The StatCan Dialogue Dataset: Retrieving Data Tables through Conversations with Genuine Intents

Xing Han Lu[1 2 †]   **Siva Reddy**[1 2 3]   **Harm de Vries**[2]

[1] Mila Quebec AI Institute, McGill University
[2] ServiceNow Research
[3] Facebook CIFAR AI Chair
`statcan.dialogue.dataset@mila.quebec`

Abstract

We introduce the StatCan Dialogue Dataset[1] consisting of 19,379 conversation turns between agents working at Statistics Canada and online users looking for published data tables. The conversations stem from genuine intents, are held in English or French, and lead to agents retrieving one of over 5000 complex data tables. Based on this dataset, we propose two tasks: (1) automatic retrieval of relevant tables based on a on-going conversation, and (2) automatic generation of appropriate agent responses at each turn. We investigate the difficulty of each task by establishing strong baselines. Our experiments on a temporal data split reveal that all models struggle to generalize to future conversations, as we observe a significant drop in performance across both tasks when we move from the validation to the test set. In addition, we find that response generation models struggle to decide when to return a table. Considering that the tasks pose significant challenges to existing models, we encourage the community to develop models for our task, which can be directly used to help knowledge workers find relevant tables for live chat users.

1 Introduction

One of the longstanding goals in Natural Language Processing (NLP) is to develop conversational agents that assist people with concrete tasks, such as finding information in large collections of documents or booking restaurants and hotels. To aid the development of such virtual assistants, the research community is in need of benchmarks that reflect the intents and linguistic phenomena found in real-world applications. However, developing such real-world conversational datasets is challenging in the current research landscape. On the one hand, academic labs often struggle to come up with

U_1:	Hi, I'm looking to obtain quarterly data in regards to GDP grow (Canada), BC Housing STarts, Canada Oil Price/BBL
A_1:	Hello, my name is Kelly C. Give me one moment as I search [...]
A_1:	For GDP growth rates, please consult the following link: [...]
A_1:	What do you mean by BC Housing Starts?
U_2:	I'm required to research all of the housing starts for BC on a quarterly basis [...]
U_2:	Housing starts are the number of new residential construction projects that have begun during any particular month [...]
A_2:	I would have monthly data regarding new building permits being issued [...]
A_2:	**Building permits, by type of structure and type of work: https[...]**
A_2:	I'll have a look for oil prices. One moment.
U_3:	Do you also have data to Canada'as oil Price/BBL ("WTI")? [...]
A_3:	Are you looking for the retail prices of oil?
A_3:	If so, I found some data for smaller geographies.
A_3:	**Monthly average retail prices for gasoline and fuel oil, by geography (https[...])**
A_3:	[...] Would those geographies be enough?
A_3:	Or are you looking for Canada only?
U_4:	[...] I would need something that pertains more to all of canada
A_4:	What about this? **Monthly average retail prices for food and other selected products (https[...])**

Table 1: An example of the StatCan Dialogue Dataset in which a user (U) talks to a StatCan agent (A) to find a number of data tables. Text in **bold** indicates the title of a table retrieved by the agent.

natural use cases of task-oriented dialogue agents and collect conversations with a large number of real users. Many labs have designed artificial tasks and collected conversations from crowd workers with simulated intents (Budzianowski et al., 2018; Adlakha et al., 2022; Lee et al., 2022), often leading to datasets that do not capture the linguistic challenges of production settings (de Vries et al., 2020). On the other hand, industry labs might have access to users with genuine intents (e.g., through Siri or Alexa) but rarely release such conversational

---

Dataset	Intent	Dialogue	Query	Result(s)	Source(s)	Lang.
Our work	Genuine	✓	Question, Request	Table link, Dial. act	StatCan	En, Fr
NQ (2019)	Mixed	✗	Question	Span excerpt	Google, Wiki.	En
DuReader (2018)	Mixed	✗	Question	Span excerpt	Baidu	Zh
OTT-QA (2020)	Simulated	✗	Question	Table/Span excerpt	Wikipedia	En
TAPAS-NQ (2021)	Mixed	✗	Question	Table excerpt	Google, Wiki.	En
CoQA (2019)	Simulated	✓	Question	Span excerpt*	Multiple	En
QuAC (2018)	Simulated	✓	Question	Dial. act	Wiki.	En
ATIS (1990)	Genuine	✓	Request	SQL query, Command	TI Corp.	En
SGD-X (2022)	Simulated	✓	Request	API call, Dial. act	Dial. Simulator	En

Table 2: Comparison with related datasets (see Section 2). (*) CoQA uses rationales to support extracted answers.

datasets due to their commercial value and user privacy concerns. Hence, we argue that the research community would benefit from a task-oriented dialogue environment where findings can be validated with real users, and, to that effect, present a unique dataset in collaboration with Statistics Canada.

Statistics Canada (StatCan) is a national statistics agency commissioned with collecting key information on Canada's economy, society, and environment. Statistics Canada conducts hundreds of surveys on virtually all aspects of Canadian life and publishes the resulting data tables on statcan.gc.ca. This website currently features 5K+ of such complex and often large data tables. Canadian citizens—and other interested individuals—come to this website to find the statistics they are looking for. The StatCan website offers a chat functionality (available in English and French) to help users in case they can not find the appropriate information.

Sourcing from these live chats, we present the StatCan Dialogue Dataset, a collection of 20K+ English and French conversations between visitors of statcan.gc.ca and agents working at Statistics Canada. Before releasing this dataset, StatCan has ran several procedures to remove Personally Identifiable Information (PII). While we observe a wide variety of user intents, ranging from table manipulation to navigation instructions, a large number of visitors use the chat functionality to find data tables on the StatCan website. Specifically, we observe 6.6K instances where agent returns a link to a data table across 4.4K conversations. In Table 1, we provide an example conversation in which an online user is looking for specific data tables.

In this work, we develop two novel tasks centered on helping users find specific tables. First, we introduce the table retrieval task, which requires a model to predict the table returned by the agent given the messages sent so far. Second, we introduce the response generation task, which requires a model to predict the agent's response given the

dialogue history. For both tasks, we investigate its difficulty by establishing strong baselines and evaluating them on various metrics.

We stress that both tasks are immediately useful in a real-world setting. The table retrieval task can help agents find relevant tables faster while the response generation task may lead to a virtual agent that can return relevant tables through an online conversation. We hope that this tight connection with a real-world scenario will bring the research community more insight into the challenges of developing practical dialogue agents and lead to faster transfer of research ideas and findings.

2 Related Work

This section presents various directions related to our work. See Table 2 for a comparative summary.

Open-domain QA This is the task of answering questions using a large and diverse collection of text documents. One of the first large-scale evaluations in open-domain QA was presented at TREC-8 (Voorhees, 2001). Since then, many studies have released large-scale open-domain QA datasets: WikiQA (Yang et al., 2015) and MS MARCO (Bajaj et al., 2018) source questions from the Bing search engine, Natural Questions (NQ) (Kwiatkowski et al., 2019) from Google search, and DuReader (He et al., 2018) source questions in Chinese from Baidu. The questions come from real users and the answers are collected from the search results through crowd workers. Although those datasets have questions with genuine intent and the answer must be retrieved from a collection of documents, our dataset emphasizes the retrieval of tables (in a conversational setting) rather than free-form documents.

Table retrieval and QA Following works on tabular pre-training (Yin et al., 2020), table-to-text generation (Parikh et al., 2020) and weak supervision for semantic parsing (Herzig et al., 2020),

Dataset	Train	Valid	Test	All
# Conv.	2573	545	557	3675
# Turns	11382	2339	2600	16321
# Messages	36147	7385	8340	51872
# Queries	3782	799	870	5451
# Tokens / Msg	32.83	33.51	29.32	32.36
# Turns / Conv.	4.42	4.29	4.67	4.44
# Msg / Conv.	14.05	13.55	14.97	14.11
# Queries / Conv.	1.48	1.47	1.57	1.49
# Tables	778	349	388	959
# New tables	0	41	145	181
# Dims / Table	3.5	3.5	3.6	3.6
# Mbrs / Table	185.5	210.8	175.6	172.1
# Notes / Table	21.1	22.7	23	20.4

Table 3: Statistics of English conversations and tables in the retrieval and generation tasks. New tables are calculated with respect to training set (see Table 11).

Chen et al. (2020) and Herzig et al. (2021) respectively propose OTT-QA and TAPAS-NQ, two novel approaches that extend open-domain QA to retrieving tables instead of documents. The former collects both the questions and answers from crowd workers and the latter extends Natural Questions by using tables from the article where the answer was taken. In both cases, the tables being retrieved are sourced from Wikipedia articles. Although our data also incorporate tabular retrieval, the tables are sourced from statcan.gc.ca, they can be significantly larger (as discussed in Appendix A.2.1), and they are being retrieved in an interactive and conversational setting.

Conversational QA Several works extended question answering to the conversational setting. CoQA (Reddy et al., 2019) and QuAC (Choi et al., 2018) introduced datasets in which multiple rounds of questions are asked about a reference passage taken from a document (such as a Wikipedia article). Subsequent works extended this setup to an open-domain setting where the reference passage is not known beforehand (Qu et al., 2020; Anantha et al., 2021; Adlakha et al., 2022). Saeidi et al. (2018) proposed a conversational QA task about regulatory texts. Aforementioned datasets are all structured in the same way: at every turn, the first speaker will ask a question, and the other speaker will give an answer. In contrast, the queries in our conversations are not restricted to questions, and the answers can be either a table, metadata, or a dialogue act.

Task-oriented Dialogue Our work is related to work on task-oriented dialogue where users converse with virtual agents to accomplish
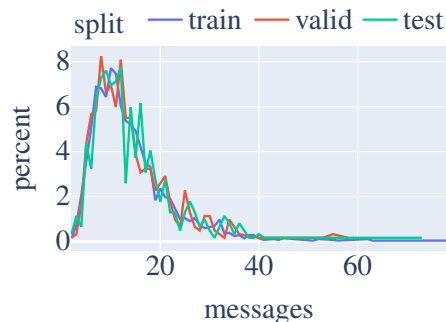

Figure 1: Histogram of messages by conversation in the both tasks (French split in Figure 7).

specific goals, such as booking a restaurant or resolving a customer issue. While early work has collected a dataset in a genuine information seeking setup (Hemphill et al., 1990), many recent datasets has collected them through a simulated setup (Budzianowski et al., 2018; Rastogi et al., 2020; Feng et al., 2020, 2021; Chen et al., 2021; Lee et al., 2022). Task-oriented models usually track the dialogue state by predicting dialogue acts that are specified through intents and slot-value pairs, e.g., findRestaurants(cuisine=Italian). While our dataset does not provide turn-based annotations, the released conversations come with an annotated goal i.e., which data table the user was looking for. Like other goal-oriented dialogue tasks, this annotation enables us to automatically evaluate the dialogue models through a task completion metric.

Chit-chat Dialogue The goal for chit-chat systems is to engage in a open-ended conversation with an end-user (Lowe et al., 2015; Dinan et al., 2018). Unlike our dataset, such conversations do not intend to assist the user with a specific task.

3 Dataset

The StatCan Dialogue Dataset consist of conversations collected from the live chat between March 1, 2019 till March 8, 2021. Although a variety of user intents can be found in the broader dataset of over 25K conversations, we focus on a single intent by selecting all conversations where the agent returns a data table. We use this subset to develop and test models for the two tasks that we introduce in Section 4. In Section 3.1, we provide basic statistics about this subset of the data and present a dialogue analysis for a small number of conversations in Section 3.2. In Section 3.3, we turn our attention to the

Merged Acts (*Example*)	%
Answer (*You can obtain on our...*)	50
Request (*...please help me retrieve data...*)	31
Time Mgmt (*Please hold*)	28
Inform (*Please take note that...*)	63
Info Seeking Ques. (*Do you have any other...?*)	25
Promise (*...please contact the Education Ministry...*)	18
Auto Feedback (*Sure*)	25
Offer (*...how may I help you?*)	11
Instruct (*Select at least one...*)	18
Clarif. Ques. (*Which of these lines would direct...*)	16
Greeting (*Hi*)	28
Self Introduction (*My name is...*)	17
Thanking (*Thanks a lot!*)	47
Accept Thanking (*you're welcome*)	15

Table 4: Frequency of merged speech acts occurring in 100 turns in conversations from the validation set.

data tables and explain what kind of information is available for them. Finally, we explain how the dataset is split into a train, validation, and test set in Section 3.4. For technical specifications, a dataset card is provided in Appendix B.

User intents The live chat was designed to fulfill specific user intents. The main intent of the chat functionality is to help users **find specific data tables**. For example, in Table 1, the agent helps the user find tables about building permit, gasoline price, and retail prices for food. Although, users might also be interested in **obtaining meta-information**, receive help in **manipulating** a table or with the **user interface**. In some cases, the user will make **out of domain requests**. Those auxiliary intents are described in Appendix A.1 since the focus of this work is on the main intent.

Messages and turns Each conversation is broken down in **turns**, which is a pair of user-agent responses. Each response can have multiple **messages** sent sequentially (e.g., in Table 1, the first agent response contains 3 back-to-back messages).

