

QUIDS: Query Intent Generation via Dual Space Modeling

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

Xiuying Chen
Mohamed bin Zayed University of
Artificial Intelligence
Abu Dhabi, United Arab Emirates
xiuying.chen@mbzuai.ac.ae

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

Abstract

Query understanding is a crucial component of Information Retrieval (IR), aimed at identifying the underlying search intent of textual queries. However, most existing approaches oversimplify this task into query classification or clustering, which fails to fully capture the nuanced intent behind the query. In this paper, we address the task of query intent generation: to automatically generate detailed and precise intent descriptions for search queries using relevant and irrelevant documents given a query. These intent descriptions can help users understand why the search engine considered the top-ranked documents relevant, and provide more transparency to the retrieval process. We propose a dual-space model that uses semantic relevance and irrelevance information in the returned documents to explain the understanding of the query intent. Specifically, in the encoding process, we project, separate, and distinguish relevant and irrelevant documents in the *representation space*. Then, we introduce a semantic decoupling model in the novel *disentangling space*, where the semantics of irrelevant information are removed from the relevant space, ensuring that only the essential and relevant intent is captured. This process refines the understanding of the query and provides more accurate explanations for the search results. Experiments on benchmark data demonstrate that our methods produce high-quality query intent descriptions, outperforming existing methods for this task, as well as state-of-the-art query-based summarization methods. A token-level visualization of attention scores reveals that our model effectively reduces the focus on irrelevant intent topics. Our findings open up promising research and application directions for query intent generation, particularly in exploratory search.

CCS Concepts

- Information systems → Users and interactive retrieval; Query intent;
- Computing methodologies → Natural language generation.

Keywords

Information Retrieval, Query intent generation, Contrastive learning, Transformers

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY'

© 2018 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-XXXX-X/18/06
<https://doi.org/XXXXXXX.XXXXXXX>

ACM Reference Format:

Yumeng Wang, Xiuying Chen, and Suzan Verberne. 2018. QUIDS: Query Intent Generation via Dual Space Modeling. In *Proceedings of (Conference acronym 'XX)*. ACM, New York, NY, USA, 13 pages. <https://doi.org/XXXXXXX.XXXXXXX>

1 Introduction

Query understanding is a focal point in the Information Retrieval (IR) community, aimed at determining the underlying search intent of a textual query input. This facet of IR is crucial for several reasons. Firstly, it enables search engines to distinguish and filter out unwanted interpretations of the current search query, ensuring relevance and accuracy in results. This is important because queries tend to be short and underspecified and are therefore often ambiguous [13]. Secondly, users benefit when search engines provide feedback, such as query expansions or more descriptive query intents [6]. This enhances the search experience by enabling users to refine and adjust their queries based on insights derived from the provided query understanding. Existing methods for query understanding can be broadly categorized into three groups. Query classification methods [3, 39] aim to categorize queries into distinct classes based on intent taxonomies. However, many of these methods tend to abstract intents into broader categories, lacking granularity in their analysis. Second, query clustering methods [15, 42] aim to group query intents into clusters, yet encounter challenges in precisely interpreting individual queries within these clusters. The third approach, which is the most effective way to capture query intent, is query intent description generation, where the intent is described in textual format. However, this task has been largely overlooked in existing work, with the most recent contribution by [47], which still relies on RNN-based methods and does not leverage the advancements in transformer models.

Figure 1 illustrates an example of query intent generation for the query ‘Freon-12’, which has an exploratory intent – aimed at exploring topics without a clearly defined goal [27]. Since the true intent behind the query is unknown to the search engine, it relies on the retrieved documents to infer it. Thus, it treats the top-ranked documents as pseudo-relevant and the low-ranked documents as irrelevant. The intent description generated from both sets of documents is highlighted with a green background, while the information directly contributing to intent formation is marked in boldface. From the relevant documents, 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 relevant and irrelevant documents, are deprioritized and excluded from the final intent description. The resulting intent generation illustrates how the search engine

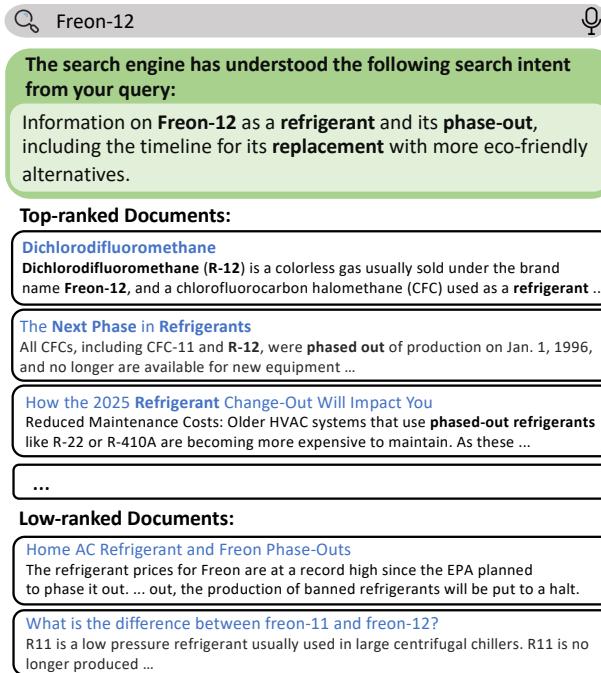


Figure 1: An application of the query intent generation task as a form of explanation to the user in exploratory search. The top-ranked documents are used as pseudo-relevant documents to generate the intent. Information relevant to forming the intent is highlighted in boldface.

interprets the user’s query based on the retrieved documents. Furthermore, this approach allows users to reflect on their interests more clearly, potentially guiding them to reformulate their query more precisely and achieve more targeted search results. This iterative process of query refinement helps bridge the gap between exploratory search and more specific information needs, enhancing the overall search experience.

