\rho$) and Kendall-Tau ($\tau$) correlations between Human evaluation and LLM evaluation of different metrics.

| Method | Intent | Fluen. | Align. | Inclu. | Exclu. |
|---|---|---|---|---|---|
| Human | Info. | 4.78 | 3.71 | 4.04 | 4.80 |
|  | Expl. | **4.89** | **4.19** | **4.14** | **4.81** |
| GPT-4o | Info. | 3.64 | 2.70 | 2.63 | 3.95 |
|  | Expl. | **4.02** | **3.17** | **2.96** | **4.20** |
| LLaMa3.1 | Info. | 3.33 | 2.96 | 3.35 | 3.99 |
|  | Expl. | **3.72** | **4.11** | **3.72** | **4.64** |

Table 8: Comparison of Human and LLM evaluation on informational and exploratory intent types on our model.

mance in this task. When diving into the factual alignment in Figure 3, while humans prefer our model for exploratory intent with 4.19 (QUIDS) vs. 3.94 (SegEnc), SegEnc is favored for informative intent with 3.71 (QUIDS) vs. 3.89 (SegEnc). Since informative intent queries dominate, this leads to a lower average score for our model on this metric. These findings indicate that our model is well-suited for exploratory search.

## E Failure Case Study

Figure 7 illustrates a failure example of filtering irrelevant information when an irrelevant document is provided. The token-level decoder cross-attention weights are compared when generating a content word in the intent, with (c) and without (d) an irrelevant document. When generating the keyword 'UK' and 'Dubai', the model mainly focuses on 'petrol' and 'cigarettes' in the relevant documents for both (c) and (d), which are also contextually important in the generated intent. However, the model fails to recognize the relationship between 'middle east' in the irrelevance document and 'Dubai', leading to the unwanted inclusion of 'Dubai' in the intent 2. This highlights that our model may struggle with excluding information that requires commonsense reasoning or domain-specific knowledge. A direction for future work is to develop advanced approaches that enhance contextual understanding for complex scenarios.

## F LLM-based Evaluation Details

Following the method of (Liu et al., 2023b), we use LLaMa3.1-8B-Instruct[3] and GPT-4o[4] as instruction-tuned evaluators to assess the generated intent across four qualitative metrics. Specifically, we define the evaluation task and criteria, prompting the LLM to generate chain-of-thoughts (CoT) for the 'Evaluation Steps'. For LLaMa3.1-8B-Instruct, we use the output token probabilities from the LLMs to normalize the scores and take their weighted summation as the final results:

$$score = \sum_{i=1}^{n} p(s_i) \times s_i \tag{11}$$

where $S = \{s_1, s_2, ..., s_n\}$ represents the predefined score set from the prompt, with a maximum value of 5 in our case. For the close-sourced GPT-4o, we sample 20 times to estimate the token probabilities. An example prompt for each model is presented below.

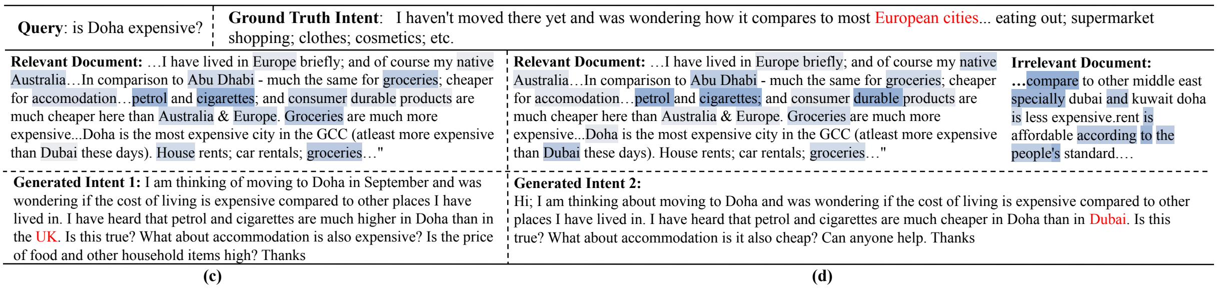
### F.1 General Evaluation Prompt

*You will be given a query, relevant and irrelevant documents with respect to the query. You will also be given a generated query intent description based on the query and documents. The ground truth query intent description will also be provided.*
*Your task is to rate the query intent description on one metric.*
*Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.*

***Evaluation Criteria:***
«MetricCriteria»

[3] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct.

[4] https://openai.com/index/hello-gpt-4o/.



**Query**: is Doha expensive?

**Ground Truth Intent**: I haven't moved there yet and was wondering how it compares to most European cities... eating out; supermarket shopping; clothes; cosmetics; etc.