3.1 Basic statistics

In total, 25397 conversations will be made available. Based on our main intent, we focus on a subset of 4468 (3675 in English and 793 in French). Out of a total of 5907 tables available in both English and French, the agents returned 959 unique tables in English 285 in French. The number of messages by conversation varies between 2 and 78 with a median of 12 for the English split (see Figure 1 for the distribution). Based on Table 3, there's on average 4.4 turns but 14.12 messages

Title: Production and value of maple products
Date range: 1924-01-01 to 2020-01-01
Dimensions: Geography, Maple products
Subject: Agriculture
Survey: Maple Products
Frequency: Annual

Table 5: Basic information (including title) of table in Section 3.3. Full version in Table 20. This can be accessed at doi.org/10.25318/3210035401-eng.

per conversation, with over 30 tokens for each message. This indicates that the speakers will express multiple sequential thoughts before the addressees respond. For the French split, we analyzed the basic statistics in Appendix A.5.

Frequently requested tables In total, 6 tables make up 13.4% of tables retrieved, covering subjects like inflation and household spending. Supplementary details can be found in Appendix A.3.1.

3.2 Dialogue Analysis

We categorize 100 turns (306 messages) from 24 conversations in the English validation set according to the speech acts defined by Bunt et al. (2010, 2020), which is also known as ISO standard 24617-2. We follow their taxonomy but merge some fine-grained acts with their broader concepts (e.g., *correction, agreement, disagreement* with *inform*). We present the speech act frequencies and examples in Table 4. See the Appendix for more information on how we merged the original acts and supplementary examples (Table 15 and Table 22, respectively).

We notice that *answers* appear twice as frequent as *information seeking questions* because an interlocutor may provide an answer to both *clarification questions* and *requests*. Additionally, *inform* acts appears 63% of the time because agents need to expand upon their answer and users tend to clarify their initial requests by informing the other. Although less frequent, *auto feedback* and *time management* are still relevant because interlocutors cannot rely on visual feedback like nodding or typing. Naturally, *time management* often co-occurs with *promises* because the agent tends to put the user on hold while promising to fulfill their request.

3.3 Table specifications

To explain the specifications, we examine a sample table with **title** *Production and value of maple products* (shown in Table 5). The table has two **dimensions**, which are groups of **member items**;
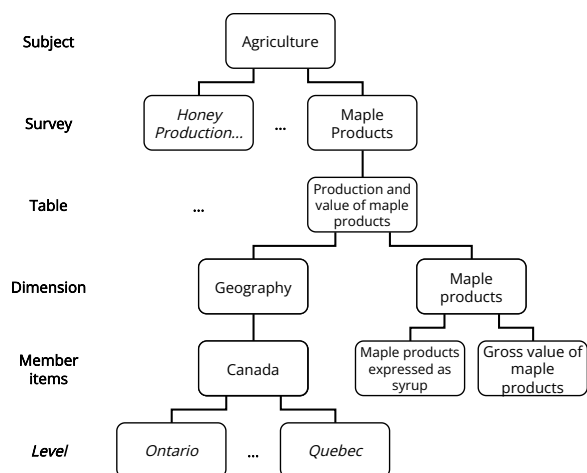
Figure 2: Diagram of the hierarchical relationship between metadata components, discussed in Section 3.3.

for *geography*, the members are the provinces producing maple syrup (Quebec, Ontario, etc.), and for *Maple products* the members are the production types (maple sugar, syrup, taffy and butter). A member item generalizes the concept of rows and columns as they are interchangeable via pivoting. Sometimes, details about those members are provided as **footnotes** at the end the page.

Basic Information   This is the core metadata and consists of the title, dimensions, subject, survey and update frequency (member items are excluded). Supplementary details are in Appendix A.2.2.

Hierarchical relation   The metadata can be viewed hierarchically. As shown in Figure 2, each subject encompasses different surveys, each survey can be used to generate one or more tables, and so on. A member item that can be nested under another member item is called *Level*.

3.4   Dataset splits

We group the conversations into a train (70%), a validation (15%) and a test (15%) set. The test set was specifically selected to be the most recent conversations by date (covering Sept 8, 2020 to Mar 8, 2021), whereas the training and validation set were randomly selected from the remaining data (covering Mar 1, 2019 to Sept 8, 2020). This lets us test a model's capability to adapt to temporal shifts in the data (such as new data releases and novel events). This is useful to understand a model's capability to generalize beyond the training distribution, but it is also a better reflection of real-world applications of a model (which will be used for future data). The same splits are used for all tasks.

4   Tasks

Based on the conversational and tabular data, we propose two tasks: (i) a table retrieval task, which requires a model to use a partial conversation to predict the table an agent will return, and (ii) a response generation task, which requires a model to use a partial conversation to generate the most probable response by the agent. The conversations in the tasks are available in both English and French.

4.1   Retrieval task

For this task, we truncate every conversation right before a link to a relevant table is shared by the agent. As a result, the product ID (PID) corresponding to that link becomes the objective of the retrieval task, as shown in Table 6. When the agent shares multiple non-repeating PIDs within a conversation, each unique occurrence is treated as a separate sample.

Recall@k   To evaluate models for retrieval, we compute the recall at k (R@k) score for $k \in \{1, 10, 20\}$, which corresponds to the rate where the correct table is among the k tables retrieved by the model (usually ranked by a relevance score). We choose $k = 1$ for real-time automatic retrieval and $k \in \{10, 20\}$ for scenarios where humans or automatic rerankers would like to use the retriever to query tables and select the best option.

4.2   Response generation task

In the first task, only the messages leading to a table retrieval are considered. For this task, each message sent by an agent is considered as a target and everything before is the source. Thus, the goal of this task is to use the source text to generate a response that matches the target (see Table 7). Since dialogue responses are challenging to evaluate, we report a wide variety of metrics for this task.

ROUGE-L and METEOR   ROUGE-L (Lin, 2004; Lin and Och, 2004) is a common text evaluation metric which naturally takes into account sentence level structure by identifying the longest overlapping word sequence between two sentences. METEOR (Banerjee and Lavie, 2005) is a word-level precision and recall scoring method that encompasses different ways to represent a word, including stems and synonyms.

BERTScore and MoverScore   Various methods were developed to leverage contextual embeddings

	Source text (on-going conversation)
	[...]
A:	What do you mean by BC Housing Starts?
U:	I'm required to research all of the housing starts for BC on a quarterly basis[...]
U:	Housing starts are the number of new residential construction projects that have begun during any particular month
A:	[...] I would have monthly data regarding new building permits being issued.
	Retrieval target (StatCan table)
	Table 34-10-0066 (Building permits, by type of structure and type of work)

Table 6: Source and targets of the retrieval task, based on Table 1. Given the on-going conversation, the goal is to retrieve a StatCan table.

from BERT (Devlin et al., 2019) to evaluate similarity between two sentences. BERTScore (Zhang et al., 2019) computes the cosine similarity at the token level, whereas MoverScore (Zhao et al., 2019) computes the earth mover distance (Rubner et al., 2000) at the word or sentence level, thus capturing the cost of transforming the distribution of the generated responses into the distribution of the original responses.

Title accuracy   In addition to the general metrics for text generation, we also explored this task-specific metric. We define it as the proportion of generated messages that contain the title of a table shared in the reference messages. Consequently, this metric only includes turns where a table is shared by an agent. To compute this, we (i) find the product ID in the reference message, (ii) look up the title, (iii) check if that title appears exactly in the generated and reference text.

5   Models

To help understand the performance of finetuned models on our tasks, this section presents an overview of the methods for the English splits, whereas the implementation details are covered in Appendix A.4. Similar architectures were used for French (described in Appendix A.5).

5.1   Retrieval

BM25   We use Robertson and Zaragoza (2009)'s algorithm to retrieve the metadata of a table (passage) similar to a given query by weighting the *idf*-scaled term frequency of query words with respect to the passages.

	Source text (on-going conversation)
U:	Hi, I'm looking to obtain quarterly data in regards to GDP grow (Canada), BC Housing STarts, Canada Oil Price/BBL
A:	Hello, my name is Kelly C. Give me one moment as I search [...]
A:	For GDP growth rates, please consult [...]
U:	I'm required to research all of the housing starts for BC on a quarterly basis [...]
	Generation target (next response by agent)
A:	I would have monthly data regarding new building permits being issued. [...]

Table 7: Source and targets of the response generation task, based on Table 1. Given the on-going conversation, the goal is to generate the agent's response.

DPR   Proposed by Karpukhin et al. (2020), Dense Passage Retrieval (DPR) is a pair of transformer models that separately encode a query and a passage, and the dot product of the resulting vectors will have a higher score if the passage is relevant to the query. We finetune this model to retrieve the metadata of a table (passage) given the on-going conversation (query).

TAPAS and TAPAS-NQ   Herzig et al. (2020) introduced a model that learned to encode flattened tables cells in a self-supervised manner during pretraining. We finetuned it to retrieve the truncated content of a table given an on-going conversation. Subsequently, Herzig et al. (2021) finetuned TAPAS to perform open-domain table retrieval on 12K questions-answer-table triplets extracted from NQ; we further finetune this variant in the same way and report the results as TAPAS-NQ.

Exploring table representation   In the simplest scenario, only the title is given to BM25 and DPR. Moreover, we evaluate variants that encode the basic information, member items, footnotes, or a combination of them. For TAPAS and TAPAS-NQ, we also finetuned a variant that retrieves the title, dimensions and member items, since the original TAPAS could attend titles and column names.

5.2   Response generation

T5   We finetuned the large variant of T5 (Raffel et al., 2020) (named *No aug.* in Table 10) to autoregressively decode the target (agent reply) after first encoding the source (on-going conversation).

Augmenting T5 with top-k title(s)   For every partial conversation, we use DPR (basic+member) to retrieve the top-k tables (where $k \in \{1, 5\}$), and append their titles to the partial conversation. This

Metadata	R@1	R@10	R@20
Basic	14.7	45.0	55.0
Basic + member	15.7	46.2	56.3
Basic + footnotes	13.9	44.4	54.2
Member	10.7	35.0	46.3
Title	13.9	43.8	53.4

Table 8: Retrieval results of DPR for the English test split with varying table representations. Overview of metadata in Section 3.3.

Model	R@1	R@10	R@20
BM25	0.3	2.3	3.8
DPR	14.3	45.1	54.2
TAPAS	6.1	22.1	31.5
TAPAS-NQ	7.4	30.0	39.3

Table 9: Retrieval results for the English test split when encoding title and member items. DPR and TAPAS were run 3 times and averaged.

allows T5 to decide between using one of the suggested titles and generating something else (e.g., clarification question). This is similar to the agents' behavior, as they tend to return a title with the URL when sharing a relevant table. Furthermore, supervising T5 to ignore or return a title is equivalent to an implicit binary classification.

6 Results and Discussions

Based on our baselines and data, we report the results and analyze the challenges that our dataset and tasks pose for existing models. For the English splits, the main retrieval results are reported in Table 9 and Table 8, and main generation results are in Table 10. Full results can be found in Appendix A, respectively in Table 16 and Table 33, and relevant statistical tests in Appendix A.7.

Impact of table representation In Table 9, we observe that the metadata representation affects the retrieval recall. Although DPR can achieves respectable results when it only retrieves the title, including basic information (defined in Section 3.3) yields slight improvements, and further adding member items results in a significant difference from only using title ($p = 0.014$). However, only using member item result in drastic decrease in recall ($p = 0.00086$), indicating the importance of the title. Moreover, footnotes do not yield any improvement, which may be because they often exceed the maximum context span (see Table 20). Thus, concisely but meaningfully representing metadata will be crucial to achieve a good recall on the retrieval task.

Metrics	No aug.	Top-1 Title	Top-5 Titles
METEOR	23.35	24.07	24.41
ROUGE-L	30.65	30.76	30.88
MoverScore	59.82	60.23	60.31
BERTScore	86.04	86.11	86.17
Title Acc.	6.96	7.99	10.82

Table 10: Response generation results for the English test split.

Transfer to table retrieval task Our experiments allow us to analyze the effectiveness of open-domain QA fine-tuning (NQ) and tabular pre-training when transferring to our table retrieval task. We observe in Table 9 that DPR outperforms TAPAS and TAPAS-NQ by respectively 23.0% and 15.1% in test recall@10. Moreover, TAPAS-NQ achieves a better performance when it only retrieves the title and member items instead of the full table ($p = 0.016$), likely due to repetitions and truncation due to context size limits. Although both DPR and TAPAS-NQ were trained on NQ, the latter was trained on a small subset (12K vs 320K) that contains tables. Our experiments indicate that TAPAS transfers poorly from one task (NQ-Tables) to another (StatCan).

Response generation We compare the performance of fine-tuned T5-large models with and without DPR-augmented table titles. In Table 10, we notice that retrieval-augmented models show modest improvements on the 4 non-task specific metrics. However, the top-5 augmented model achieves an absolute improvement of 3.86% in title accuracy, indicating that the information provided by DPR does help T5 in generating the desired title. It is nevertheless surprising that T5 without augmentation achieves a score of 6.96%, suggesting that the T5 model is capable of storing the titles seen at training time, and, to a limited extent, is able to recall and return them at test time. Finally, we point out that the title accuracy is still 5.91% lower than top-1 recall of the DPR retriever (Table 8), indicating that T5 fails to learn when to return a table (despite the agent retrieving a table in 23.4% of all turns). In the case of top-5 titles, T5 struggles to decide which table title to return.