In scenarios where the ground truth intent is available, a related task to train models for generating query intent is query-focused summarization (QFS). This task is to generate summaries tailored to specific user queries or information needs, drawing from pseudo-relevant (high-ranked) documents. However, the focus in QFS is on summarizing content rather than explicitly identifying or reproducing query intents. Typically, QFS does not involve irrelevant information, which means the resulting summaries may include unfiltered or irrelevant details that are not truly reflective of the query’s intent. On the other hand, previous work [47] on the Q2ID task incorporates irrelevant documents for contrast, but it does not establish a direct connection between the query and irrelevant documents. Our approach explicitly models both relevant and irrelevant intent spaces in a query-aware manner. Specifically, we construct the final intent space by subtracting the irrelevant intent space from the relevant one. This subtraction ensures that the final intent description is closely aligned with the query while effectively filtering out irrelevant information.

We propose a novel pipeline, implementing contrastive learning through dual space modeling in transformer-based [38] models, particularly T5 [29] and BART [17]. Specifically, we use a dual-encoder architecture and perform contrastive learning in the representation space using encoder embeddings of both relevant and irrelevant documents. During the decoding phase, we adapt the transformer decoder to conduct contrastive learning using decoder hidden states in the newly proposed disentangling space, where we model query-aware relevant and irrelevant intent spaces. This dual-space contrastive learning approach ensures that the model learns to differentiate between relevant and irrelevant documents, while also guiding it to produce high-quality intent descriptions with attention to the learned relevant and irrelevant intent space. Moreover, we introduce a method to augment hard negative documents driven by intent, to expand the irrelevant information representation space during training.

We evaluate our model using the benchmark dataset Q2ID, sourced from TREC and SemEval, which includes two types of query intent. Our results demonstrate that our model outperforms state-of-the-art baselines in terms of ROUGE metrics and BERTScore. Additionally, we perform human evaluations and LLM-based evaluations using tailored metrics, and provide a comprehensive analysis of our approach. To further illustrate the effect of irrelevant documents, we conduct a case study in which we highlight the information used to generate a semantically rich intent token based on the decoder’s cross-attention weights at the token level for both relevant and irrelevant documents. In summary, our contributions are as follows:

- 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.
- Our model generates high-quality intent descriptions, with performance significantly enhanced by incorporating hard negative data augmentation during training.
- 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.

2 Related Work

2.1 Query Understanding

The most closely related tasks in query understanding include query classification, query clustering, and query expansion. Query classification [13, 18] aims to classify queries into one or more pre-defined target taxonomies, based on factors like intent [3, 34], topic [2, 19], information type [28] or other classification dimensions. However, due to the typically short and ambiguous nature of search queries, as well as their evolving meanings, these taxonomies may lack sufficient granularity and become outdated over time [13]. This can lead to the loss of detailed information in the classification process. Query clustering attempts to discover search topics or intents by mining clusters of queries. The similarity between queries is typically measured using data from user click-through logs [23, 31, 42] or top-ranked documents [15]. In a recent application, query clustering has been employed in search engines to display real-time trending user queries, capturing shifts in queries and news articles

over consecutive timestamps [23]. Earlier studies on question answering demonstrate that an effective way to expand a query is to extract answer patterns or select terms that could be possible answers as expansion terms. Recently, some generation-augmented retrieval methods [5, 24] focus on exploiting the knowledge captured in pre-trained language models (PLMs) [4, 33] to generate the potential answer as expansion terms.

Inspired by benchmark datasets containing detailed descriptions provided by human annotators, Zhang et al. [47] propose the Query-to-Intent-Description (Q2ID) task that aims to generate a natural language intent description based on both relevant and irrelevant documents of a given query. Unlike their method, we directly model a query-aware irrelevant intent space in transformer models, leading to a more precise intent description.

2.2 Query-Focused Summarization

Query-focused summarization (QFS) is a subtask within text summarization that aims to generate a summary of a given text based on a specific query. Traditionally, the predominant approaches are unsupervised extraction methods [11, 41] that generate summaries by ranking and extracting textual segments according to their similarity to other segments and their relevance to the query. In the past years, numerous QFS datasets [10, 16, 49] have been introduced, primarily based on answer selection or summaries for specific queries, leading to the development of many recent QA-motivated methods [9, 36, 37]. Other works focus on query modeling through approaches such as the learning evidence ranking model [45], jointly optimizing a latent query model and a conditional language model [46], or designing pipeline models for abstractive summarizations. For instance, [44] introduce a coarse-to-fine pipeline consisting of three estimators (relevance, evidence, and centrality estimators) that gradually discard less relevant segments.

Effectively extracting key insights from long documents has become a critical challenge. Vig et al. [40] evaluate a two-step extract-then-abstract approach, where a scoring model is trained to select relevant portions of the text based on a query, which are then used as input to an abstraction model to generate a final summary. Additionally, they propose end-to-end methods that handle longer input texts by employing sparse attention mechanisms. Additional approaches focus on the use of a question-driven pretraining objective [26], contrastive learning based on the scored segments [35], and joint modeling of token and utterance based on query-utterance attention [21]. In this work, we use state-of-the-art QFS methods as baselines for our proposed model.

3 Methods

In this section, we present our approach QUIDS that generates descriptive query intents via dual space modeling. Figure 2 shows an overview. We first introduce the pipeline framework in section 3.1, then the data augmentation method to select intent-driven hard negative samples in section 3.2, and finally the contrastive mechanism via dual space modeling in section 3.3.