**Relevant Document:** …I have lived in Europe briefly; and of course my native Australia…In comparison to Abu Dhabi - much the same for groceries; cheaper for accomodation…petrol and cigarettes; and consumer durable products are much cheaper here than Australia & Europe. Groceries are much more expensive...Doha is the most expensive city in the GCC (atleast more expensive than Dubai these days). House rents; car rentals; groceries…"

**Generated Intent 1:** I am thinking of moving to Doha in September and was wondering if the cost of living is expensive compared to other places I have lived in. I have heard that petrol and cigarettes are much higher in Doha than in the UK. Is this true? What about accommodation is also expensive? Is the price of food and other household items high? Thanks

**(c)**

**Relevant Document:** …I have lived in Europe briefly; and of course my native Australia…In comparison to Abu Dhabi - much the same for groceries; cheaper for accomodation…petrol and cigarettes; and consumer durable products are much cheaper here than Australia & Europe. Groceries are much more expensive...Doha is the most expensive city in the GCC (atleast more expensive than Dubai these days). House rents; car rentals; groceries…"

**Irrelevant Document:** **…**compare to other middle east specially dubai and kuwait doha is less expensive.rent is affordable according to the people's standard.…

**Generated Intent 2:** Hi; I am thinking about moving to Doha and was wondering if the cost of living is expensive compared to other places I have lived in. I have heard that petrol and cigarettes are much cheaper in Doha than in Dubai. Is this true? What about accommodation is it also cheap? Can anyone help. Thanks

**(d)**

Figure 7: Failure case study. Token-level decoder cross-attention weights for a generated intent token (red) are shown with (c) and without (d) an irrelevant document. Deeper color indicates a higher value.

***Evaluation steps:***
«EvaluationSteps»

*Query:*

`{{Query}}`

*Relevant documents:*

`{{Relevant documents}}`

*Irrelevant documents:*

`{{Irrelevant documents}}`

*Generated Intent:*

`{{Generated intent}}`

*Ground Truth Intent:*

`{{Gound truth intent}}`

***Evaluation Form (scores ONLY):***
*- «MetricName»:*

### F.2 Evaluation Prompt on Fluency

***Evaluation Criteria:***
*Fluency Score (1-5) - This metric measures if the generated query intent description reads naturally, understandably, and without noticeable errors or disruptions.*

***Evaluation steps:***
*1. Carefully review the provided query, relevant, and irrelevant documents to understand the context and content.*
*2. Read the ground truth query intent description to understand the ideal response. This serves as a benchmark for evaluating the fluency of the generated description.*
*3. Carefully read the generated query intent description. Focus on the fluency aspect, considering factors such as grammatical correctness, naturalness, clarity, coherence, and readability.*
*Assign a rating from 1 to 5 based on the level of fluency.*

***Evaluation Form (scores ONLY):***
*- Fluency:*

### F.3 Evaluation Prompt on Factual Alignment

***Evaluation Criteria:***
*Factual Alignment (1-5) - This metric measures if the generated query intent description is factually aligned with the ground truth intent. Ensuring the facts presented in the generated description are correct and match those in the ground truth description. Verifying that all key facts and points mentioned in the ground truth are covered in the generated description without omission. Any hallucination that diverges from the ground truth should be flagged.*

***Evaluation steps:***
*1. Review the ground truth intent description for the central facts and points that convey the query's purpose.*
*2. Read the generated intent description and list the main facts and points it conveys.*
*3. Compare the lists from the ground truth and generated descriptions for consistency in content. Look for alignment in terms of content, completeness, and accuracy.*
*4. Identify any key facts or points from the ground truth that are missing in the generated description (omissions) and note any information in the generated description that is not present or diverges from the ground truth (hallucinations).*
*Assign a rating from 1 to 5 based on the level of factual alignment.*

***Evaluation Form (scores ONLY):***
*- Factual Alignment:*

### F.4 Evaluation Prompt on Inclusion Score

***Evaluation Criteria:***

*Inclusion Score (1-5) - This metric measures how well the generated query intent includes important details from the query and relevant documents. Assessing whether the generated description captures key elements that are directly relevant to the query. Evaluating if the generated description thoroughly includes significant points from the relevant documents. Ensuring that the included details are integrated in a way that maintains the context and importance as presented in the relevant documents.*

***Evaluation steps:***

*1. Review the query and relevant documents to extract the main facts, significant points, and key elements that directly address the query.*
*2. Read the generated query intent description and list the key details it includes.*
*3. Compare the key details and elements from the generated description with those identified from the query and relevant documents, checking for inclusion and alignment.*
*4. Assess how well the included details are integrated into the generated description, ensuring they maintain the context and importance as presented in the relevant documents.*
*Assign a rating from 1 to 5 based on the thoroughness and relevance of the included details.*

***Evaluation Form (scores ONLY):***
*- Inclusion Score:*

### F.5 Evaluation Prompt on Exclusion Score

***Evaluation Criteria:***

*Exclusion Score (1-5) - This metric measures if the generated query intent description excludes information present in the irrelevant documents that is not relevant to the query and relevant documents. Evaluating whether the description effectively filters out information that is irrelevant to the query. Ensuring that the description does not include misleading or incorrect information found in the irrelevant documents. Evaluating whether the description effectively filters out information present in the irrelevant documents but focus on topics different from those in relevant documents.*

***Evaluation steps:***

*1. Carefully read through the irrelevant documents to pinpoint details, facts, or topics that are not relevant to the query and relevant documents.*
*2. Read the generated query intent description and extract the key details and points included in the description.*
*3. Compare the extracted content from the generated description with the irrelevant information identified in the irrelevant documents to check for the presence of any irrelevant details.*
*4. Assess how effectively the generated description filters out irrelevant information, ensuring it focuses only on the query and relevant documents.*
*Assign a rating from 1 to 5 based on the level of exclusion of irrelevant details.*

***Evaluation Form (scores ONLY):***
*- Exclusion Score:*