Qualitative analysis of generated responses We examine various conversations to understand what type of responses are generated by T5. We find that it can generate simple speech acts like greetings, but can struggle with context-specific speech acts such as clarification questions. Moreover, it can
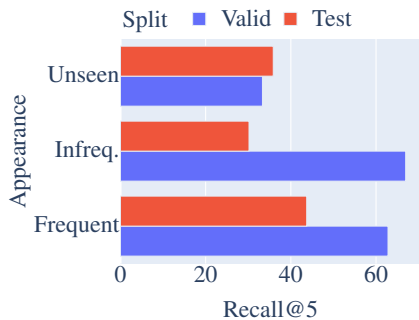
Figure 3: Results for DPR B+M for tables appearing in the training set frequently (10+ times), infrequently (1-10 times), and unseen at train time.

Figure 4: Validation and test recall for a selected set of retrieval models. We observe a significant drop in performance. *B+M* denotes Basic + member.

reliably reply with the title of a common table, struggles for uncommon ones, and is sometimes capable of generating unseen titles with the help of DPR. The full analysis is in Appendix A.6.

Temporal drifts   As explained in Section 3.4, we use a temporal split to test the model's ability to generalize to future conversations. We observe a significant drop in recall (13%-28.3%) in Figure 4 when we compare the validation and test set performance, even when the models are trained with varying metadata representations. Similarly, T5 achieves low scores on the test split for the response generation task (Table 10). This large gap suggests that trained models struggle to generalize to future conversations. First, we found that this is likely caused by the number of new tables that appear in the test split (145) compared to validation (41), as shown in Table 3. Moreover, the subjects of the conversations have significantly changed: users started to care more about businesses, health and IT, and less about demography, income and pensions. This is likely motivated by real-world events affecting the users, which are more difficult to implicitly capture from simulated environments, but desirable in order to understand a model's robustness in temporal shift and for real world applications. In the Appendix, Figure 8 displays the differences between the training and test splits for all subjects.

Generalizing to unseen tables   As shown in Figure 3, DPR performs well for tables appearing frequently in the validation split, but poorly in the test split, which could be caused by temporal drift. As expected, tables that were not seen during training resulted in poor recall@5 in either splits. Moreover, the difference in recall between valid and test for infrequent tables could be caused by many potential reasons (learning bias, temporal overfitting,
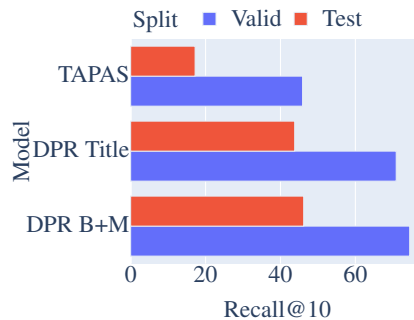
spurious correlation with hidden factors). Thus, future models should aim to close the gap between unseen and frequent tables and within the temporal spectrum of infrequent tables.

French results   In both tasks, we see a drop across all metrics for all models, likely due to the smaller dataset size. Some observations remain valid: temporal drift, poor BM25 performance, and augmentations benefit mT5 for certain metrics. However, others differ: adding member items hurts test results and mT5 performs poorly on *title accuracy*. Modeling details and results can be found in the Appendix A.5.

7   Conclusion

In this paper, we introduce the StatCan Dialogue Dataset, a novel corpus consisting of 20K+ English and French conversations between online visitors of statcan.gc.ca and operators of Statistics Canada. Based on this dataset, we propose two tasks centered on helping users find specific data tables: the table retrieval task and the response generation task. For the table retrieval task, we experiment with various DPR and TAPAS variants, finding that DPR strongly outperforms its TAPAS counterpart, as well as the BM25 baseline. For the response generation task, we investigate fine-tuned T5-large models and explore variants where the input is augmented with table titles from DPR. We find that retrieval-augmented T5 models more frequently return the correct tables, although its title accuracy is still lower than the corresponding recall of the DPR retriever. This result suggests that the generation models struggle to decide when to return a table. We also find that retrieval and generation have difficulty generalizing to future conversations, as our temporal test split revealed a big performance gap

between the validation and test set. All in all, we believe that our tasks pose significant challenges to currently available models and encourage the research community to further explore this dataset and build conversational models that help users of Statistics Canada.

Limitations

Tasks and models limitations The tables in the retrieval task are sourced from statcan.gc.ca, which means that the content is primarily about Canadian demographics[2] and are professionally edited by StatCan employees. Moreover, the generation task is specifically designed to model responses with high fidelity based on retrieved tables, so this task should not be directly used in an unintended or non-research setting (e.g., deploying a virtual assistant) as they pose risks of hallucination that could negatively impact stakeholders. Furthermore, those limitations can be reflected in the models we trained, so we will share those limitations in the model cards (Mitchell et al., 2019) on release.

Environment impact We acknowledge the models in Section 5 used hardware with significant energy consumption. We purposefully chose models of reasonable sizes that can be reproduced on one GPU. Additionally, our hardware is powered by renewable energy.

Artifacts and computational experiments We trained models using libraries based on their intended use and we will release the relevant artifacts following the original licenses. The computational details of the experiments are described in Appendix A.4.

Ethics Statement

Privacy and data access As discussed in Section 1, significant efforts were made to remove Personally Identifiable Information (PII). However, we do not rule out the possibility that certain details could have been missed in that process. Thus, any user that wishes to use the data will need to authenticate and accept the terms of use through an institutional data repository; the terms will require the user to report any instance of PII leak, which will be removed with a dataset update. Additionally, we request any derivative or modifications

---

to be published in the same data repository with the original terms of use and licenses preserved or extended.

Risk of toxicity in online discourse StatCan agents are trained to work with online users in a professional manner. Moreover, since the users access statcan.gc.ca anonymously and virtually, it is more likely to observe toxic online disinhibition (Lapidot-Lefler and Barak, 2012), which could translate to toxicity in users' utterances. Thus, we request dataset users to report any instance of toxicity in conversations, which will be reviewed in the same manner as PII leaks.

References

Vaibhav Adlakha, Shehzaad Dhuliawala, Kaheer Suleman, Harm de Vries, and Siva Reddy. 2022. TopiOCQA: Open-domain Conversational Question Answering with Topic Switching. *Transactions of the Association for Computational Linguistics*.

Raviteja Anantha, Svitlana Vakulenko, Zhucheng Tu, Shayne Longpre, Stephen Pulman, and Srinivas Chappidi. 2021. Open-domain question answering goes conversational via question rewriting. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 520–534, Online. Association for Computational Linguistics.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2018. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. *arXiv:1611.09268 [cs]*. ArXiv: 1611.09268.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

---

[2]More information can be found here: `https://www.statcan.gc.ca/en/subjects-start/population_and_demography`

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.

Harry Bunt, Jan Alexandersson, Jean Carletta, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Kiyong Lee, Volha Petukhova, Andrei Popescu-Belis, Laurent Romary, Claudia Soria, and David Traum. 2010. Towards an ISO Standard for Dialogue Act Annotation. In *Seventh conference on International Language Resources and Evaluation (LREC'10)*.

Harry Bunt, Volha Petukhova, Emer Gilmartin, Catherine Pelachaud, Alex Fang, Simon Keizer, and Laurent Prevot. 2020. The iso standard for dialogue act annotation. In *12th Edition of its Language Resources and Evaluation Conference (LREC 2020)*, pages 549–558. European Language Resources Association (ELRA).

Derek Chen, Howard Chen, Yi Yang, Alexander Lin, and Zhou Yu. 2021. Action-based conversations dataset: A corpus for building more in-depth task-oriented dialogue systems. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3002–3017, Online. Association for Computational Linguistics.

Tianqi Chen, Bing Xu, Chiyuan Zhang, and Carlos Guestrin. 2016. Training Deep Nets with Sublinear Memory Cost. *arXiv:1604.06174 [cs]*. ArXiv: 1604.06174.

Wenhu Chen, Ming-Wei Chang, Eva Schlinger, William Yang Wang, and William W. Cohen. 2020. Open Question Answering over Tables and Text. In *International Conference on Learning Representations*.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question Answering in Context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184, Brussels, Belgium. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Martin d'Hoffschmidt, Wacim Belblidia, Quentin Heinrich, Tom Brendlé, and Maxime Vidal. 2020. FQuAD: French Question Answering Dataset. In

Findings of the Association for Computational Linguistics: EMNLP 2020, pages 1193–1208, Online. Association for Computational Linguistics.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of Wikipedia: Knowledge-Powered Conversational Agents. In *International Conference on Learning Representations*.

Etalab Lab-IA. 2021. DPR CamemBERT trained on French Q&A.

Song Feng, Siva Sankalp Patel, Hui Wan, and Sachindra Joshi. 2021. MultiDoc2Dial: Modeling dialogues grounded in multiple documents. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6162–6176, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Song Feng, Hui Wan, Chulaka Gunasekara, Siva Patel, Sachindra Joshi, and Luis Lastras. 2020. doc2dial: A goal-oriented document-grounded dialogue dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8118–8128, Online. Association for Computational Linguistics.

Timnit Gebru, Jamie H. Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumé, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64:86 – 92.

Wei He, Kai Liu, Jing Liu, Yajuan Lyu, Shiqi Zhao, Xinyan Xiao, Yuan Liu, Yizhong Wang, Hua Wu, Qiaoqiao She, Xuan Liu, Tian Wu, and Haifeng Wang. 2018. DuReader: a Chinese Machine Reading Comprehension Dataset from Real-world Applications. In *Proceedings of the Workshop on Machine Reading for Question Answering*, pages 37–46, Melbourne, Australia. Association for Computational Linguistics.

Charles T. Hemphill, John J. Godfrey, and George R. Doddington. 1990. The ATIS spoken language systems pilot corpus. In *Proceedings of the workshop on Speech and Natural Language - HLT '90*, pages 96–101, Hidden Valley, Pennsylvania. Association for Computational Linguistics.

Jonathan Herzig, Thomas Müller, Syrine Krichene, and Julian Eisenschlos. 2021. Open Domain Question Answering over Tables via Dense Retrieval. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 512–519, Online. Association for Computational Linguistics.

Jonathan Herzig, Pawel Krzysztof Nowak, Thomas Müller, Francesco Piccinno, and Julian Eisenschlos. 2020. TaPas: Weakly Supervised Table Parsing via Pre-training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*,

pages 4320–4333, Online. Association for Computational Linguistics.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

Ali Kabbadj. 2021. French-SQuAD : French Machine Reading for Question Answering. Original-date: 2019-09-22T17:54:45Z.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Rachel Keraron, Guillaume Lancrenon, Mathilde Bras, Frédéric Allary, Gilles Moyse, Thomas Scialom, Edmundo-Pavel Soriano-Morales, and Jacopo Staiano. 2020. Project PIAF: Building a native French question-answering dataset. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5481–5490, Marseille, France. European Language Resources Association.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural Questions: A Benchmark for Question Answering Research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Noam Lapidot-Lefler and Azy Barak. 2012. Effects of anonymity, invisibility, and lack of eye-contact on toxic online disinhibition. *Computers in Human Behavior*, 28(2):434–443.

Harrison Lee, Raghav Gupta, Abhinav Rastogi, Yuan Cao, Bin Zhang, and Yonghui Wu. 2022. Sgd-x: A benchmark for robust generalization in schemaguided dialogue systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, et al. 2021. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics - ACL '04*, pages 605–es, Barcelona, Spain. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 285–294, Prague, Czech Republic. Association for Computational Linguistics.

Marco Lui and Timothy Baldwin. 2011. Cross-domain Feature Selection for Language Identification. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 553–561, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 220–229, New York, NY, USA. Association for Computing Machinery.

Ankur Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. 2020. ToTTo: A controlled table-to-text generation dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1173–1186, Online. Association for Computational Linguistics.

Chen Qu, Liu Yang, Cen Chen, Minghui Qiu, W. Bruce Croft, and Mohit Iyyer. 2020. Open-Retrieval Conversational Question Answering. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, pages 539–548, New York, NY, USA. Association for Computing Machinery.

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):1–67.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8689–8696.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A Conversational Question Answering Challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Radim Rehurek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *In Proceedings of the Lrec 2010 Workshop on New Challenges for Nlp Frameworks*, pages 45–50.

Stephen Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. 2000. The Earth Mover's Distance as a Metric for Image Retrieval. *International Journal of Computer Vision*, 40(2):99–121.

Marzieh Saeidi, Max Bartolo, Patrick Lewis, Sameer Singh, Tim Rocktäschel, Mike Sheldon, Guillaume Bouchard, and Sebastian Riedel. 2018. Interpretation of natural language rules in conversational machine reading. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2087–2097, Brussels, Belgium. Association for Computational Linguistics.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4596–4604. PMLR.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Ellen M. Voorhees. 2001. The TREC question answering track. *Natural Language Engineering*, 7(4):361–378.

Harm de Vries, Dzmitry Bahdanau, and Christopher Manning. 2020. Towards Ecologically Valid Research on Language User Interfaces. *arXiv:2007.14435 [cs]*. ArXiv: 2007.14435.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven

Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. WikiQA: A Challenge Dataset for Open-Domain Question Answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2013–2018, Lisbon, Portugal. Association for Computational Linguistics.

Pengcheng Yin, Graham Neubig, Wen-tau Yih, and Sebastian Riedel. 2020. TaBERT: Pretraining for Joint Understanding of Textual and Tabular Data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8413–8426, Online. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text Generation Evaluating with Contextualized Embeddings and Earth Mover Distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578, Hong Kong, China. Association for Computational Linguistics.