3.1 Pipeline Framework

We define the contrastive generation task as follows. Given a dataset $\mathcal{D} = \{(q, \mathcal{R}, \mathcal{I}, y)_j\}$ with L samples, where $j \in \{0, 1, \dots, L\}$: q is a

query, $\mathcal{R} = \{r_1, r_2, \dots, r_{|\mathcal{R}|}\}$ is a collection of relevant documents for the query, $\mathcal{I} = \{i_1, i_2, \dots, i_{|\mathcal{I}|}\}$ is a collection of irrelevant documents and y is the ground truth query intent (i.e. the expected output). The goal is to learn the distinctions between relevant and irrelevant inputs based on a query while generating an 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)*: mining 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 $\mathcal{I}' = \{i'_1, i'_2, \dots, i'_h\}$ with h the expected number of irrelevant documents. (2) *Dual Space Modeling (DSM)*: contrastively generating a descriptive intent for the query by modeling query-aware relevant and irrelevant intent spaces, i.e.,

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

3.2 Intent-Driven Negative Augmentation

Inspired by Liu et al. [20] 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 [32] 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 using the following steps: For each query q in the training set,

- (1) We rank its relevant documents \mathcal{R} based on their cosine similarity to the concatenated embedding of the query and its ground truth intent $(q; y)$ where $(:)$ denotes concatenation, resulting in \mathcal{R}^* .
- (2) We concatenate the ranked relevant documents \mathcal{R}^* into a single document R^* and obtain its embedding h_{R^*} .
- (3) The original irrelevant documents are initially ranked based on their cosine similarities to the newly concatenated relevant embedding h_{R^*} .
- (4) If the number of irrelevant documents is smaller than an expected size, the irrelevant document collection \mathcal{I}' is augmented by selecting hard negative samples from all other documents in the whole dataset \mathcal{D} according to their cosine similarities to the new embedding h_{R^*} .

For step 1, 3 and 4, we use a pre-trained Sentence Transformers model [32] trained on the MSMARCO Passage Ranking Dataset [25] to encode the documents. In practice, during step 4, we select documents whose cosine similarities exceed a predefined threshold 0.8 until the expected size is reached.

3.3 Dual Space Modeling

We use a Transformer-based encoder-decoder architecture, with the BART-large model [17] and the T5-large model [29] as the backbone, as illustrated in Figure 2 (b). To capture the relationship between a query and its relevant and irrelevant documents, we adopt a Siamese dual encoder architecture. Based on the encoder outputs,

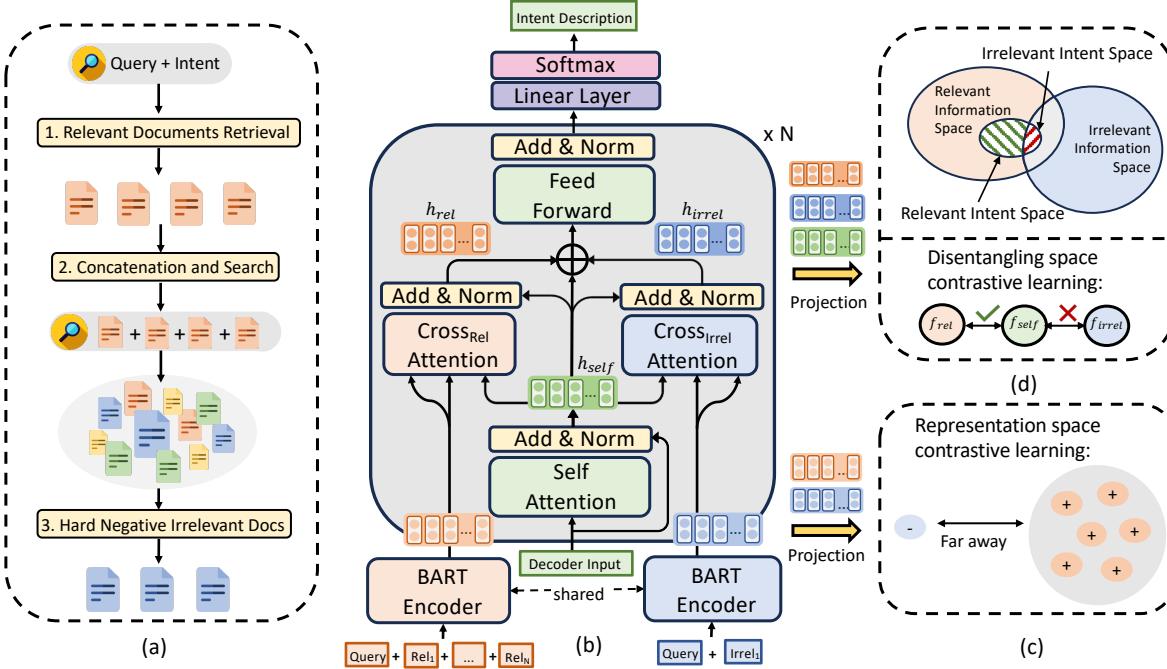


Figure 2: Overview of our proposed pipeline. From left to right, we show (a) Intent-Driven Negative Augmentation method, (b) Contrastive decoder structure with dual cross-attention layers, (c) and (d) Contrastive learning via dual space modeling.

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. Naturally, we expect the encoder to model and distinguish the relevant documents and irrelevant documents concerning a query in a feature representation space. To this end, we implement a dual encoder architecture, with each encoder being a cross-encoder, so that each cross-encoder directly models the relevance of a relevant or irrelevant document given a query by jointly encoding the query-document pair.

Intuitively, relevant documents are assumed to share similar topics, reflecting their connection to the specific information need behind a query to varying degrees. Therefore, it is natural to encode relevant documents collectively by concatenating them according to the ranking outlined in Section 3.2, Step 1. 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. Instead, 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}' . We aim to model a feature representation space for relevant documents, reflecting their shared topics, while ensuring the irrelevant feature representation space is distant from the relevant one due to the dissimilarity of their topics. To achieve this, we project both of the encoder last hidden states into a feature representation space via a linear layer. The feature representation for each relevant document

is obtained by applying average pooling over the token representations corresponding to the document. We then pull the relevant feature representations closer together while pushing the irrelevant representation further away from each relevant one (Figure 2 (c)). The optimization objective of the relevant feature representation space is to minimize:

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

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:

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

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:

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

3.3.2 Disentangling space modeling. In the previous encoding phase, we obtain contextual feature representations for relevant and irrelevant documents given a query. In the decoder part, we want to make sure the decoder generates intent descriptions based on these information sources. Hence, we design a contrastive decoder by adding an additional cross attention so that it can attend to both relevant and irrelevant documents. To effectively disentangle the

contrastive information learned by the model during the decoding process, we apply additional contrastive learning in disentangling space. As shown in Figure 2 (d), the model is designed to capture the contrast between relevant and irrelevant information spaces, allowing it to model both relevant and irrelevant intent spaces. This enables the generation of more precise and nuanced intent descriptions by focusing on the relevant intent space while minimizing attention on the irrelevant one.

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 $h_{self,z}^l$ by employing masked self-attention layers, to make sure the prediction of position z depends only on the predictions before z . Based on $h_{self,z}^l$, we compute relevant document hidden states $h_{rel,z}^l$ by applying multi-head attention with cross-attention (MHAtt) to relevant encoder output:

$$h_{rel,z}^l = \text{MHAtt}(h_{self,z}^l, h_{R^*}) \quad (4)$$

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

$$h_{irrel,z}^l = \text{MHAtt}(h_{self,z}^l, h_{i'}) \quad (5)$$

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

$$h_{combine,z}^l = h_{self,z}^l + h_{rel,z}^l - h_{irrel,z}^l \quad (6)$$

$$P_z^{vocab} = \text{Softmax}(W(h_{combine,z}^N)) \quad (7)$$

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

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

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_{self}^N , h_{rel}^N , and h_{irrel}^N , 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 [7] and use from-batch negative samples \mathcal{B} in the InfoNCE loss [14]:

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

where τ is the temperature and $\cos(,)$ 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:

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

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

4 Experimental Settings

4.1 Dataset

Q2ID [47] is a benchmark dataset for query-to-intent description, derived from existing TREC and SemEval collections: **TREC** is an ongoing series of benchmark activities in IR. Q2ID comprises the TREC Dynamic Domain track 2015, 2016, 2017 on dynamic, exploratory search and the TREC 2004 Robust Track on consistency of retrieval technology. These tracks together contribute to 510 entries in the Q2ID benchmark dataset. **SemEval** is a series of workshops on Semantic Evaluation. Q2ID comprises the English SemEval-2015 Task3 and SemEval-2016 Task3 on Community Question Answering, contributing 4878 entries to the Q2ID dataset.

For each query in the collections, documents with multi-graded relevance labels were converted into binary labels, indicating whether they are relevant or not to the query. In total, the dataset comprises 5,358 entries, each formatted as a quadruple: <query, relevant documents, irrelevant documents, intent description>, with intent descriptions being human-written narratives that provide detailed descriptions of the search intent behind the queries. The length statistics are an average of 7.2 tokens per query and 45.5 tokens per intent description. We split the 5338 queries into training (5000), validation (100), and test (258) query sets according to the original split setting of the Q2ID benchmark dataset.

4.2 Baselines

Since query-focused summarization task and query intent generation task are closely aligned in their focus on user queries and the extraction of relevant information, we evaluate our model against state-of-the-art models from the Q2ID benchmark, QFS task, and additionally include LLM zero-shot and two-shot baselines.

- *Pretrained sequence-to-sequence model baselines:* **T5** [30]: a Transformer-based encoder-decoder [38] 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** [17]: 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** [47]: 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** [1]: We use the Llama3.1-8B instruction-tuned text-only model in 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.
- *Query Focused Summarization (QFS) baselines:* **RelReg** [40] and **RelRegTT** [40]: 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** [40]: 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** [26]: an unsupervised, question-driven pretraining approach designed to tailor generic language models for controllable summarization tasks. **Qontsum** [35]: 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.

4.3 Implementation Details

We build our method based on the BART-large [17] model using Huggingface Transformers [43]. Cross_{Rel} attention and Cross_{Irrel} attention layers in the decoder blocks share the same initial weights from the pre-trained BART-large model. For the IDNA step, we set the expected number of irrelevant documents per query to three in the training set. To ensure this, we augment 1,984 queries so that each query includes at least three irrelevant documents. We train the model for 10 epochs using the Adam optimizer with a learning rate of 0.0001. The final checkpoints are chosen based on the average ROUGE-{1, 2, L} scores against the ground truth query intent obtained from the validation set. During decoding, we set the maximum length to 256 tokens and apply beam search with a beam size of 4, using a no-repeat n-gram size of 3. We optimize λ during training and fix on ($\lambda_0 = 0.2, \lambda_1 = 0.2, \lambda_2 = 0.6$) to balance the three losses. All experiments are conducted five times, and we report the average results across these runs. The implementation details for the other baseline models can be found in Appendix A.1.

4.4 Evaluation Metrics

We conduct three kinds 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. [47], along with BERTScore [48], which assesses semantic and syntactic similarity beyond exact word matches. Additionally, we conduct a human evaluation study using 50 randomly selected test samples. Three 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), and the final scores were averaged across the three annotators.