A Appendices

A.1 Complete User Intents

To provide insight into what kind of help is offered by StatCan's live chat, we qualitatively analyze the conversations and highlight examples of the main user intents below.

Finding a table The main intent of the chat functionality is to help users find specific data tables. For example, one user was looking for the population numbers in certain regions of Montreal for 2012-2016. This intent is the focus of our work.

Obtaining meta-information of table Instead of finding data tables some users are interested in meta information of a specific table. For example, one visitor wanted to know when the next Census is released. Another user was interested in understanding the definition of Workforce Availability (WFA) and Labour Market Availability (LMA).

Manipulating a data table Some users would like to obtain the data tables in a different format or representation. For example, one user was looking at a specific data table and asked if they can see annual instead of monthly values.

Help with user interface Some users are looking for help with the user interface. For example, one user wanted to download a specific data table but they were unable to find the download link.

Out of domain requests We find many conversations that are outside of the scope of StatCan's live chat. For example, some user asked what documentation needs to be provided to ship a specific product to a foreign country.

The first intent is covered in Table 1, and subsequent intents are in Table 18. While we believe all intents are interesting directions for dialogue research, we focus on the table retrieval intent because (i) there are many conversations available for them and (ii) there is a clear measure of task success i.e., whether the correct table is retrieved. Throughout the rest of this paper, we work with conversations where the agent returns a table URL.

A.2 Supplementary Table Description

A.2.1 Formatting and size

Full tables are stored in a *long* format, where each column corresponds to a dimension, except the last one which corresponds to the value. The number of rows corresponds to the Cartesian product of the dimensions' members. This means that for n_i members across d dimensions, we have a total of $\prod_{i=0}^{d} n_i$ rows in the full table. When a table has many member items, the full table can become extremely large. Moreover, tables shown on the statcan.gc.ca are usually a pivoted and filtered view of the full table, which means certain members will become columns, others will become rows, and many are simply omitted.

A.2.2 Detailed Specifications

This section provides supplementary details for Section 3.3.

Product ID (PID) Unique 8 to 10-digit identifier given to each published data table. Although other types of tables might be shared by the agent, the PID will always be given for the official data tables; as a result, any table that does not have a PID in the URL is not considered for this task. The first two digits (1-2) represent the code of the subject associated with the table (this can be found in the basic information), then 3-4 represent the product type, which in our case are tables and are common coded as "10". Digits 4-8 is a unique identifier representing that table for the given subject. Digits 9-10 indicate the view of the table and will commonly be "01", which represents the default view; they are needed when constructing the URL but they are otherwise optional, and are omitted in the released dataset.

Member items Labels for individual tables, and can either be represented as a column or a row index through a pivot operation. Each table will have member items different from other tables.

Dimensions Non-overlapping sets of member items; each member item must belong to a dimension. For a table with d dimensions, each data value is associated with a single d-tuple in the d-fold Cartesian product of the dimensions, but not every tuple in the product will have an associated data value (if it was not recorded or if the tuple is invalid).

Basic information Metadatum consisting of the title, the date range, the frequency, the dimensions, the subject category, and the survey[3] from which the data was sourced.

Footnotes Unstructured comments often included with a table if supplementary details need to be given; usually, those notes are associated with a specific member item that requires more explanations.

Full table For all tables with a PID, the full tables (as a CSV file), their complete metadata, and their basic information are made available as part of the data release and can be used in our proposed tasks. A sample table can be found in Section 3.3.

---

[3]The list of surveys and link to detailed information can be found at https://www.statcan.gc.ca/en/survey/list

A.2.3 Table updates and archives

Tables on Statcan will sometimes be updated regularly, whereas in other times they may only be released once. When they stop being updated and the information becomes outdated, they will be marked as archived. In some cases, a new version of an archive table may be created with substantial changes (such as new columns). As shown in Table 14, tables are released without a predetermined schedule, but has been on average increasing since 2000, with major peaks in 2017, which was caused by many health-related tables, and 2021, which was caused by the release of many labour, science, and income tables, as shown in Figure 6. Moreover, Figure 5 shows that a majority of tables released between 2019 and 2021 are still up-to-date (current), whereas most of the tables before then have been archived.

A.3 Supplementary Statistics

A.3.1 Frequently requested tables

The most frequent tables are summarized in Table 21. Whereas 2 of them are sourced from the consumer price index survey (commonly used to track inflation), the 4 other tables cover more general and broad subjects like income, demography, business performance and crime. Each of those 4 tables are sourced from different surveys. One table is updated monthly, another semi-annually, and the rest are updated annually. The oldest table was updated in 2019, which is when the conversations started being recorded.

A.3.2 Table Frequency Statistics

In tables 3 and 12, we can calculate that a table is returned on average 5.68 times (with standard deviation of 12.86) in English conversations and 4.25 times (standard deviation of 9.73) in French conversations. Thus, the most requested tables are disproportionately represented compared to less popular tables, and there's a very high variance in the number of time a table is used. In fact, there are 294 tables that appear only once in either splits (i.e., 28.9%).

A.3.3 Fine-grained Conversation Statistics

In Table 13, we observe that the number of messages and turns will vary significantly around the mean, with over 68% conversations lasting between 2 and 7 turns. In extreme cases, a conversation can last up to 28 turns. Moreover, we also notice that, although most messages will have around 32 tokens, the longest message can have up to 1374 tokens; in those scenarios, we will see agents write a large body of text, and sometimes also copy and paste large amount of text (for example, from a database of templates) when responding to the user. Although those are usually sent in multiple consecutive messages within a turn, they may decide to send everything all at once.

A.4 Implementation Details

This section provides the details for implementing the models in Section 5.

Implementing transformer models All models based on the transformer architecture (Vaswani et al., 2017) were implemented using Hugging-Face's library (Wolf et al., 2020).

BM25 To facilitate reproducibility, we implemented the model in Gensim (Rehurek and Sojka, 2010).

DPR and TAPAS-NQ We used the base variant of DPR and the large variant of TAPAS-NQ. We use the DPR checkpoints that were trained on 320K questions from Natural Questions (Kwiatkowski et al., 2019) (NQ). During training, the networks were optimized with AdamW (Loshchilov and Hutter, 2017) at a learning rate of 10^{-5} and zero weight decay. Based on the original work, the networks were trained for 30 epochs, with a batch size of 64 queries, positive passages, and hard negative passages (the latter are retrieved with BM25). Negative in-batch sampling was used to increase negative examples. To ensure reproducibility, the networks were trained on a single 32GB GPU and used gradient checkpointing (Chen et al., 2016) to reduce memory usage. The conversation lengths was 512 tokens, and the metadata token lengths were 128 for title, 256 for basic information (defined in Section 3.3), and 512 for the rest.

T5 We used an Adafactor optimizer (Shazeer and Stern, 2018) with a learning rate of 0.001. We used batch sizes of 16 with 8 steps of gradient accumulation and gradient checkpointing to reproduce the batch size of 128 samples in the original implementation. The models were trained on a single 32GB GPU for 10 epochs. The source and target lengths were respectively 512 and 256 tokens, where the source was truncated from the right to ensure that the latest messages remained after truncation. We

used a beam size of 4 and length penalty of 0.6 following the original implementation.

Training time For the English split, each variant of DPR can be trained in 68 minutes on a A100 GPU. Each large variant of TAPAS-NQ takes 15h to train on a V100 GPU. Each variant of T5 can be trained in 16h on a V100 GPU. All DPR results can be reproduced in 8h, TAPAS in 90h, and T5 in 96h, and proportionally less time would be needed for the French split.

A.5 Modeling the French subsets

Basic statistics The number of messages by conversation varies between 2 and 59 with a median of 11 for the English split (see Figure 7 for the distribution). Based on Table 12, there's on average 3.9 turns but 12.3 messages. On average, there are over 30 tokens for each message (using the T5 tokenizer).

Language splitting In order to determine the language of each conversation, we used two popular language identification libraries: *langid.py* (Lui and Baldwin, 2011) and a fasttext network fine-tuned for language detection (Joulin et al., 2017). After apply the models on every conversation, we only retain the conversations with matching language labels (both English or both French).

Training and evaluation The training procedure and evaluation on the French subsets follow exactly the tasks specified in Section 4.

Modeling response generation Instead of T5, we used the multilingual T5 model by Xue et al. (2021) as it naturally handles text in French.

Modeling retrieval We used a variant of DPR derived from CamemBERT (Martin et al., 2020) and trained on three French Q&A datasets (Keraron et al., 2020; d'Hoffschmidt et al., 2020; Kabbadj, 2021) by Etalab Lab-IA (2021).

Retrieval results In Table 17, we observe that, unlike the English split, adding member items to the basic information or to the title improves validation results but not test results, which likely indicates overfitting. However, we notice a high variance between the runs, which makes it difficult to determine whether member items is helpful. Both overfitting and high variance are likely caused by the smaller size of the training set. Moreover, BM25 perform extremely poorly on any metadata view, which can also be linked to the dataset size.

Generation results In Table 33, we notice a significant decrease across all metrics, with the title accuracy being consistently 0%. This is likely because the French split is significantly smaller, yet remains as complex as the English split, which becomes challenging for mT5 to model. In the case of title accuracy, we found 55 instances in the French test split where the title is in the target text (i.e., returned by an agent). However, in 54 cases, the augmented mT5 returned a generic reply (e.g., "Veuillez patienter pendant que j'effectue une recherche.") instead of the expected title, which indicates that mT5 is incapable of determining when it is relevant to return a title and can't generate non-templated responses.

A.6 Responses generated by T5

In this section, we select a few conversations from the validation set and examine the messages generated by T5 and T5 augmented with DPR-retrieved titles (T5+D).

Common and uncommon responses In Table 23, we notice that both T5 and T5+D are capable of generating common speech acts like "Thank you" and "Please wait...", but struggles when faced with an unfamiliar situation (having to ask for clarification for a user that has been accidentally disconnected).

Common table In Table 24, among the retrieved tables (R_i) titles, the first one was partially correct. Both T5 and T5+D extended the title and also output the desired ID, matching the expected agent's response, which is one of the most popular table in the training set (see Table 21).

Multiple tables, date selection In Table 28, we notice that T5 only returns one of the two tables that the agent returned. On the other hand, the correct tables were retrieved by DPR, but T5+D failed to select the ones with the correct dates (it selected June 2019 instead of December 2019) but the select were otherwise relevant.

Verbosity of explanations T5+D additionally provided a paragraph of explanation while linking to relevant resources, both on the StatCan website (non-tabular) and external resources. This is because T5+D memorized this information during training, and simply replaced the tables with the updated dates (Table 27).

Uncommon tables In the conversations shown in Tables 25, 26, 29, the retrieved table appears 10 times in the training set, which is significantly less common than the table retrieved in Table 24. For Table 25, the table returned by the agent was not retrieved by DPR, leading to T5+D returning the first title retrieved. However, in Table 26, the correct title was retrieved by DPR (title #4), yet T5+D failed to use that correct title in the generated message. As for Table 29, the agent gave a hint by stating "As a standard product, we have tables about employment by industry", which was correctly acknowledged by DPR as the second retrieved title perfectly matches the PID of the table in the agent's response. However, T5+D fails again at selecting the correct title, instead opting to return a generic response ("Please hold while I find the information"), and T5 hallucinates a PID that is different from the title it generated (both of which are wrong).

Tables unseen during training Among the tables that do not appear in the training set (see Table 11 for more information), there are seven that appear 3 or more times in the validation or test sets (Table 32). In Table 30, we see a conversation where DPR retrieves the correct title, which is correctly returned by T5+D, whereas T5 fails to return it. On the other hand, when DPR also correctly retrieves the title in Table 31, T5+D fails to return it, as it was likely mislead by the agent saying "Unfortunately,...".

A.7 Statistics Tests

To back the claims in Section 6, we performed multiple single-tailed Welch t-tests, using the mean and corrected standard deviation from 16. The null hypotheses are that means of experiments A are different from the means of experiments B, across 3 runs. Unless otherwise specified, we use R@1 on the test split.

1. Claim: Adding basic information and member items to title results in a significant difference for DPR. With A being the model using only title, and B using basic + member, our p-value is 0.014.

2. Claim: For DPR, using member item result in drastic decrease. With A being the model using only member and B using title + member, our p-value is 0.00086.

3. Claim: TAPAS-NQ performs better with title and member items compared to the full table. With A being the model using the full table, and B using title + member, our p-value is 0.016.

4. Claim: In Figure 4, the validation recall@10 are higher than the test split for TAPAS, DPR Title and DPR Basic + member. With A being the validation score and B the test scores, the p-values are respectively $0.00197, 2.18 \times 10^{-5}, 0.00014$.

2814

B Dataset Card

This section presents a dataset card that follows the format proposed by Lhoest et al. (2021), which was inspired by Mitchell et al. (2019) and Gebru et al. (2021).

Summary The StatCan Dialogue Dataset consists of over 20K+ conversations between agents working at Statistics Canada (StatCan) and users who are visiting StatCan's website and need support via the official live chat system.

Tasks A subset of 19K conversation turns is used to build two tasks:

1. Automatic retrieval of relevant tables based on a on-going conversation. For each partial conversation, the task is to return the ID of the most likely table returned by an agent. This is evaluated using the recall@k metric.