- **Fluency:** to what extent the generated query intent description reads naturally, understandably, and without noticeable errors or disruptions.
- **Factual Alignment:** to what extent the generated query intent description is factually aligned with the ground truth intent.
- **Inclusion score:** how well the generated query intent includes important details from the query and relevant documents.

Table 1: Performance between our model and baselines in terms of automatic evaluation (%). [†] 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. The best results are highlighted in bold, while the best baseline results are underlined.

Models	RG-1	RG-2	RG-L	BS
CtrsGen [†]	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 zero-shot	29.28	7.42	20.90	57.26
LLaMa3.1 2-shot	32.75	9.54	24.34	57.89
RelReg	26.67	12.83	21.99	59.24
RelRegTT	27.21	12.77	22.25	59.60
SegEnc	<u>31.83</u>	<u>14.29</u>	<u>25.18</u>	<u>62.15</u>
+ 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*

- **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.

However, conventional automatic evaluation metrics tend to show weak correlation with human judgments. Recent study [12, 22] propose using LLMs as evaluators for natural language generation (NLG) systems, scoring generated outputs based on generation probability rather than reference targets. Following the method of [22], we use LLaMa3.1-8B¹ and GPT-4o² 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, we use the output token probabilities from the LLMs to normalize the scores and take their weighted summation as the final results: $score = \sum_{i=1}^n p(s_i) \times s_i$ where $S = \{s_1, s_2, \dots, s_n\}$ 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. Example prompts are in Appendix B.

5 Experimental Results

5.1 Overall Results

We compare model performance between QUIDS and baselines in Table 1. The results show that: (1) Our model QUIDS outperforms all other baselines; (2) Our approach is compatible with both T5 and BART architectures, which both have a transformer-based encoder-decoder structure. This demonstrates the effectiveness of our design in the query intent generation task. Notably, BART-large

¹<https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct>

²<https://openai.com/index/hello-gpt-4/>

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

Models	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

outperforms T5-large despite having nearly half the model size, highlighting its efficiency in this task. (3) We implemented QFS models for the query intent generation task, which already outperform the baseline model, CtrsGen. However, our design further improves ROUGE-1 by 2.64 and ROUGE-L by 1.59 compared to the best QFS model, SegEnc; (4) In our experiments, the two-shot setting with the LLaMa3.1 8B model significantly outperforms the zero-shot setting in ROUGE scores, while showing only minor improvements in BERTScore. This suggests that without fine-tuning, intent generation may be lexically similar but semantically misaligned.

5.2 Ablation Study

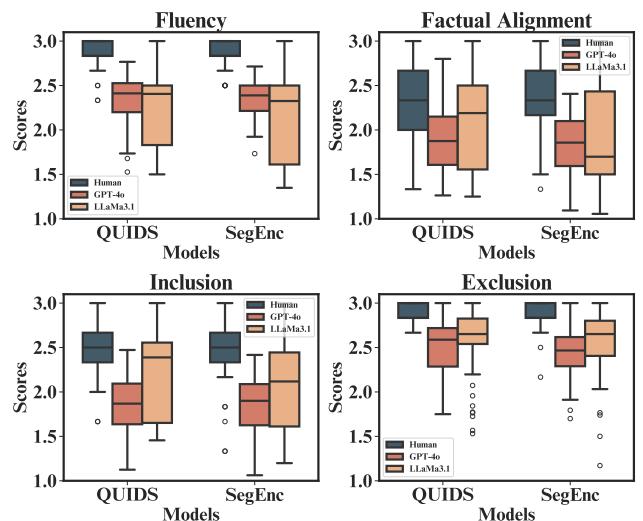
We conduct an ablation study on the test set to assess the impact of various modules in our proposed approach under three conditions: (1) We remove the IDNA module to examine the effect of data preprocessing (w/o IDNA); (2) We remove contrastive learning from the encoder and (3) decoder separately to assess the importance of representation space modeling (w/o RSM) and disentangling space modeling (w/o DSM), respectively. Table 2 shows the results. Removing IDNA leads to a 1.16-point drop in BERTScore, suggesting that incorporating harder irrelevant documents helps the model capture relevant intent information more effectively. Removing contrastive learning from the decoder leads to a significant drop in all metrics, highlighting its importance in modeling a distinguishable intent space and generating accurate intent descriptions. Finally, removing contrastive learning from the encoder results in a slight decrease, indicating that representation space modeling aids the model in distinguishing relevant from irrelevant information. Due to the similar architecture and experimental performance, we omit the T5 model results.

5.3 Human and LLM Evaluation

We analyze the quality of the generated intents by our model QUIDS and the best baseline model SegEnc with human evaluation and LLM-based approaches. Human inter-annotator agreement is assessed with weighted Cohen’s kappa, revealing fair to moderate consistency across the four metrics. As shown in table 3, QUIDS consistently receives higher scores from all three evaluators, except for Factual Alignment, where human evaluators prefer SegEnc’s outputs. To examine the performance differences between humans and LLMs in this task, we present boxplots in Figure 3 that display the score distributions for each metric. Additionally, we calculate

Table 3: Comparison of human evaluation and LLM evaluation in terms of Fluency (Fluen.), Factual Alignment (Align.), Inclusion score (Inclu.) and Exclusion score (Exclus.).

Method	Model	Fluen.	Align.	Inclu.	Exluc.
Human	SegEnc	4.71	3.70	3.88	4.71
	QUIDS	4.73	3.61	3.97	4.73
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

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

the Spearman and Kendall-Tau correlations between the evaluations from LLMs and human judgments, as shown in Table 5 in Appendix A.2. From this analysis, we derive the following insights:

- Humans tend to assign higher scores than GPT-4o and LLaMa 3.1 across all metrics, particularly for Fluency and Exclusion, where scores nearly reach 5. The variability in fluency scores suggests humans are more tolerant of less fluent text.
- Human evaluations exhibit the narrowest range, while GPT-4o shows a similar range but with more critical scores. As supported by the correlation table in Appendix A.2, GPT-4o aligns more closely with human evaluations, particularly in Factual Alignment. However, a difference in model preferences for this metric prompts further analysis of sub-dataset differences in Section 5.4.
- For the metrics Factual Alignment and Inclusion, all methods tend to exhibit a wider distribution with generally lower scores. This suggests variability in the assessments of these two metrics, potentially indicating inconsistency in how both human and LLM evaluators interpret the scoring criteria.

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.
<p>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...</p>	<p>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...</p>
<p>Generated Intent 1: How will the price of Freon-12 be impacted by the phasing out of this refrigerant?</p>	<p>Generated Intent 2: Identify documents that discuss the effects of the international agreement to phase out Freon-12 as a refrigerant.</p>

(a)

(b)

Figure 4: (a) Visualizes the token-level decoder cross-attention weights for a relevant document snippet when generating the token ‘impacted’ in Intent 1, *without an irrelevant document provided*. (b) for relevant document and irrelevant document when generating the token ‘effects’ in Intent 2, *with an irrelevant document provided*. Deeper color indicates a higher value.

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

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

5.4 Analysis of Intent Types

For analysis of the effect of intent type, 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. 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 [47], 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 [47], 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. We additionally conduct a breakdown analysis by intent types for the human and LLM results in Appendix A.3.

5.5 Analysis of Document Length

In real-world scenarios, relevant documents associated with a query can vary in length. To assess the impact of document length on the generated output, we categorize the documents into three groups based on the BART-large input limit (1024 tokens): documents

shorter than 512 tokens are classified as short documents; those ranging from 512 to 1024 tokens are termed medium-sized documents; and documents longer than 1024 tokens, which the model will truncate, are labeled as long documents. In our test dataset, there are 204 short, 25 medium, and 29 long documents. The overall differences in performance among all the input length categories are minimal across all metrics (also see Figure 6 in Appendix A.4). This indicates that 1024 tokens provide sufficient information for our model to effectively capture the underlying query intent. This may explain why QFS approaches, particularly those designed to handle lengthy documents, do not demonstrate significant advantages in the query intent generation task.

5.6 Case Study

To intuitively understand how irrelevant documents influence our model, we compare the generated intents for a query ‘Freon-12’ from test set with and without providing an irrelevant document. Figure 4 illustrates the token-level decoder cross-attention weights for both a relevant document snippet, and an irrelevant document snippet when provided. As shown in Figure 4 (a), without an irrelevant document provided, when generating the word ‘impacted’, the model’s focus is centered on the economic effects on cars, as indicated by the mention of ‘price’ in Intent 1. With an irrelevant document provided in Figure 4 (b), when generating the word ‘effects’ in Intent 2, similar attentions can be detected on ‘prices’, ‘\$', ‘costs’, ‘auto’ for both relevant and irrelevant documents. However, as shown in Generated Intent 2, our model now treats tokens related to prices and cars in relevant documents as irrelevant. This demonstrates that our model effectively generates intent descriptions that excludes this unrelated information.

6 Conclusion

In this paper, we introduced a novel dual-space modeling approach for the task of query intent generation. Our approach implements contrastive learning in both the encoding and decoding phases, combined with intent-driven hard negative augmentation during data preprocessing, to automatically generate detailed and precise

intent descriptions. Experimental results show that our model effectively filters out irrelevant information from the relevant intent space, leading to more accurate intent descriptions and greater transparency in the retrieval process. For future work, we aim to extend our approach to conversational search by first mining exploratory needs and then explaining the understanding of query intents.

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).