2. Automatic generation of appropriate agent responses at each turn. For each partial conversation, the task is to return the most likely response by an agent, including link to a relevant table. This is evaluated using four metrics described in Section 4.2.

Leaderboard The leaderboard and submission instructions can be found on the project webpage. Each submission will be accompanied with a tag indicating if:

- It was self-reported;
- The submissions were externally evaluated;
- The inference was reproduced following provided instructions;
- The complete training process was independently reproduced.

Languages The conversations were held in Canadian English (en-CA) and Canadian French (fr-CA).

B.1 Dataset Structure

B.1.1 Data Instances

Conversation A full example of a conversation can be found in Table 19. Instances for each user intent can be found in Table 18, and two conversations with annotated dialogue acts can be found in Table 22. For our case study in A.6, we show partial conversations in Tables 24, 25, 26, 27, 28, 29, 30, 31.

Tables The complete metadata of a table can be found in Table 20, which can be access at doi.org/10.25318/3210035401-eng. Table 21 shows the basic information for the most popular tables.

B.1.2 Data Fields

Full dataset A CSV file with the following fields is provided:

- `conversation`: The partial conversation (before a table is returned) in JSON format.

- `conversation_index`: A unique index that serves at identifying the conversation outside of this task.

- `conversation_processed`: The conversation converted into a readable text format, with extra information (such as timestamp) removed, the URLs replaced with a special tag, and separation tags (`</s>`) added.

Retrieval task CSV files with the following fields is provided for each split:

- `conversation`

- `conversation_index`

- `conversation_processed`

- `target_pid`: The product ID of the table that is returned by the agent

- `language`: The language reported by the live chat system, which may not always be accurate due to mislabeling.

- `ft_detected_lang`: The language predicted by fastText.

- `ft_detected_lang`: The score output by fastText.

- `lid_detected_lang`: The language predicted by langid.py.

- `lid_detected_prob`: The score output by langid.py.

Metadata The metadata that was used during retrieval is provided as a CSV file with the following fields (one for each of the 5907 tables):

- `pid`: The product ID of the table

- `title`: The title of the table

- `basic_info`: The basic information in the textual format

- `member_info`: The member items enumerated as text

- `footnote_info`: The footnotes enumerated as text

- `full_info`: The basic information, member items and footnotes all in a single entry

- `x_and_y`: An combination of two items above, for example `x=title` and `y=footnotes`.

- `*_fr`: All of the above are also available in French, indicated by the suffix `_fr`.

Generation task CSV files with the following fields is provided for each split:

- `source`: Equivalent to `conversation`.

- `source_processed`: Equivalent to `conversation_processed`.

- `target`: The message written by the agent following the conversation.

- `target_processed`: The message written by the agent following the conversation, with URLs replaced with a special tag.

- `conversation_index`

An augmented variant of each CSV file for the conversation task is provided with all of the above as well as the following fields:

- `source_augmented`: The same content as `conversation`, appended with the title of the top-5 tables retrieved by best DPR variant trained on the basic information.

- `target_augmented_1`: The same content as `source_augmented`, but using only the first table instead of top-5.

Conversation JSON formatting A conversation follows the following JSON format:

```
[
  ...
 {'timestamp': '13.03.2019 17:03:22',
  'actor': 'user',
  'actor_name': '<NAME>',
  'text': "I'll take a look at that",
  'urls': []},
 {'timestamp': '13.03.2019 17:04:12',
```

```
  'actor': 'operator',
  'actor_name': 'Kelly C',
  'text': 'Building permits...',
  'urls': ['https:...']},
  ...
]
```

B.1.3 Data Split

The retrieval splits has the following number of samples:

- Train: 3782 (en); 869 (fr)

- Validation: 799 (en); 201 (fr)

- Test: 870 (en); 141 (fr)

They correspond to the number of queries in Table 3 and Table 12 because each query results in a table being retrieved, which can happen more than once in a conversation. In such cases, the partial conversations will be truncated at different turns in the conversation.

The generation splits have the following number of samples:

- Train: 21582 (en); 3977 (fr)

- Validation: 4464 (en); 861 (fr)

- Test: 4850 (en); 884 (fr)

There are fewer samples than the number of messages in Table 3 and Table 12 because the goal of the task is only to predict the messages that will be written by the agent. Just like the retrieval task, the partial conversations will be truncated at different turns in the conversation.

B.2 Dataset Creation

B.2.1 Curation Rationale

Section 1 extensively motivates the curation of the dataset. To summarize, we enumerate the major points:

- Data from real users: We wanted a dataset that captures the linguistic challenges that exist in the real world

- Task-oriented dialogue: We wanted tasks with the specific goal of helping live chat users in their search of statistics.

- Real-world applications: Our model can be directly applied other statistics offices that want to set up a chat system, and our dataset will be useful for any organization that has (1) a chat system, and (2) a database of tables.

- Multilingual dialogue: We wanted to build models that can handle languages beyond English, which is why we also offer French versions of our tasks.

B.2.2 Source Data

Conversations The data was retrieved from the live chat system on statcan.gc.ca, which was anonymized by the development team at StatCan. The conversations happened between March 2019 and March 2021.

Tables The tables are publicly available and were downloaded following the instructions in the Web Data Service user guide: statcan.gc.ca/en/developers/wds/user-guide. The tables were released by Statistics Canada between 2000 and 2021. The data are either collected directly by Statistics Canada (e.g., through a census or a survey) or were compiled from existing sources (such as private sector organizations and government agencies) into official statistics. Existing sources include:

- Administrative data: Collected by government or the private sector as part of ongoing operations, and include records of birth and death, taxes, border control, and satellite data.

- Microdata Linkage: Existing information is linked to create new data. The existing information may not always be available publicly (for privacy purposes), thus linkage could add new information that was previously unavailable, while protecting the confidentiality of the public.

- Open data: Machine-readable and freely available data sourced from various channels (e.g., OpenStreetMap).

- Web scraping: Data from the internet that were scraped by Statistics Canada (this excludes personal information and "any information that will not be used to produce statistical output").

B.2.3 Annotations

The dataset does not contain any additional annotations beyond the ones collected through the live chat system and included in the table metadata.

B.2.4 Personal and Sensitive Information

Personal and sensitive information were removed programmatically from the conversations, and officially published tables only contain aggregated information that preserve the confidentiality of the participants. Although the removal process is highly advanced, there is a non-zero chance that some information can be used to reconstruct the profile of a user. For this reason, the access to the data will require researchers to sign-up and agree to the terms of use, and any derivative must be shared on the same platform and include the same terms.

B.3 Considerations for Using the Data

B.3.1 Social Impact of Dataset

The purpose of this dataset is develop and evaluate models that can assist knowledge workers in finding relevant tables from a data source. By providing a specialized retrieval system capable of returning more relevant results compared to general purpose, the productivity of the knowledge workers can be increased. For public agencies and statistics offices, this would benefit many live chat users interested in statistics related to a certain community.

B.3.2 Discussion of Biases

As discussed in Section 7, there are always risks of toxicity in online discourses, which means that the live users may exhibit negative biases in their messages. However, the StatCan agents are tasked to communicate with online users in a professional manner. Thus, researchers should not use this dataset to build models that generate messages written by live users, and any model trained on the dataset should not be used in scenarios where biases can negatively impact stakeholders.

B.3.3 Other Known Limitations

Section 7 describes other known limitations.

B.4 Additional Information

B.4.1 Dataset Curators

The dataset was curated by the authors of this paper based on the original data collected and processed by StatCan developers and agents.

B.4.2 Licensing Information

The conversations use a custom license, which needs to be accepted by researchers interested in accessing the conversation. The tables are released under the the Statistics Canada Open Licence: statcan.gc.ca/en/reference/licence.
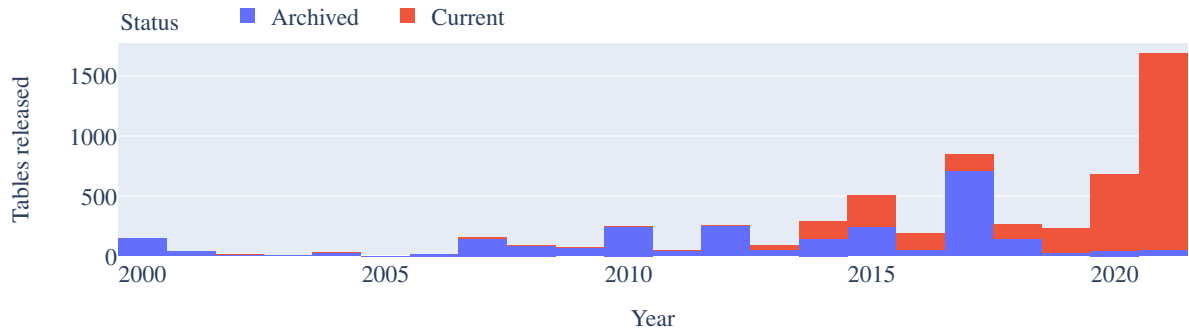
Figure 5: The release year of all tables available on statcan.gc.ca
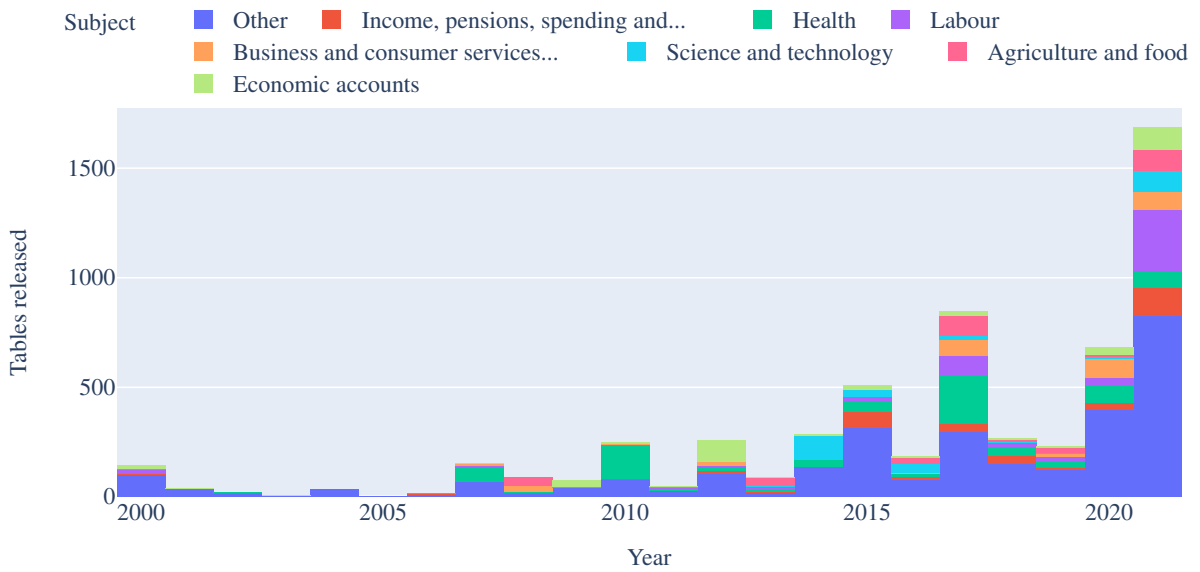

Figure 6: The release year of all tables by subject. Only the top 8 subjects are shown for readability.
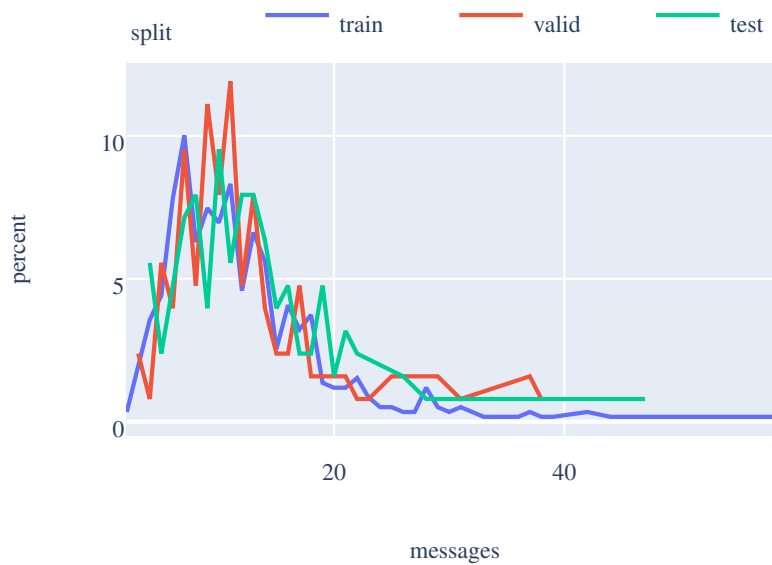

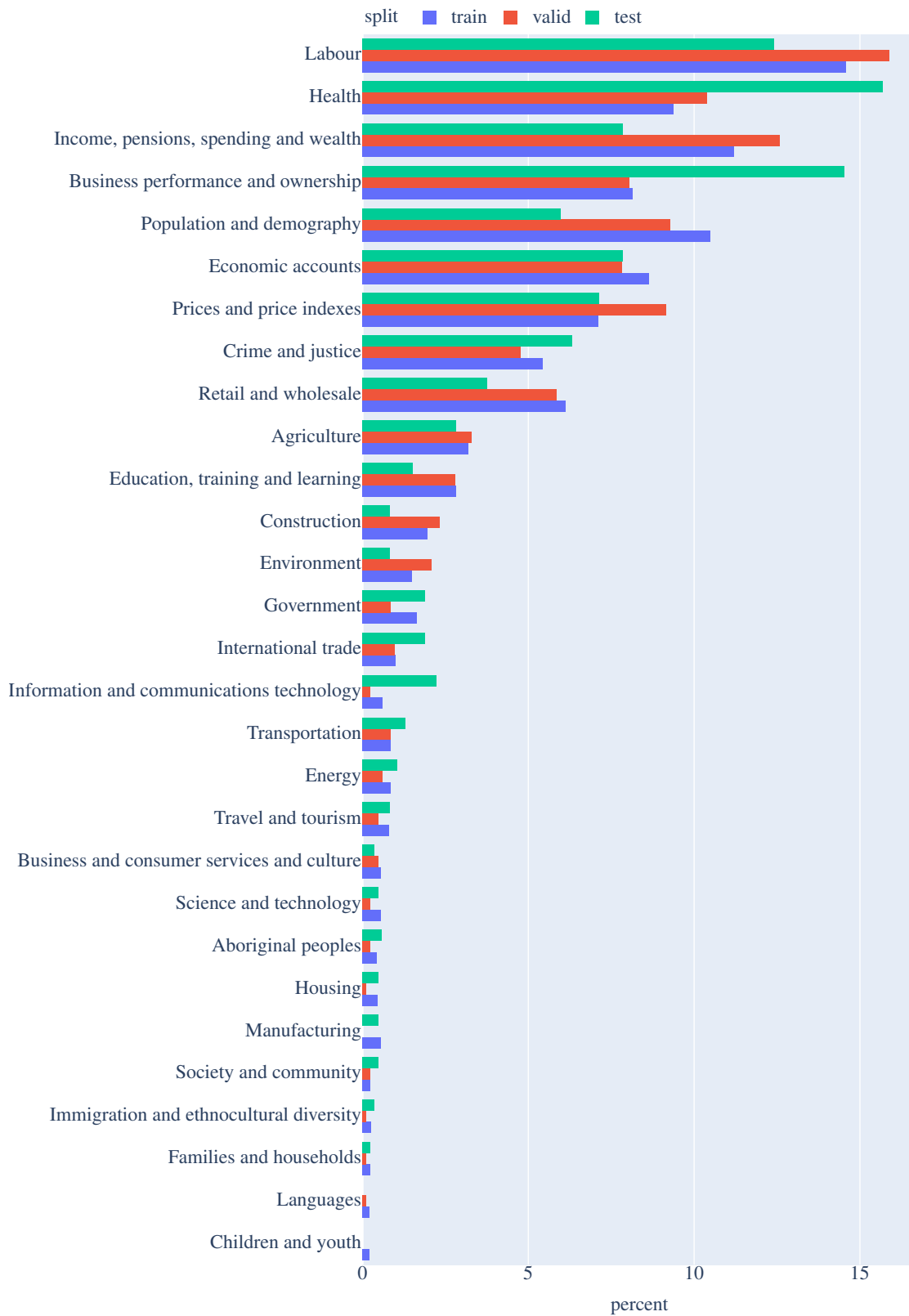Figure 7: Histogram of messages by conversation in the French task splits. English split in Figure 1.

Figure 8: Distribution of subject categories for tables in the retrieval task, colored by the split.

	Train	Valid	Test	Overall
Train	0	470	535	0
Valid	41	0	199	0
Test	145	238	0	0
Overall	181	610	571	0

(a) Differences

	Train	Valid	Test	Overall
Train	778	308	243	778
Valid	308	349	150	349
Test	243	150	388	388
Overall	778	349	388	959

(b) Overlaps

	Train	Valid	Test	Overall
Train	0	154	184	0
Valid	18	0	64	0
Test	44	60	0	0
Overall	61	197	201	0

(c) Differences (French)

	Train	Valid	Test	Overall
Train	224	70	40	224
Valid	70	88	24	88
Test	40	24	84	84
Overall	224	88	84	285

(d) Overlaps (French)

Table 11: Number of tables (a,c) differing and (b,d) overlapping between each split (subset used for both tasks). The difference is computed as row - column. Summarized results in Table 3.

Dataset	Train	Valid	Test	All
Max Tokens / Msg	1243	857	912	1243
Max Turns / Conv.	28	18	21	28
Max Msg / Conv.	78	58	73	78
Max Qs / Conv.	15	11	11	15
Std Tokens / Msg	1	1	1	1
Std Turns / Conv.	2.58	2.36	2.62	2.56
Std Msg / Conv.	8.53	7.55	8.98	8.47
Std Qs / Conv.	1.27	1.24	1.18	1.25
Max Tokens / Msg	1374	1373	352	1374
Max Turns / Conv.	15	10	15	15
Max Msg / Conv.	59	38	47	59
Max Qs / Conv.	15	15	5	15
Std Tokens / Msg	1	2	2	1
Std Turns / Conv.	2.21	2.08	2.1	2.17
Std Msg / Conv.	7.02	6.74	6.61	6.91
Std Qs / Conv.	1.84	2.18	0.77	1.79

Table 13: More statistics (max and standard deviation) at the conversation level (top: English, bottom: French), following Table 3.

Dataset	Train	Valid	Test	All
# Conv.	562	122	109	793
# Turns	2147	472	439	3058
# Messages	6807	1492	1389	9688
# Queries	869	201	141	1211
# Tokens / Msg	47.25	49.86	40.14	46.63
# Turns / Conv.	3.82	3.87	4.03	3.86
# Msg / Conv.	12.11	12.23	12.74	12.22
# Queries / Conv.	1.55	1.65	1.31	1.53
# Tables	224	88	84	285
# New tables	0	18	44	61
# Dims / Table	3.4	3.5	3.5	3.4
# Mbrs / Table	205.8	249.9	164.1	190
# Notes / Table	20.3	26.3	23.4	21

Table 12: Statistics for the retrieval and generation tasks for the French split. See Table 3 for the English split.

Year Released	# Archived	# Up-to-date
2000	145	0
2001	39	0
2002	14	5
2003	8	0
2004	31	2
2005	1	0
2006	18	0
2007	142	11
2008	85	7
2009	72	5
2010	243	8
2011	45	3
2012	248	8
2013	49	40
2014	142	145
2015	247	264
2016	52	134
2017	710	136
2018	148	122
2019	27	203
2020	43	640
2021	52	1633

Table 14: Release of Statcan tables over the years

Merged Acts	Count	Original Speech Acts
answer	50	answer, confirm, disconfirm
request	31	request, acceptOffer, declineOffer
timeManagement	28	stalling, pausing
inform	63	inform, agreement, disagreement, correction
infoSeekingQuestions	25	propositionalQuestion, checkQuestion, setQuestion, choiceQuestion, directQuestion
promise	18	promise, acceptRequest, declineRequest
autoFeedback	25	positiveAutoFeedback, negativeAutoFeedback
offer	11	offer
instruct	18	instruct
clarificationQuestion	16	clarificationQuestion
greeting	28	initialGreeting, returnGreeting
selfIntroduction	17	initialSelfIntroduction, returnSelfIntroduction
thanking	47	thanking
acceptThanking	15	acceptThanking

Table 15: Merged and original speech acts occurring in 100 turns in conversations from the validation set. This table is summarized in Table 4.

Model	Encoded	Valid R@1	R@10	R@20	Test R@1	R@10	R@20
DPR	Basic	37.8 ± 1.82	73.1 ± 1.02	80.8 ± 0.51	14.7 ± 1.13	45.0 ± 0.24	55.0 ± 1.67
	Basic + footnote	40.0 ± 0.64	73.3 ± 0.85	81.3 ± 1.19	13.9 ± 0.37	44.4 ± 2.73	54.2 ± 2.65
	Basic + member	40.0 ± 0.44	74.5 ± 1.38	81.6 ± 0.55	15.7 ± 0.7	46.2 ± 0.44	56.3 ± 0.58
	Member	28.0 ± 0.47	62.1 ± 1.07	71.4 ± 1.06	10.7 ± 0.88	35.0 ± 0.57	46.3 ± 0.83
	Title	37.8 ± 0.21	70.9 ± 0.67	78.2 ± 1.18	13.9 ± 0.59	43.8 ± 1.22	53.4 ± 1.42
	Title + footnote	39.5 ± 0.69	73.2 ± 0.81	80.8 ± 0.47	13.3 ± 0.7	43.6 ± 0.46	53.6 ± 1.5
	Title + member	38.6 ± 1.19	72.1 ± 0.92	79.4 ± 0.72	14.3 ± 0.59	45.1 ± 0.44	54.2 ± 1.07
TAPAS	Table	16.9 ± 4.26	45.9 ± 5.46	55.5 ± 3.13	3.9 ± 0.65	17.2 ± 2.83	24.6 ± 3.45
	Title	21.5 ± 0.33	54.7 ± 3.39	63.7 ± 1.38	4.7 ± 0.53	20.2 ± 0.85	28.7 ± 1.98
	Title + member	24.6 ± 3.6	54.9 ± 3.34	63.7 ± 2.74	6.1 ± 0.98	22.1 ± 2.66	31.5 ± 3.02
TAPAS-NQ	Table	25.3 ± 3.46	60.0 ± 2.43	70.0 ± 2.89	5.3 ± 0.8	22.8 ± 2.32	32.0 ± 2.48
	Title	26.5 ± 1.85	64.4 ± 1.95	71.6 ± 2.32	6.3 ± 0.64	27.4 ± 1.84	36.1 ± 2.53
	Title + member	29.8 ± 2.74	62.9 ± 1.35	71.5 ± 0.63	7.4 ± 0.73	30.0 ± 1.46	39.3 ± 0.86
BM25	Basic	1.0	7.5	10.3	1.0	7.2	9.8
	Basic + footnote	0.9	2.4	3.1	0.9	2.8	3.9
	Basic + member	0.1	1.3	2.6	0.2	2.4	3.8
	Full	0.3	1.3	1.9	0.2	1.1	2.0
	Title	2.0	10.3	13.8	2.5	10.9	14.0
	Title + footnote	0.8	2.3	3.1	0.7	2.9	4.0
	Title + member	0.0	1.4	2.4	0.3	2.3	3.8

Table 16: Full retrieval results for the English splits. The values reported are in recall % at k. DPR and TAPAS were run 3 times and averaged (standard deviation given after ±). Selected results in Table 8.

Model	Encoded	Valid R@1	R@10	R@20	Test R@1	R@10	R@20
DPR	Basic	19.7 ± 1.52	57.4 ± 2.55	67.0 ± 5.01	12.5 ± 1.78	40.0 ± 1.08	48.5 ± 4.27
	Basic + footnote	23.2 ± 3.8	55.4 ± 1.88	66.8 ± 1.04	13.9 ± 1.48	35.9 ± 0.82	49.9 ± 3.91
	Basic + member	22.7 ± 4.35	57.9 ± 2.45	67.7 ± 3.48	12.8 ± 2.13	38.1 ± 2.17	47.0 ± 1.78
	Title	23.7 ± 2.92	58.5 ± 2.24	69.8 ± 2.35	14.7 ± 2.28	35.7 ± 3.2	48.2 ± 2.84
	Title + footnote	23.2 ± 2.74	54.7 ± 3.59	64.7 ± 2.28	14.4 ± 1.48	36.6 ± 2.69	45.4 ± 2.13
	Title + member	24.2 ± 1.52	55.4 ± 4.51	65.3 ± 2.01	14.4 ± 2.28	39.5 ± 0.82	48.0 ± 2.05
BM25	Basic	0.0	0.0	0.0	0.0	0.0	0.0
	Basic + footnote	0.0	0.0	0.0	0.0	0.0	0.0
	Basic + member	0.5	1.0	1.0	0.0	0.7	0.7
	Full	0.0	0.0	0.5	0.0	0.0	0.0
	Title	0.0	0.5	0.5	0.0	0.0	0.7
	Title + footnote	0.0	0.0	0.0	0.0	0.0	0.0
	Title + member	0.0	0.5	1.0	0.0	0.7	0.7

Table 17: Full retrieval results for the French splits. The values reported are in recall % at $k \in \{1, 10, 20\}$. DPR was run 3 times and averaged (standard deviation given after ±).