References

- [1] 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.
- [2] Steven M Beitzel, Eric C Jensen, Ophir Frieder, David D Lewis, Abdur Chowdhury, and Aleksander Koltcz. 2005. Improving automatic query classification via semi-supervised learning. In *Fifth IEEE International Conference on Data Mining (ICDM'05)*. IEEE, 8–pp.
- [3] Andrei Broder. 2002. A taxonomy of web search. In *ACM Sigir forum*, Vol. 36. ACM New York, NY, USA, 3–10.
- [4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.
- [5] Haonan Chen, Zhicheng Dou, Yutao Zhu, Zhao Cao, Xiaohua Cheng, and Ji-Rong Wen. 2022. Enhancing user behavior sequence modeling by generative tasks for session search. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 180–190.
- [6] Jia Chen, Yiqun Liu, Jiaxin Mao, Fan Zhang, Tetsuya Sakai, Weizhi Ma, Min Zhang, and Shaoping Ma. 2021. Incorporating query reformulating behavior into web search evaluation. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 171–180.
- [7] 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*. PMLR, 1597–1607.
- [8] Richard Csaky and Gábor Recsiki. 2020. The Gutenberg dialogue dataset. *arXiv preprint arXiv:2004.12752* (2020).
- [9] Elozino Eggonwan, Vittorio Castelli, and Md Arifat Sultan. 2019. Cross-task knowledge transfer for query-based text summarization. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*. 72–77.
- [10] Alexander Fabbri, Xiaojian Wu, Srinidhi 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*, Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (Eds.). Association for Computational Linguistics, Seattle, United States, 2508–2520. <https://doi.org/10.18653/v1/2022.naacl-main.180>
- [11] 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*. 961–964.
- [12] Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2024. GPTScore: Evaluate as You Desire. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Kevin Duh, Helena Gomez, and Steven Bethard (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 6556–6576. <https://doi.org/10.18653/v1/2024.naacl-long.365>
- [13] Jiafeng Guo and Yanyan Lan. 2020. Query classification. *Query Understanding for Search Engines* (2020), 15–41.
- [14] 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*. 9729–9738.
- [15] Yuan Hong, Jaideep Vaidya, Haibing Lu, and Wen Ming Liu. 2016. Accurate and efficient query clustering via top ranked search results. In *Web Intelligence*, Vol. 14. IOS Press, 119–138.
- [16] 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* (2020).
- [17] 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*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetraault (Eds.). Association for Computational Linguistics, Online, 7871–7880. <https://doi.org/10.18653/v1/2020.acl-main.703>
- [18] 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*. 3315–3319.
- [19] Ying Li, Zijian Zheng, and Honghua Dai. 2005. KDD CUP-2005 report: Facing a great challenge. *ACM SIGKDD Explorations Newsletter* 7, 2 (2005), 91–99.
- [20] 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*, Eneko Agirre, Marianna Apidianaki, and Ivan Vulic (Eds.). Association for Computational Linguistics, Dublin, Ireland and Online, 100–114. <https://doi.org/10.18653/v1/2022.deelio-1.10>
- [21] Xingxian Liu, Bin Duan, Bo Xiao, and Yajing Xu. 2023. Query-Utterance Attention With Joint Modeling For Query-Focused Meeting Summarization. In *ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.
- [22] Yang Liu, Dan Iter, Yichong Xu, Shuo Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-Eval: NLG Evaluation using Gpt-4 with Better Human Alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 2511–2522. <https://doi.org/10.18653/v1/2023.emnlp-main.153>
- [23] Chang Lu, Liqiu 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)*. IEEE, 5334–5340.
- [24] Yuning Mao, Pengcheng He, Xiaodong Liu, Yelong Shen, Jianfeng Gao, Jiawei Han, and Weizhu Chen. 2021. Generation-Augmented Retrieval for Open-Domain Question Answering. 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)*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 4089–4100. <https://doi.org/10.18653/v1/2021.acl-long.316>
- [25] 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. (2016).
- [26] Artidor 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)*, Anna Rogers, Jordan Boyd-Graber, and Naoki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 12737–12755. <https://doi.org/10.18653/v1/2023.acl-long.713>
- [27] 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*. 3–8.
- [28] Marius A Pasca and Sandra M Harabagiu. 2001. High performance question/answering. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*. 366–374.
- [29] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research* 21, 140 (2020), 1–67.
- [30] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research* 21, 140 (2020), 1–67.
- [31] Manukonda Sumathi Rani and Geddati China Babu. 2019. Efficient query clustering technique and context well-informed document clustering. In *Soft Computing and Signal Processing: Proceedings of ICSCSP 2018, Volume 1*. Springer, 261–271.
- [32] 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. <http://arxiv.org/abs/1908.10084>
- [33] Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How Much Knowledge Can You Pack Into the Parameters of a Language Model?. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 5418–5426. <https://doi.org/10.18653/v1/2020.emnlp-main.437>
- [34] Daniel E Rose and Danny Levinson. 2004. Understanding user goals in web search. In *Proceedings of the 13th international conference on World Wide Web*. 13–19.

- [35] Sajad Sotudeh and Nazli Goharian. 2023. Qontsum: On contrasting salient content for query-focused summarization. *arXiv preprint arXiv:2307.07586* (2023).
- [36] 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*, Karin Verspoor, Kevin Bretonnel Cohen, Michael Conway, Berry de Brujin, Mark Dredze, Rada Mihalcea, and Byron Wallace (Eds.). Association for Computational Linguistics, Online. <https://doi.org/10.18653/v1/2020.nlpcovid19-2.14>
- [37] 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*. 3124–3131.
- [38] 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 (2017).
- [39] 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 (2013), 2224–2237.
- [40] Jesse Vig, Alexander Richard Fabbri, Wojciech Kryściński, Chien-Sheng Wu, and Wenhai Liu. 2022. Exploring Neural Models for Query-Focused Summarization. In *Findings of the Association for Computational Linguistics: NAACL 2022*. 1455–1468.
- [41] 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.
- [42] Ji-Rong Wen, Jian-Yun Nie, and Hong-Jiang Zhang. 2002. Query clustering using user logs. *ACM Transactions on Information Systems* 20, 1 (2002), 59–81.
- [43] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierrick Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface’s transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771* (2019).
- [44] Yumo Xu and Mirella Lapata. 2020. Query focused multi-document summarization with distant supervision. *arXiv preprint arXiv:2004.03027* (2020).
- [45] 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)*. 6096–6109.
- [46] Yumo Xu and Mirella Lapata. 2022. Document summarization with latent queries. *Transactions of the Association for Computational Linguistics* 10 (2022), 623–638.
- [47] 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*. 1823–1832.
- [48] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675* (2019).
- [49] Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, et al. 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*. 5905–5921.

A Appendix A

A.1 Implementation details for baseline models

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 [8] from the Huggingface Model Hub³ 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.

³<https://huggingface.co/Salesforce/socratic-books-30M>

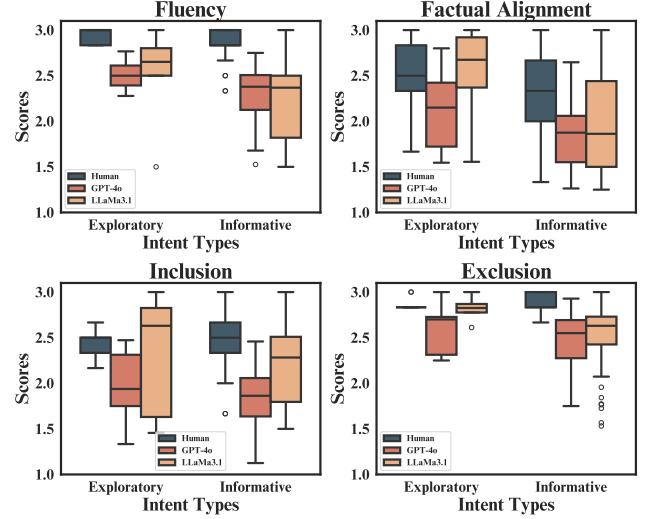


Figure 5: Boxplot of evaluation scores on 4 metrics of our model on different intent types.