	Obtaining meta-information of table
U:	Hi I have a question on the date of new census
A:	Hello my name is Graham, how may I help you?
U:	Hi Graham
U:	I am currently looking at : Census of Canada and National Household Survey
U:	I see that the last census were made in 2011 and 2016
U:	should I expect a new census in the coming month in 2020 ?
A:	no - the next Census is going to be taking place in May 2021
A:	it's every 5 years
U:	Ok perfect thank you
A:	you're welcome! have a good day
	Help with user interface
U:	Hello Graham - is there way I can see total annual values instead of monthly?
U:	for - https://www150.statcan.gc.ca/t1/tbl1/en/cv!recreate.action?...
A:	please hold - I will see if a table like this with an annual frequency is available
U:	wonderful thank you
A:	unfortunately no, the only active table on construction investment is the monthly one I sent you - there is an inactive quarterly table Investment in Building Construction (https://www150.statcan.gc.ca/n1/en/surveys/5014)
A:	it is also possible to download the monthly data in .CSV form, add together the months in a spreadsheet to get annual totals
A:	or if you'd prefer, you can make a custom request to our analysts (a fee may apply).Please send us your request, clearly indicating the variables, geographic areas, time frame and data frequency (i.e., annually, monthly or quarterly) needed at infostats@canada.ca (mailto:infostats@canada.ca) and refer to Case number: 986005 so we can access the transcript of this live chat. Your request must include your name, address, postal code, and phone number (please do not share your personal information on the live chat.)An agent will get in touch with you to provide an estimate and proceed with the payment, if applicable.You may also submit your request by telephone at 1-800-263-1136.
U:	thats fine! thank you very much :)
U:	very helpful
U:	:)!
U:	have a good day!
A:	glad I could help! you too
	Help with user interface
U:	Hello, Do you know where I can download the dataset for General Social Survey Cycle 32?
A:	Please wait while I try to find this information for you.
U:	Thank you
U:	(I'm referring to this one: https://www23.statcan.gc.ca/imdb/p2SV.pl?...)
A:	General Social Survey: Caregiving and Care Receiving, Public Use Microdata File (https://www150.statcan.gc.ca/...)
U:	I'm sorry, but im having trouble finding the download link for cycle 32 (2018)
A:	you have to click on 'More information' and that will bring you to the order form
U:	ohhhh, I see
U:	thank you for letting me know
U:	thank you for your help!
A:	you're welcome - happy to help!
	Out of domain requests
U:	I'm wondering how to obtain a copy of an autopsy that's been done for a family member. Doctor office says they don't get the autopsy report.
A:	Statistics Canada does not issue certificates for birth, death or marriage. Civil registration is a provincial jurisdiction. Therefore, you must contact the government ministry, registry or vital statistics office of the province or territory in which the birth, death or marriage occurred. To obtain the contact information for your provincial or territorial government, please call Service Canada at 1-800-O-Canada (1-800-622-6232) or visit the following page: Vital Statistics contacts (https://www.statcan.gc.ca/eng/health...) .
A:	also: Coroner's report (https://spvm.qc.ca/en/...)

Table 18: Examples of user intents described in Section A.1.

U:	Hi, I'm looking to obtain quarterly data in regards to GDP grow (Canada), BC Housing STarts, Canada Oil Price/BBL
A:	Hello, my name is Kelly C. Give me one moment as I search
U:	thanks!
A:	For GDP growth rates, please consult the following link: Gross domestic product (GDP) at basic prices, by industry, monthly, growth rates
A:	I'll continue searching for the rest. One moment
A:	What do you mean by BC Housing Starts?
U:	I'm required to research all of the housing starts for BC on a quarterly basis
A:	Define "starts".
U:	unit basis
U:	Housing starts are the number of new residential construction projects that have begun during any particular month
A:	Perfect. Give me one moment.
A:	I'm still searching. Sorry about the wait.
U:	no worries
A:	I would have monthly data regarding new building permits being issued.
U:	I'll take a look at that
A:	Building permits, by type of structure and type of work (https://www150.statcan.gc.ca/t1/...)
A:	Do play with the variables to get what you are looking for.
U:	thank you
A:	To view/ manipulate the variables available in a data table: • Click on "Add/Remove data" • Select at least one variable in each tab and click on "Apply" in order to view your customized table To download the data: • Click on "Download" • Select one of the three output formats then click on the hyperlink and save the table The following video may also be helpful to you: How to use the data tables (https://www.statcan.gc.ca/eng/sc/video/howto) .
A:	I'll have a look for oil prices. One moment.
U:	Do you also have data to Canada'as oil Price/BBL ("WTI")?
U:	okay thanks!
A:	Are you looking for the retail prices of oil?
A:	If so, I found some data for smaller geographies.
A:	Monthly average retail prices for gasoline and fuel oil, by geography (https://www150.statcan.gc.ca/t1/...)
U:	yes
A:	Would those geographies be enough?
A:	Or are you looking for Canada only?
U:	hmm
U:	I would need something that pertains more to all of canada
A:	What about this? Monthly average retail prices for food and other selected products (https://www150.statcan.gc.ca/t1/...)
A:	Is there anything else I can do for you?
U:	I thinks that's all
U:	I
U:	i'll try and use this thanks!
A:	No problem. Have a good one!

Table 19: Full Example (Sample #42) taken from the conversations dataset. URLs were updated to link to default view (rather than filtered). See Table 1 for truncated conversation.

Title: Production and value of maple products

Date range: 1924-01-01 to 2020-01-01
Dimensions: Geography, Maple products
Subject: Agriculture
Survey: Maple Products
Frequency: Annual

Geography:
ID: 1, Parent: None, Name: Canada
ID: 2, Parent: 1, Name: Nova Scotia
ID: 3, Parent: 1, Name: New Brunswick
ID: 4, Parent: 1, Name: Quebec
ID: 5, Parent: 1, Name: Ontario
Maple products:
ID: 9, Parent: None, Name: Maple products expressed as syrup, total
ID: 10, Parent: None, Name: Gross value of maple products
ID: 1, Parent: None, Name: Production of maple sugar
ID: 2, Parent: None, Name: Value of maple sugar
ID: 3, Parent: None, Name: Production of maple syrup
ID: 4, Parent: None, Name: Value of maple syrup
ID: 5, Parent: None, Name: Production of maple taffy
ID: 6, Parent: None, Name: Value of maple taffy
ID: 7, Parent: None, Name: Production of maple butter
ID: 8, Parent: None, Name: Value of maple butter

ID: 0, Note: Conversion factors: 1 gallon of syrup equals 10.0 pounds of maple sugar. One gallon of syrup weighs 13.24760 pounds. One gallon of syrup equals 10.4 pounds of taffy. Maple taffy is reported by Quebec and Nova Scotia only and commenced reporting in 1965 and 1983 respectively. The conversion of maple taffy to syrup varies with the density of syrup that year.
ID: 0, Note: Commercial production and value figures exclude inventory.
ID: 2, Note: Estimates produced by Nova Scotia horticulture industry specialists.
ID: 3, Note: Beginning in 1986, survey data for New Brunswick and Ontario are weighted using the number of taps made on maple trees in the spring of the year of the most recently completed census of agriculture.
ID: 4, Note: Estimates produced by l'Institut de la statistique du Québec, Groupe AGÉCO and Fédération des producteurs acéricoles du Québec.
ID: 5, Note: Beginning in 1986, survey data for New Brunswick and Ontario are weighted using the number of taps made on maple trees in the spring of the year of the most recently completed census of agriculture.
ID: 9, Note: "These products are represented by North American Product Classification System (NAPCS) Canada 2017 category 115136

Maple syrup and other maple products. Data collected for maple butter, maple taffy and maple sugar are converted into a maple syrup equivalent and are included in this category."
ID: 10, Note: "These products are represented by North American Product Classification System (NAPCS) Canada 2017 category 115136 – Maple syrup and other maple products. Data collected for maple butter, maple taffy and maple sugar are converted into a maple syrup equivalent and are included in this category."
ID: 0, Note: "These products are represented by North American Product Classification System (NAPCS) Canada 2017 category 115136 – Maple syrup and other maple products. Data collected for maple butter, maple taffy and maple sugar are converted into a maple syrup equivalent and are included in this category."

Table 20: Full metadata of table in Section 3.3. Single lines were added to delimit the scope: Title, basic information, item names, and footnotes. The double line was added to delimit the truncation limit of the DPR and TAPAS model (512 tokens). The basic information is presented in Table 5.

PID	18100004 (count = 170)
Title	Consumer Price Index, monthly, not seasonally adjusted
Dates	1914-01-01 to 2021-08-01
Dim.	Geography, Products and product groups
Subject	Prices and price indexes
Survey	Consumer Price Index
Freq.	Monthly

PID	11100222 (count = 125)
Title	Household spending, Canada, regions and provinces
Dates	2010-01-01 to 2019-01-01
Dim.	Geography, Statistic, Household expenditures, summary-level categories
Subject	Income, pensions, spending and wealth
Survey	Survey of Household Spending
Freq.	Annual

PID	17100005 (count = 123)
Title	Population estimates on July 1st, by age and sex
Dates	1971-01-01 to 2021-01-01
Dim.	Geography, Sex, Age group
Subject	Population and demography
Survey	Annual Demographic Estimates Canada, Provinces and Territories
Freq.	Annual

PID	33100214 (count = 117)
Title	Canadian Business Counts, with employees, June 2019
Dates	2019-01-01 to 2019-01-01
Dim.	Geography, Employment size, North American Industry Classification System (NAICS)
Subject	Business performance and ownership
Survey	Business Register
Freq.	Semi-annual

PID	35100177 (count = 104)
Title	Incident-based crime statistics, by detailed violations, Canada, provinces, territories and Census Metropolitan Areas
Dates	1998-01-01 to 2020-01-01
Dim.	Geography, Violations, Statistics
Subject	Crime and justice
Survey	Uniform Crime Reporting Survey
Freq.	Annual

PID	18100005 (count = 91)
Title	Consumer Price Index, annual average, not seasonally adjusted
Dates	1914-01-01 to 2020-01-01
Dim.	Geography, Products and product groups
Subject	Prices and price indexes
Survey	Consumer Price Index
Freq.	Annual

Table 21: Most frequently retrieved tables across all splits in the retrieval task. The *basic information* is provided for each table. Analysis presented in Section A.3.1.

U: hi, i was wondering if you vae any statistics on video game sales, ot high school drop out rates? infoSeekingQuestions; request;

A: Hello my name is Sylvain, how may I help you? offer;

U: hello, I am looking for high school drop out r ates and teen depression rates, but cant seem to find anything that goes back more than a year request;

A: Data for High School dropouts is compiled by the Provincial Education Ministry. To obtain this infor please contact the Education Ministry of your province inform; promise;

A: Data for mood disorder (depression) is available on our website from the Data Table 13- 10-0096-18 (https://www150.statcan.gc.ca/...) answer;

A: To view/ manipulate the variables available in a data table: • Click on "Add/Remove data" • Select at least one variable in each tab and click on "Apply" in order to view your customized table To download the data: • Click on "Download" • Select one of the three output formats then click on the hyperlink and save the table The following video may also be helpful to you: How to use the data tables (https://www.statcan.gc.ca/eng/sc/video/...). inform; request;

U: Hi team, can you please help me retrieve data of public deposits of chartered banks? infoSeekingQuestions; request;

A: Please hold timeManagement;

U: Surely. autoFeedback;

A: You can obtain this info on our website from our Data Table 10-10-0116-01 (https://www150.statcan.gc.ca/t1/...) answer;

A: To view/ manipulate the variables available in a data table: • Click on "Add/Remove data" • Select at least one variable in each tab and click on "Apply" in order to view your customized table To download the data: • Click on "Download" • Select one of the three output formats then click on the hyperlink and save the table The following video may also be helpful to you: How to use the data tables (https://www.statcan.gc.ca/eng/sc/video/...) . inform; instruct;

A: Please take note that this is a non-Statistics Canada Table. This info is provide from The Bank of Canada inform;

U: Thanks a lot! So, how should I get down to Public deposits? or rather which of these lines would direct me to Public deposits? clarificationQuestion;

A: Under the category Chartered bank deposits. For more info please contact The Bank of Canada on their website by using their Online Request Form (https://www.bankofcanada.ca/...) answer; inform;

Table 22: Samples of the dialogue analysis in Section 3.2, which are summarized in Table 15. The speech acts are in blue at the end of each message. The remaining annotations are in the supplementary materials. Sample examples are given in Table 4.

U:	**Thank you for your information, but I have done that already. Your previous website had a "manipulate" function to calculate the percentage change. I don't see the button anymore, and wondered if you have other schedule prepared as that function is gone.**
A:	**Hello, my name is Jimmy N., how may I help you?**
T5:	Hello, my name is Jimmy N., how may I help you?
T5+D:	Hello, my name is Jimmy N., how may I help you?
U:	**I was talking with Sylvain, will you be able to see the log?**
A:	**Yes, please hold while I verify.**
T5:	Yes, please hold while I verify.
T5+D:	Yes, please hold while I verify.
A:	**Just to clarify, you're looking for all-items ?**
T5:	You can obtain this information at the following hyperlink : Table: 18-10-0004-13 Consumer Price Index by product group, monthly, percentage change, not seasonally adjusted, Canada, provinces, Whitehorse, Yellowknife and Iqaluit ( )
T5+D:	You can obtain this information at the following hyperlink : Table: 18-10-0004-01 Consumer Price Index by product group, monthly, [...] and Iqaluit ( )

Table 23: Conversation #4890. Each message is separated by a horizontal line. The original conversation, in **bold**, only contains messages by the user (U) and agent (A). The generated responses by T5 and T5 with top-5 DPR-retrieved titles (T5+D) predict the corresponding agent's message (in **bold**).

	[...]
U:	**Ideally, I want the percentage change for these items**
R_1:	*Consumer Price Index, monthly, not seasonally adjusted*
R_2:	*Monthly adjusted price index, provisional*
R_3:	*Consumer Price Index, monthly, seasonally adjusted*
R_4:	*Historical (real-time) releases of the industrial product price index, by major product group, monthly, inactive*
R_5:	*Consumer Price Index (CPI) statistics, alternative measures, unadjusted and seasonally adjusted, Bank of Canada*
A:	**You can obtain this information at the following hyperlink : Table: 18-10-0004-13 Consumer Price Index by product group, monthly, percentage change, not seasonally adjusted, Canada, provinces, Whitehorse, Yellowknife and Iqaluit (https://www150.statcan.gc.ca/t1/tbl1/...)**
T5:	You can obtain this information at the following hyperlink : Table: 18-10-0004-13 Consumer Price Index [...] and Iqaluit ( )
T5+D:	You can obtain this information at the following hyperlink : Table: 18-10-0004-13 Consumer Price Index [...] and Iqaluit ( )

Table 24: The second part of conversation #4890, immediately after Table 23. R_i indicates the i-th table retrieved by DPR, which is only seen by T5+D.

U:	This is pretty cool!
A:	Hello, my name is Martine, how may I help you?
U:	I have a question - how would I be able to access data for 2019 on the amount of money the federal government spent (money out) to Canadians through social programs like EI / OAS etc...
A:	Letme check..
U:	Thank you
R_i:	*Federal extramural expenditures on science and technology, type of science, performing sector, type of payment and geography*
R_i:	*Historical: Current transfers to other levels of government, 1968 System of National Accounts (SNA), annual, 1926 - 1986*
R_i:	*Private non-profit organization outsourced research and development expenditures*
R_i:	*Capital expenditures, sources of funding, by geography*
R_i:	*Historical: Government transfer payments to persons, 1968 System of National Accounts (SNA), annual, 1926 - 1986*
A:	**Revenue, expenditure and budgetary balance - General governments (x 1,000,000) (https://www150.statcan.gc.ca/t1/...3610047701)**
T5:	Data Tables (https://www150.statcan.gc.ca/n1/en/type/data?text=federal+debt)
T5+D:	Federal extramural expenditures on science and technology, type of science, performing sector, type of payment and geography ( )

Table 25: Conversation #8960.

U:	How can I find information relating to Kelowna population estimates, demography, population projection estimates, population by age and sex
A:	hello how may i help you today?
U:	How can I find information relating to Kelowna population estimates, demography, population projection estimates, population by age and sex
A:	let me see what i can find on the subject for the projections information our tables only provide it to the provincial level it does not go lowe then that <extra_id_0> (<extra_id_0>)
R_1:	*Estimates and projections of population aged 0 to 29, by age group, Canada, provinces and territories*
R_2:	*Projected population, by projection scenario, age and sex, as of July 1*
R_3:	*Population estimates on July 1st, by age and sex*
R_4:	*Components of projected population growth, by projection scenario*
R_5:	*Population estimates, quarterly*
A:	**https://www150.statcan.gc.ca/t1/tbl1/en/cv.action?pid=1710005801**
T5:	https://www150.statcan.gc.ca/n1/daily-quotidien/190619/dq190619f-cansim-eng.htm
T5+D:	https://www150.statcan.gc.ca/n1/pub/71-607-x/71-607-x2018005-eng.htm

Table 26: Conversation #21533.

U:	Hi there, do you know approximately how many craft distillers are operating in Canada?
A:	Hello my name is Sylvain, how may I help you?
U:	I was wondering if you had any information about craft distillers in Canada. Specifically, how many craft distillers are operating in Can
A:	The industries are classified according to the North American Industry Classification System (NAICS). To find the NAICS code for an indu refer to this page (http://www23.statcan.gc.ca/imdb/...1181553) . You may start your search by selecting one of the 20 sectors, until a six-d obtained. [...] You will find the number of business locations by NAICS code in the following data tables: Table 33-10-0105-01 Canadian Business Counts, with employees, December 2018 (https://www150.statcan.gc.ca/t1/...3310010501) [...] I also suggest consulting the financial data of small and medium enterprise in Canada; these organizations have annual earning revenues ranges of $30,000 to $20 million. T information is available by industry, province, territory and legal status of companies (incorporated or unincorporated). This data is not available on our website; however, i available on Innovation, Science and Economic Development Canada's website, using their Financial Performance Data (http://www.ic.gc.ca/...)[...]

Table 27: Conversation #1628

U:	Hi. I want to know how many small, medium and big companies are there in canada. Thank you
A:	Hello, my name is Olivier C. Please wait while i find the information you're looking for.
R_1:	*Canadian Business Counts, without employees, December 2019*
R_2:	*Canadian Business Counts, with employees, December 2019*
R_3:	*Canadian Business Counts, without employees, June 2019*
R_4:	*Canadian Business Counts, with employees, June 2019*
R_5:	*Canadian Business Counts, without employees, December 2017*
A:	**Thank you for waiting, you will find the information about the number of business per business size if you click the following hyperlinks : Canadian Business Counts, with employees, December 2019 (https://www150.statcan.gc.ca/t1/...pid=3310022201) Canadian Business Counts, without employees, December 2019 (https://www150.statcan.gc.ca/t1/...pid=3310022301)**
T5:	Thank you for waiting. You will find the information you're looking for if you click the following hyperlink : Canadian Business Counts, with employees, June 2019 ( )
T5+D:	Thank you for waiting. The industries are classified according to the North American Industry Classification System (NAICS). To find the NAICS code for an industry, please refer to this page (http://www23.statcan.gc.ca/imdb/...1181553). You may start your search by selecting one of the 20 sectors, until a six-digit NAICS is obtained. You will find the number of business locations by NAICS code in the following data tables: Table 33-10-0214-01 Canadian Business Counts, with employees, June 2019 ( ) Table 33-10-0215-01 Canadian Business Counts, without employees, June 2019 ( ) I also suggest consulting the financial data of small and medium enterprise in Canada; these organizations have annual earning revenues ranges of $30,000 to $20 million. This information is available by industry, province, territory and legal status of companies (incorporated or unincorporated). This data is not available on our website; however, it is freely available on Innovation, Science and Economic Development Canada's website, using their Financial Performance Data (http://www.ic

Table 28: Conversation #8500.

U:	Do you have any data showing labour in manufacturing companies and how it is allocated between direct and indirect labour?
A:	Hello, my name is Jimmy N. Could you please specify what you mean by direct and indirect labour.
U:	Direct = labour for building product Indirect = administrative, supervision, etc.
A:	The data you are requesting is not available as a standard product. However, it may be available through a custom order (a fee may apply). [...] Would you like to make this request?
A:	As a standard product, we have tables about employment by industry.
U:	I would just like to look at the standard product. Can you tell me where I can find it, or what it is called?
R_1:	*Labour force characteristics by industry, monthly, seasonally adjusted, last 5 months*
R_2:	*Employment by industry, monthly, seasonally adjusted and unadjusted, and trend-cycle, last 5 months*
R_3:	*Employment by industry, annual*
R_4:	*Employment by industry, three-month moving average, unadjusted for seasonality, provinces and economic regions, inactive*
R_5:	*Usual hours worked by industry, annual*
A:	**You can obtain this information at the following hyperlinks : Table: 14-10-0355-02 Employment by industry, monthly, seasonally adjusted (x 1,000)Table: 14-10-0023-01 Labour force characteristics by industry, annual (x 1,000)Employed labour force who worked full year, full time and reported employment income in 2015, 2016 Census (https://www150.statcan.gc.ca/t1/...1410035502)**
T5:	Table: 14-10-0296-01 Labour force characteristics by industry, annual (x 1,000) ( )
T5+D:	Please hold while I find the information.

Table 29: Conversation #19568.

U:	I am on webpage https://www150.statcan.gc.ca/n1/daily-quotidien/201224/dq201224beng.htm where can I find the same data set for previous years, for example 2019 2018 2017 etc
A:	Hello, can you tell me on which table you are : https://www150.statcan.gc.ca/n1/daily-quotidien/201224/dq201224b-cansim-eng.htm
U:	Provisional death counts and excess mortality, January to October 2020 Adjusted number of deaths, expected number of deaths and estimates of excess mortality, by week
A:	That is the report, but there is 5 different tables: https://www150.statcan.gc.ca/n1/daily-quotidien/201224/dq201224b-cansim-eng.htm Let me verify this for you. One moment please. To view/ manipulate the variables available in a data table: • Click on "Add/Remove data" • [...] Select at least one variable in each tab and click on "Apply" in order to view your customized table Data are available for 2016 up to 2020.
U:	The reference period stops at January 2020 I can't seem to get data for past years
A:	Click on the schedule icon and select the year you want. (2016 - 2020)
U:	I did and all years are 'greyed out' meaning they do not populate
A:	Let me verify this. One moment please.
U:	screenshot.lzsc
R_1:	*Weekly death counts, by age group and sex*
R_2:	*Previous releases and revisions to weekly death counts*
R_3:	*Adjusted number of deaths, expected number of deaths and estimates of excess mortality, by week, age group and sex*
R_4:	*Death counts, by week, inactive*
R_5:	*Labour force characteristics by sex and detailed age group, monthly, unadjusted for seasonality*
A:	**Have you already consulter the table that replaces table# 13-10-0785-01 [...] https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1310081001 - On this table you can go back to 2016.**
T5:	You can manipulate the table by following these steps: • Click on "Add/Remove data" • Click on the " + " icon next to a checkbox to expand the list of variables (if applicable) • Select at least one variable in each tab and click on "Apply" in order to view your customized table To download the data: • Click on "Download" • Select one of the three output formats then click on the hyperlink and save the table The following video may also be helpful to you: How to use the data tables (https://www.statcan.gc.ca/eng/sc/video/howto).
T5+D:	Please consult the following data table(s): Weekly death counts, by age group and sex ( )

Table 30: Conversation #15376.

U:	Hello	
A:	Hello, my name is Jimmy N., how may I help you?	
U:	[...] We saw this article from the Vancouver Sun which mentioned StatsCan data on business closures. [...] is there any way for us to collect localized data for the Township, or at least the Fraser Valley	
A:	Unfortunately, the information is available on a national and provincial/territorial level : And for some Census metropolitan areas	
R_i:	*Experimental estimates for business openings and closures for Canada, provinces and territories, census metropolitan areas, seasonally adjusted*	
R_i:	*Quarterly estimates of business entry and exit*	
R_i:	*Of businesses where at least one employee was laid off, percentage of workforce laid off and rehired due to COVID-19, by business characteristics*	
R_i:	*Canadian Business Counts, with employees, census metropolitan areas and census subdivisions, December 2020*	
R_i:	*Manufacturing sales, by industry for 12 Selected Census Metropolitan Areas*	
A:	**Table: 33-10-0270-01 - Experimental estimates for business openings and closures for Canada, provinces and territories, census metropolitan areas, seasonally adjusted (https://www150.statcan.gc.ca/t1/...3310027001)**	
T5:	However, it may be available through a custom order (a fee may apply). To obtain a free cost estimate for this service, please send us your request, clearly indicating the variables, geographic areas, time frame and data frequency needed, making sure to include your complete contact information (name of organization, address, postal code, and phone number) at the following email address: infostats@canada.ca (mailto:infostats@canada.ca) or by telephone at 1-800-263-1136. Our agents are available Monday to Friday (except holidays) from 7:30 am to 7:30 pm (Eastern Time). An agent will get in touch with you to verify the information, provide an estimate and proceed with the payment, if applicable.	
T5+D:	However, it may be available through a custom order (a fee may apply). To obtain a free [...] An agent will get in touch with you to verify the information, provide an estimate and proceed with the payment, if applicable. Please do not share any personal information through the live chat. Would you like to make this request?	

Table 31: Conversation #11869.

PID	Title	Train	Valid	Test	Total
13100810	Selected grouped causes of death, by week	0	0	6	6
33100270	Experimental estimates for business openings and closures for Canada, provinces and territories, census metropolitan areas, seasonally adjusted	0	1	3	4
13100763	Health characteristics of children and youth aged 1 to 17 years, Canadian Health Survey on Children and Youth 2019	0	0	4	4
35100060	Number of homicide victims and persons accused of homicide, by Indigenous identity, age group and gender	0	1	2	3
32100042	Estimated areas, yield, production of corn for grain and soybeans, using genetically modified seed, Quebec and Ontario, in metric and imperial units	0	1	2	3
14100084	Labour force characteristics of immigrants by sex and age group, three-month moving average, unadjusted for seasonality	0	0	3	3
13100143	Deaths, by cause, Chapter V: Mental and behavioural disorders (F00 to F99)	0	0	3	3

Table 32: List of tables that do not appear in the training set, but appear at least 3 times in the validation or test sets.

Language	Split	Augmentation	METEOR	ROUGE-L	BERTScore	MoverScore	Title Acc.
English	Test	None	23.35	30.65	86.04	59.82	6.96
		Top-1 Title	24.07	30.76	86.11	60.23	7.99
		Top-5 Titles	24.41	30.88	86.17	60.31	10.82
	Valid	None	31.95	43.31	88.31	65.50	30.39
		Top-1 Title	33.11	44.22	88.60	66.14	33.66
		Top-5 Titles	33.38	44.20	88.55	65.94	38.89
French	Test	None	5.98	7.21	61.74	51.88	0.00
		Top-1 Title	9.71	11.14	65.86	54.81	0.00
		Top-5 Titles	8.96	10.75	64.76	54.38	0.00
	Valid	None	6.84	8.85	60.97	52.21	0.00
		Top-1 Title	8.09	9.03	64.22	53.62	0.00
		Top-5 Titles	7.37	8.75	62.55	52.97	0.00

Table 33: Full response generation results. Selected results in Table 10.