A.2 Correlation with Human Evaluation

We assess the correlation between human and LLM evaluators across four qualitative evaluation criteria, presenting the Spearman (ρ) and Kendall-Tau (τ) correlations for the best SOTA model SegEnc and our QUIDS model in Table 5. 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.

A.3 Breakdown Analysis 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 human evaluations on different intent types. Table 6 present human and LLM evaluations on our model regarding two intent types. While humans prefer our model for exploratory intent with 4.04 (QUIDS) vs. 3.81 (SegEnc), SegEnc is favored for informative intent with 3.51 (QUIDS) vs. 3.68 (SegEnc). Since informative intent queries dominate, this leads to a lower average score for our model. We also present the evaluation scores distribution on different intent types over 50 the same test samples

Table 5: Spearman (ρ) and Kendall-Tau (τ) correlations between Human evaluation and LLM evaluation of different metrics.

Correlation	Model	Fluen.		Align.		Inclu.		Exluc.	
		ρ	τ	ρ	τ	ρ	τ	ρ	τ
Corr (Human, GPT-4o)	SegEnc	0.377	0.304	0.409	0.310	0.184	0.159	0.369	0.306
	QUIDS	0.476	0.375	0.517	0.412	0.252	0.199	0.361	0.286
Corr (Human, LLaMa3.1)	SegEnc	0.243	0.187	0.459	0.344	0.143	0.111	0.224	0.177
	QUIDS	0.343	0.278	0.553	0.423	0.325	0.248	0.158	0.126

Table 6: Comparison of Human and LLM evaluation on informational and exploratory intent types on our model.

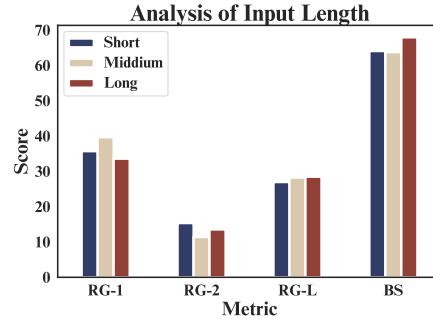
Method	Intent	Fluen.	Align.	Inclu.	Exluc.
Human	Info.	4.71	3.51	3.98	4.73
	Expl.	4.85	4.04	3.88	4.74
GPT-4o	Info.	3.64	2.70	2.63	3.95
	Expl.	4.02	3.17	2.96	4.20
LLaMa3.1	Info.	3.33	2.96	3.35	3.99
	Expl.	3.72	4.11	3.72	4.64

in subsection 5.4. Figure 5 is a breakdown of Figure 3 on our model for four metrics. Based on Table 6 and Figure 5, we have following insights:

- Exploratory intent types generally outperform informational intent types across all metrics, except for human evaluations on inclusion. This finding, based on 50 test samples from both human and LLM evaluations, is consistent with the automatic evaluation results for the full test dataset, as shown in Table 4.
- The boxplots for informational intent on the right side of each metric in Figure 5 exhibit a distribution similar to the overall performance of our model presented in Figure 3, suggesting that informational intent queries dominate the dataset and largely drive performance outcomes. However, exploratory intent queries outperform in this particular task, indicating their stronger performance despite being less frequent.

A.4 Breakdown Analysis by document length

Figure 6 shows the results of the automatic evaluation, broken down based on document length.

**Figure 6:** Comparison of automatic evaluation on our model QUIDS on different input lengths.

B Appendix B

B.1 Evaluation prompt on Fluency

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:

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.

Query:

{ {Query} }

Relevant documents:

`{{{Relevant documents}}}`

Irrelevant documents:

`{{{Irrelevant documents}}}`

Generated Intent:

`{{{Generated intent}}}`

Ground Truth Intent:

`{{{Gound truth intent}}}`

Evaluation Form (scores ONLY):

- Fluency:

B.2 Evaluation prompt on Factual Alignment

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:

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.

Query:

`{{{Query}}}`

Relevant documents:

`{{{Relevant documents}}}`

Irrelevant documents:

`{{{Irrelevant documents}}}`

Generated Intent:

`{{{Generated intent}}}`

Ground Truth Intent:

`{{{Gound truth intent}}}`

Evaluation Form (scores ONLY):

- Factual Alignment:

B.3 Evaluation prompt on Inclusion Score

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:

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.

Query:

`{{{Query}}}`

Relevant documents:

`{{{Relevant documents}}}`

Irrelevant documents:

`{{{Irrelevant documents}}}`

Generated Intent:

`{{{Generated intent}}}`

Ground Truth Intent:

`{{{Gound truth intent}}}`

Evaluation Form (scores ONLY):

- Inclusion Score:

B.4 Evaluation prompt on Exclusion Score

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:

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.

Query:

`{{Query}}`

Relevant documents:

`{{Relevant documents}}`

Irrelevant documents:

`{{Irrelevant documents}}`

Generated Intent:

`{{Generated intent}}`

Ground Truth Intent:

`{{Ground truth intent}}`

Evaluation Form (scores ONLY):

- Exclusion Score: