arXiv:2304.01412v2 [cs.CL] 5 Apr 2023

加拿大统计局对话数据集：通过具有真实意图的对话检索数据表

Xing Han Lu [1 2†] **Siva Reddy** [1 2 3] **Harm de Vries** [2]魁北克米拉人工智能研究所、麦吉尔大学[2]ServiceNow Research[3]
Facebook CIFAR 人工智能讲席
statcan.dialogue.dataset@mila.quebec

摘要

我们介绍加拿大统计局对话数据集[1]该数据集包含 19,379 个对话轮次，来自加拿大统计局工作人员与寻找已发布数据表的在线用户之间的对话。这些对话源于真实意图，以英语或法语进行，并促使智能体检索 5,000 多个复杂数据表之一。基于该数据集，我们提出两项任务：（1）根据正在进行的对话自动检索相关表格；（2）在每个轮次自动生成恰当的智能体回复。我们通过建立强有力的基线，研究各项任务的难度。在时间数据划分上的实验表明，所有模型都难以泛化到未来对话：当我们从验证集转向测试集时，两项任务的性能均显著下降。此外，我们发现，回复生成模型难以判断何时应返回表格。鉴于这些任务对现有模型构成了重大挑战，我们鼓励社区为这些任务开发模型，以便直接帮助知识工作者为在线聊天用户找到相关表格。

1 引言

自然语言处理（NLP）长期以来的目标之一，是开发能够帮助人们完成具体任务的对话式智能体，例如从大量文档中查找信息，或预订餐馆和酒店。为了推动此类虚拟助手的发展，研究社区需要能够反映真实世界应用中常见意图和语言现象的基准数据集。然而，在当前的研究环境中，开发这类真实世界对话数据集颇具挑战。

U_1：嗨，我想获取加拿大国内生产总值增长、 不列颠哥伦比亚省住房开工量以及加拿大石油价格/桶的季度数据。A_1：您好，我叫 Kelly C。请稍等，我查一下。[...]A_1：有关 GDP 增长率，请查看以下链接： [...]A_1：您所说的不列颠哥伦比亚省住房开工量是什么意思？U_2：我需要研究不列颠哥伦比亚省所有新屋开工的季度数据。 [...]U_2：新屋开工量是指在某个月份开始建设的住宅建设项目数量。 [...]A_2：我可以提供每月新发放建筑许可证的数据。 [...]A_2：按建筑结构类型和工作类型划分的建筑许可证：https[...]A_2：我来查一下石油价格，请稍等。U_3：您还需要加拿大的石油价格/桶（“WTI”）数据吗？ [...]A_3：您要找的是石油零售价格吗？A_3：如果是，我找到了按较小地理区域划分的一些数据。A_3：按地理区域划分的汽油和燃料油月平均零售价格（https[...]）A_3：[...] 这些地理区域的数据够用吗？A_3：还是您只需要加拿大范围的数据？U_4：[...] 我需要更能代表整个加拿大的数据。A_4：这个怎么样？食品及其他精选产品的月平均零售价格（https[...]）

表1：StatCan 对话数据集示例，其中用户（U）与加拿大统计局智能体（A）对话，以查找若干数据表。**粗体** 文本表示智能体检索到的表格标题。

另一方面，学术实验室往往难以为面向任务的对话代理构思自然的使用场景，也难以收集大量真实用户的对话。许多实验室设计了人为任务，并从具有模拟意图的众包工作者那里收集对话（Budzianowski等人，2018年；Adlakha 等人，2022年；Lee等人，2022年），这往往导致数据集无法反映实际生产环境中的语言挑战（de Vries 等人，2020年）。另一方面，行业实验室可能能够接触到具有真实意图的用户（例如，通过

[†]作为 ServiceNow Research 访问研究员完成的工作[l]网站： mcgill-nlp.github.io/statcan-dialogue-dataset

数据集	意图	对话	查询	结果	来源	语言
我们的工作	真实	✓	问题，请求	表格链接，对话行为	加拿大统计局	英语、法语
自然问题（NQ）（2019年）	混合的	✗	问题	文本片段摘录	Google、维基百科	En
DuReader (2018年)	混合的	✗	问题	文本片段摘录	百度	Zh
OTT-QA (2020年)	模拟	✗	问题	表格/跨度摘录	维基百科	En
TAPAS-NQ（2021年）	混合的	✗	问题	表格摘录	Google，维基百科	En
CoQA (2019年)	模拟	✓	问题	文本片段摘录*	多个	En
QuAC（2018年）	模拟	✓	问题	对话行为	维基百科	En
ATIS（1990）	真实	✓	请求	SQL 查询、命令	TI Corp.	En
SGD-X (2022年)	模拟	✓	请求	API 调用、对话行为	对话模拟器	En

表2：与相关数据集的比较（参见 第 2 节）。(*) CoQA 使用解释依据来支持抽取的答案。

Sirior Alexa），但由于其商业价值和用户隐私问题，很少发布此类对话数据集。因此，我们认为，研究社区将受益于一种面向任务的对话环境，使研究发现能够通过真实用户得到验证。为此，我们与加拿大统计局合作，推出了一个独特的数据集。

加拿大统计局（StatCan）是一个国家统计机构，负责收集有关加拿大经济、社会和环境的关键信息。加拿大统计局针对加拿大人生活的几乎所有方面开展数百项调查，并将由此产生的数据表发布在statcan.gc.ca上。该网站目前收录了 5K+ 个此类复杂且通常规模庞大的数据表。加拿大公民以及其他感兴趣的人士会访问该网站，查找他们需要的统计数据。加拿大统计局网站提供英语和法语聊天功能，以便在用户找不到合适信息时提供帮助。

基于这些在线聊天内容，我们推出了 StatCan 对话数据集，其中收录了 20K+段英语和法语对话，参与者包括statcan.gc.ca 的访客和加拿大统计局的工作人员。在发布该数据集之前，加拿大统计局采取了多项措施来删除个人可识别信息（PII）。我们观察到种类繁多的用户意图，从表格操作到导航说明都有；许多访客使用聊天功能，在加拿大统计局网站上查找数据表。具体而言，我们观察到 4.4K 段对话中，智能体返回数据表链接的情况共有 6.6K 次。在表 1 中，我们提供了一个在线用户查找特定数据表的对话示例。

在这项工作中，我们开发了两项新任务，旨在帮助用户找到特定表格。首先，我们提出表格检索任务，要求模型根据目前已发送的消息，预测智能体返回的表格。其次，我们提出响应生成任务，要求模型根据对话历史预测智能体的响应。对于这两项任务，我们通过建立强大的基线并使用各种指标进行评估，来研究其难度。

我们强调，这两项任务在现实应用场景中都具有直接的实用价值。表格检索任务能够帮助智能体更快速地找到相关表格，而响应生成任务则有望推动虚拟代理的发展，使其能够通过线上对话返回相关表格。我们希望，这种与现实应用场景的紧密联系能够帮助研究社区进一步认识开发实用型对话代理所面临的挑战，并加速研究理念与成果向实际应用的转化。

2 相关工作

本节介绍与我们的工作相关的各个研究方向。详见表2 以了解比较概要。

开放域问答 这项任务是使用大量且多样化的文本文档集合来回答问题。开放域问答领域最早的大规模评测之一发表于 TREC-8（Voorhees，2001年）。此后，许多研究发布了大规模开放域问答数据集：WikiQA（Yang 等人， 2015年）和 MS MARCO（Bajaj 等人， 2018年）从必应搜索引擎获取问题，Natural Questions（自然问题（NQ））（Kwiatkowski 等人， 2019年）从 Google 搜索获取问题，而 DuReader（He 等人， 2018年）则从百度获取中文问题。这些问题来自真实用户，答案由众包工作者从搜索结果中收集。尽管这些数据集中的问题都具有真实意图，且答案必须从文档集合中检索得出，但我们的数据集侧重于检索表格（在对话场景中），而非自由形式的文档。

表格检索与问答 在关于表格预训练的后续研究中（Yin 等人，2020年）、表格转文本

数据集	训练集	验证集	Test	All
# 对话	2573	545	557	3675
# 轮次	11382	2339	2600	16321
# 消息	36147	7385	8340	51872
# 查询	3782	799	870	5451
# 每条消息的词元数	32.83	33.51	29.32	32.36
# 每段对话的轮次	4.42	4.29	4.67	4.44
# 每段对话的消息数	14.05	13.55	14.97	14.11
# 每段对话的查询数	1.48	1.47	1.57	1.49
# 表格	778	349	388	959
# 新增表格	0	41	145	181
# 每个表格的维度数	3.5	3.5	3.6	3.6
# 每个表格的成员数	185.5	210.8	175.6	172.1
# 每个表格的注释数	21.1	22.7	23	20.4

表3：检索与生成任务中的英语对话和表格的统计数据。新增表格相对于训练集计算（见表11）。

生成（Parikh 等人，2020年）以及用于语义解析的弱监督（Herzig 等人，2020年），Chen 等人(2020年) 和 Herzig 等人(2021年) 分别提出了 OTT-QA 和 TAPAS-NQ 这两种新方法，将开放域问答扩展到检索表格而非文档。前者通过众包工作者收集问题和答案，后者则利用答案所在文章中的表格扩展 Natural Questions。在这两种情况下，检索的表格都来自维基百科文章。尽管我们的数据也包含表格检索，但表格来源于statcan.gc.ca，其规模可能大得多（如附录A.2.1所述），并且是在交互式对话场景中检索的。

对话式问答 一些研究将问答扩展到了对话场景。CoQA（Reddy 等人，2019年）和 QuAC（Choi 等人，2018年）引入了这样的数据集：围绕取自某份文档（例如维基百科文章）的参考段落提出多个轮次的问题。后续研究将这一设置扩展到了开放域场景，在该场景中，参考段落事先并不确定（Qu 等人，2020年；Anantha 等人，2021年；Adlakha 等人，2022年）。Saeidi 等人（2018年）提出了一个关于监管文本的对话式问答任务。上述数据集的结构都相同：在每一轮中，第一个说话者提出一个问题，另一个说话者给出一个答案。相比之下，我们的对话中的查询并不局限于问题，答案可以是表格、元数据或对话行为。

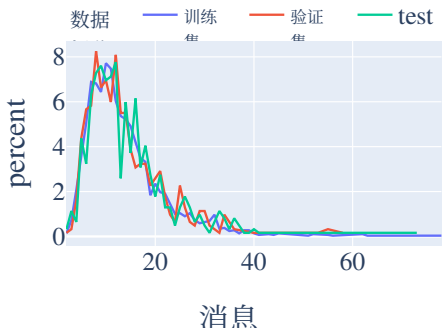

图1：两项任务中按对话统计的消息直方图（法语数据划分见图7）。

面向任务的对话 我们的工作与面向任务的对话研究相关，在这类对话中，用户与虚拟代理进行交流，以实现特定目标，例如预订餐馆或解决客户问题。尽管早期研究在真实的信息寻求设置中收集了一个数据集（Hemphill 等人，1990），但许多近期数据集都是通过模拟设置收集的（Budzianowski等人，2018年；Rastogi 等人，2020年；Feng等人，2020年，2021年；Chen 等人，2021年；Lee等人，2022年）。面向任务的模型通常通过预测由意图和槽值对指定的对话行为来跟踪对话状态，例如，`findRestaurants(cuisine=Italian)`。尽管我们的数据集不提供基于轮次的标注，但发布的对话附带了一个已标注目标，即用户正在查找的数据表。与其他面向目标的对话任务一样，这一标注使我们能够通过任务完成度指标自动评估对话模型。

闲聊对话 闲聊系统的目标是与终端用户进行开放式对话（Loweet al.，2015年；Dinanet al.，2018年）。与我们的数据集不同，这类对话并非旨在帮助用户完成特定任务。

3 数据集

StatCan 对话数据集由2019年3月1日至2021年3月8日期间在线聊天中收集的对话组成。尽管规模超过2.5万段对话的更大数据集中包含各种用户意图，我们只关注一种意图，选取了所有智能体返回数据表的对话。我们使用这个子集来开发和测试第4节中介绍的两项任务模型。第4节。在第3.1节中，我们提供该数据子集的基本统计数据

合并后的言语行为（示例）	%
回答（您可以在我们的网站上获取……）	50
请求（……请帮我检索数据……）	31
时间管理（请稍候）	28
告知（请注意……）	63
信息寻求问题（您还有其他……？）	25
承诺（……请联系教育部……）	18
自动反馈（当然）	25
提议（……我能为您提供什么帮助？）	11
指示（至少选择一项……）	18
澄清问题（以下哪条语句会指引……）	16
问候（嗨）	28
自我介绍（我的姓名是……）	17
致谢（非常感谢！）	47
接受致谢（不客气）	15

表4：验证集对话中每100轮次出现的合并言语行为的频率。

，并对第3.2节中的少量对话进行对话分析。第3.2节。在第3.3节中，我们将注意力转向数据表，并说明其中提供了哪些信息。最后，我们说明数据集在第3.4节中如何划分为训练集、验证集和测试集。技术规格见第3.4节。数据集卡片见附录B。

用户意图 在线聊天旨在满足特定的用户意图。聊天功能的主要目的是帮助用户 **查找特定的数据表格**。例如，在 表1中，智能体帮助用户查找有关建筑许可证、汽油价格和食品零售价格的表格。不过，用户也可能希望 **获取元信息**、获得有关 **操作**表格或使用 **用户界面**方面的帮助。在某些情况下，用户会提出 **域外请求**。由于本文的重点是主要意图，因此这些辅助意图将在附录 A.1 中进行描述。

消息与轮次 每段对话都被拆分为**轮次**，每个轮次由一组用户与代理的回复构成。每条回复可以包含多个**消息**，并按顺序依次发送（例如，在表1中，智能体的第一条回复包含连续发送的3条消息）。

3.1 基本统计

总计将提供 25397 段对话。根据我们的主要意图，我们聚焦于一个包含 4468 段对话的子集（英语 3675 段，法语 793 段）。在英语和法语中总共可用的 5907 张表格中，智能体返回了 959 张独特的英语表格和 285 张法语表格。英语数据集划分中，每段对话的消息数从 2 到 78 不等，中位数为 12（见

标题：枫树产品的产量与价值
日期范围：1924-01-01 至 2020-01-01
维度：地理、枫树产品
主题：农业
调查：枫糖产品
频率：年度

表5：第3.3节中表格的基本信息（包括标题）第3.3节。完整版本见表20。可通过以下网址访问：doi.org/10.25318/3210035401-eng。

图1 表示分布）。根据 表3，每段对话平均有 4.4 个轮次，但包含 14.12 条消息，每条消息超过 30 个词元。这表明说话者会连续表达多个想法，然后收话人才会回应。对于法语数据划分，我们在附录 A.5中分析了基本统计数据。

高频请求表格总计，检索到的表格中有6张（占13.4%），涵盖通货膨胀和家庭支出等主题。补充详情见附录A.3.1。

3.2 对话分析

我们根据Bunt等人（2010年，2020年）定义的言语行为，对英语验证集中24段对话的100个轮次（306条消息）进行分类。这套分类法也称为ISO标准24617-2。我们沿用其分类体系，但将一些细粒度行为归并到更广泛的类别中（例如，纠正、同意、不同意归入告知）。我们在表4中列出言语行为的频率和示例。有关如何合并原始行为及补充示例的更多信息，请参阅附录（表15和表22，分别对应）。

我们注意到，答案 出现的频率是信息寻求问题 的两倍，因为对话者可能会同时回答澄清问题和请求。此外，告知 行为出现的频率高达 63%，因为智能体需要进一步阐述自己的回答，而用户往往会通过告知对方来澄清最初的请求。尽管自动反馈 和时间管理 出现的频率较低，但仍然很重要，因为对话者无法依赖点头或打字等视觉反馈。自然地，时间管理 常与承诺 同时出现，因为智能体往往会让用户稍候，同时承诺会满足其请求。

3.3 表格规范

为了说明这些规范，我们来查看一个包含**标题** 枫树的生产和价值的示例表格

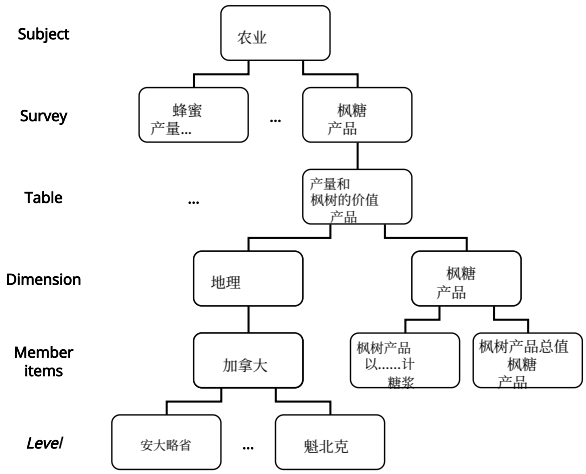

图2：元数据组件之间层级关系的示意图，相关内容见第3.3节。

产品（见 表5）。该表有两个**维度**，每个维度由一组**成员项**构成；就地理而言，成员是生产枫糖浆的省份（魁北克、安大略省等）；就枫树产品而言，成员是不同的产量类型（枫糖、糖浆、太妃糖和黄油）。成员项将行和列的概念泛化，因为通过透视操作，两者可以互换。有时，关于这些成员的详细信息会以**脚注**的形式列在页面末尾。

基本信息 这是核心元数据，由标题、维度、主题、调查和更新频率组成（不包括成员项目）。补充详情见附录A.2.2。

层级关系 元数据可以按层级查看。如图 2所示，每个主题包含不同的调查，每项调查可用于生成一个或多个表格，依此类推。可嵌套在另一成员项下的成员项称为层级。

3.4 数据集划分

我们将对话划分为训练集（70%）、验证集（15%）和测试集（15%）。测试集专门按日期选取最新的对话（涵盖2020年9月8日至2021年3月8日），而训练集和验证集则从剩余数据中随机选取（涵盖2019年3月1日至2020年9月8日）。这使我们能够测试模型适应数据中时间变化的能力（例如新的数据发布和新事件）。这有助于了解模型在训练分布之外进行泛化的能力，同时也更能反映模型在真实世界应用中的表现（模型将用于未来的数据）。所有任务均使用相同的数据划分。

4 任务

基于对话和表格数据，我们提出两项任务：(i) 表格检索任务，要求模型利用部分对话预测智能体将返回的表格；(ii) 响应生成任务，要求模型利用部分对话生成智能体最有可能给出的响应。这些任务中的对话提供英语和法语两种版本。

4.1 检索任务

对于此任务，我们会在智能体分享相关表格的链接之前截断每个对话。因此，与该链接对应的产品ID（PID）就成为检索任务的目标，如表格 6所示。当智能体在一次对话中分享多个不重复的PID时，每个唯一出现的PID都会被视为一个单独的样本。

Recall@k 为了评估用于检索的模型，我们计算 k (R@k) 分数，用于 $k \in \{1, 10, 20\}$；该分数表示在模型检索出的 k 个表格中包含正确表格的比例（通常按相关性分数排序）。我们为实时自动检索选择 $k = 1$，为人类或自动重排序器希望使用检索器查询表格并选择最佳选项的场景选择 $k \in \{10, 20\}$ 。

4.2 响应生成任务

在第一个任务中，仅考虑引向表格检索的消息。在此任务中，智能体发送的每条消息都被视为目标文本，而此前的所有内容则是源文本。因此，此任务的目标是利用源文本生成与目标文本相匹配的响应（见表格 7）。由于对话响应难以评估，我们为此任务报告了种类繁多的指标。

ROUGE-L 和 METEOR ROUGE-L（Lin，2004年；Lin 和 Och，2004年）是一种常见的文本评估指标，它通过识别两个句子之间的最长重叠词序列，自然地考虑句子级结构。METEOR（Banerjee 和 Lavie，2005年）是一种词级精确率和召回率评分方法，涵盖了表示一个词的不同方式，包括词干和同义词。

	源文本（正在进行的对话）
	[...]
A:	你所说的不列颠哥伦比亚省住房开工量是什么意思?
U:	我需要研究所有 不列颠哥伦比亚省按季度统计的新屋开工量[...]
U:	住房开工量是指新建住宅的数量 在任意月份开始的建筑业项目 的
A:	[.....] 我希望获取有关新建筑许可证签发情况的月度数据。 发放情况。
	检索目标（加拿大统计局表格）
	表34-10-0066（按结构类型划分的建筑许可证，征和工作类型）

表6：检索任务的源文本和目标文本，依据 表1。给定正在进行的对话，目标是检索一张加拿大统计局表格。

BERTScore 和 MoverScore 研究人员开发了多种方法，利用 BERT（Devlin 等人，2019年）的上下文嵌入来评估两个句子之间的相似度。BERTScore（Zhang 等人，2019年）在词元级别计算余弦相似度，而 MoverScore（Zhao 等人，2019年）在词或句子级别计算推土机距离（Rubner 等人，2000年），从而捕捉将生成回复的分布转换为原始回复的分布所需的代价。

标题准确率 除了文本生成的通用指标外，我们还探讨了这一任务专用指标。我们将其定义为生成的消息中包含参考消息所分享表格标题的比例。因此，该指标仅涵盖智能体分享表格的轮次。为计算此指标，我们：(i) 在参考消息中查找产品ID；(ii) 查询标题；(iii) 检查该标题是否在生成文本和参考文本中原样出现。

5 模型

为帮助了解微调模型在我们各项任务上的表现，本节概述英语数据划分所采用的方法，具体实现细节见附录 A.4。法语采用了类似的架构（详见附录A.5）。

5.1 检索

BM25 我们采用 Robertson和Zaragoza(2009）的算法，通过对查询词相对于段落的词频进行加权，检索与给定查询相似的表格（段落）的元数据。

	源文本（进行中的对话）
U:	您好，我想获取有关季度数据的内容。 国内生产总值增长（加拿大），不列颠哥伦比亚省住房开工量，加拿大石油价格/桶
A:	您好，我的姓名是Kelly C.。请稍等片刻，我将... 我正在搜索 [...]。
A:	关于GDP增长率，请查阅 [...]
U:	我需要研究所有的新屋开工数据， 不列颠哥伦比亚省按季度统计的 [...]
	生成目标文本（智能体的下一条回复）
A:	我将获得有关新建筑 许可证签发情况的每月数据。 [...]

表7：响应生成任务的源文本和目标文本，基于表1。给定正在进行的对话，目标是生成智能体的回答。

idf 缩放后的查询词相对于段落的词频。

DPR 由Karpukhin 等人提出 (2020年)，密集段落检索（DPR）由一对 Transformer 模型组成，分别对查询和段落进行编码；如果段落与查询相关，所得向量的点积得分就会更高。我们对该模型进行微调，使其能够根据正在进行的对话（查询）检索表格（段落）的元数据。

TAPAS 和 TAPAS-NQ Herzig 等人(2020年) 提出了一种模型，该模型在预训练期间以自监督方式学习对展平后的表格单元格进行编码。我们对其进行微调，使其能够根据正在进行的对话检索表格的截断内容。随后，Herzig 等人 (2021年) 对 TAPAS 进行微调，使其在从自然问题（NQ）中提取的 12K 个问题-答案-表格三元组上执行开放域表格检索；我们进一步以相同方式微调这一变体，并将结果报告为 TAPAS-NQ。

探索表格表示在最简单的场景中，BM25 和 DPR 仅使用标题。此外，我们还评估了对基本信息、成员项、脚注或它们的组合进行编码的变体。对于 TAPAS 和 TAPAS-NQ，我们还微调了一个检索标题、维度和成员项的变体，因为原始 TAPAS 能够关注标题和列名。

5.2 响应生成

T5 我们对 T5 的大型变体（Raffel 等人，2020年）（命名为 无增强，见 表10）进行了微调，使其先对源文本（正在进行的对话）进行编码，再自回归地解码目标文本（智能体回复）。

元数据	R@1	R@10	R@20
基础	14.7	45.0	55.0
基础 +成员	15.7	46.2	56.3
基础 +脚注	13.9	44.4	54.2
成员	10.7	35.0	46.3
标题	13.9	43.8	53.4

表8：针对英语测试集数据划分，在不同表格表示下 DPR 的检索结果。元数据概览见第3.3节。

模型	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

表9：对标题和成员项进行编码时，英语测试集划分的检索结果。DPR 和 TAPAS 各运行 3 次并取平均值。

使用排名靠前的标题扩充 T5k 对于每段部分对话，我们使用 DPR（基础+成员）检索排名靠前的k 个表格（其中 $k \in \{1, 5\}$），并将其标题附加到部分对话中。这使 T5 能够决定是使用建议的标题之一，还是生成其他内容（例如，澄清问题）。这与智能体的行为类似，因为它们在分享相关表格时，往往会返回一个标题及其 URL。此外，监督 T5 忽略或返回标题，相当于进行隐式二元分类。

6 结果与讨论

基于我们的基线和数据，我们报告结果，并分析数据集和任务给现有模型带来的挑战。对于英语数据划分，主要检索结果见表9 和表8，主要生成结果见表10。完整结果分别见附录A中的表16 和表 33，相关统计检验见附录A.7。

表格表示的影响在表9中，我们观察到元数据表示会影响检索召回率。尽管 DPR 仅检索标题时也能取得不错的结果，但加入基本信息（定义见第 3.3 节）带来的提升较小，而进一步加入成员项后，与仅使用标题相比则会产生显著差异（$p = 0.014$）。然而，仅使用成员项会导致召回率大幅下降（$p = 0.00086$），这表明标题至关重要。此外，脚注并未带来任何提升，这可能是因为

指标	无增强	排名第一的标题	Top-5 标题
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
标题准确率	6.96	7.99	10.82

表10：英语测试集划分的响应生成结果。

它们通常会超出最大上下文长度（参见表20）。因此，要在检索任务中获得较高的召回率，简洁且有意义地表示元数据至关重要。

迁移到表格检索任务 我们的实验使我们能够分析开放域问答微调（NQ）和表格预训练在迁移到我们的表格检索任务时的效果。我们在表9 中观察到，DPR 的测试集 recall@10 分别比 TAPAS 和 TAPAS-NQ 高出 23.0% 和 15.1%。此外，TAPAS-NQ 在仅检索标题和成员项、而非完整表格时，表现更好（$p = 0.016$），这可能是由于上下文长度限制导致重复和截断。尽管 DPR 和 TAPAS-NQ 都是在 NQ 上训练的，但后者是在包含表格的较小子集（12K 对 320K）上训练的。我们的实验表明，TAPAS 从一个任务（NQ-Tables）迁移到另一个任务（加拿大统计局）时效果不佳。

响应生成 我们比较了微调后的 T5-large 模型在使用和不使用 DPR增强的表格标题时的表现。在表10中，我们注意到，检索增强模型在4项非任务特定指标上均有小幅提升。不过，使用 Top-5 标题的增强模型在标题准确率上绝对提升了3.86%，这表明 DPR 提供的信息确实有助于 T5 生成目标标题。尽管如此，未使用增强的 T5 仍取得了6.96%的分数，这令人惊讶，说明 T5 模型能够存储训练时见过的标题，并且在一定程度上能够在测试时回忆并返回这些标题。最后，我们指出，标题准确率仍比 DPR检索器的 top-1召回率低5.91%（表8），这表明 T5 未能学会何时返回表格（尽管智能体在23.4%的所有轮次中检索到表格）。对于 Top-5 标题，T5 难以判断应返回哪个表格标题。

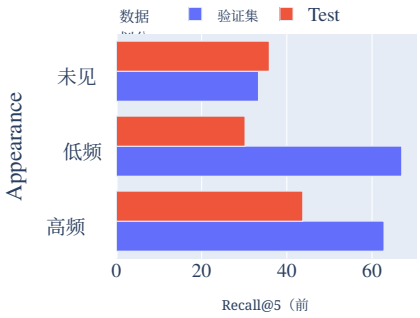

图3：DPR B+M 在训练集中高频（10+次）、低频（1-10次）出现的表格，以及训练时未见表格上的结果。

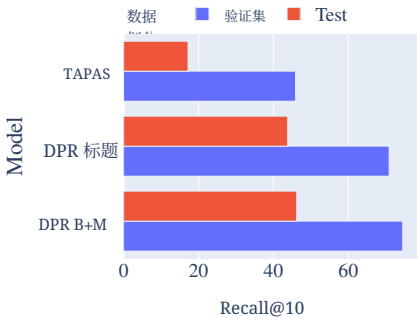

图4：所选检索模型的验证集和测试集召回率。我们观察到性能显著下降。*B+M* 表示基础 + 成员。

生成响应的定性分析我们考察了多段对话，以了解 T5 会生成何种响应。我们发现，它能够生成问候语等简单的言语行为，但在处理澄清问题等依赖上下文的言语行为时可能会遇到困难。此外，它能够可靠地回复常见表格的标题，但处理不常见表格的标题时会遇到困难；有时在 DPR 的帮助下，它也能生成未见过的标题。完整分析见附录 A.6。

时间漂移 如第3.4节所述，我们采用时间划分来测试模型对未来对话的泛化能力。通过比较验证集和测试集上的表现，我们发现图4中召回率显著下降（13%-28.3%），即使模型采用不同的元数据表示进行训练，也存在这一现象。类似地，T5在响应生成任务的测试集上的得分也较低（表10）。这一巨大差距表明，训练后的模型难以泛化到未来对话。首先，我们发现，这很可能是由于测试集中出现的新表格数量（145）远多于验证集（41）所致，如表3所示。此外，对话主题也发生了显著变化：用户开始更加关注商业、健康和信息技术，而较少关注人口统计、收入和养老金。这很可能是受到影响用户的现实世界事件驱动的，而这些事件更难从模拟环境中隐式捕捉。然而，为了理解模型在时间偏移下的鲁棒性以及其在现实世界应用中的表现，这种能力十分必要。在附录中，图8展示了所有主题在训练集和测试集之间的差异。

泛化到未见表格如图3所示，DPR 对验证集数据划分中频繁出现的表格表现良好，但在测试集数据划分中表现较差，这可能是由时间漂移导致的。正如预期，在训练期间未见过的表格在两个数据划分中的 Recall@5（前5项召回率）均较低。此外，低频表格在验证集和测试集之间的召回率差异可能由许多潜在原因造成（学习偏差、时间过拟合、与隐藏因素的虚假相关）。因此，未来的模型应致力于缩小未见表格与高频表格之间的差距，并缩小低频表格在时间维度上的差距。

法语结果 在两个任务中，由于数据集规模较小，所有模型的所有指标都出现了下降。部分观察结果仍然有效：时间漂移、BM25 表现不佳，以及数据增强在某些指- 标上有利于 mT5。然而，其他结果有所不同：添加成员项会损害测试结果，而 mT5 在 标- 题准确率方面表现不佳。建模细节和结果可在附录 A.5中找到。

7 结论

本文介绍 StatCan 对话数据集，这是一个新型语料库，包含 2 万条+ 加拿大统计局线上访客与statcan.gc.ca操作员之间的英语和法语对话。基于该数据集，我们提出了两项旨在帮助用户查找特定数据表的任务：表格检索任务和响应生成任务。对于表格检索任务，我们测试了多种 DPR 和 TAPAS 变体，发现 DPR 的表现远胜于对应的 TAPAS 变体，也优于 BM25 基线。对于响应生成任务，我们研究了经过微调的 T5-large 模型，并探索了将 DPR 检索到的表格标题加入输入的不同变体。我们发现，检索增强型 T5 模型返回正确表格的频率更高，尽管其标题准确率

仍低于 DPR 检索器对应的召回率。这一结果表明，生成模型难以判断何时应返回表格。我们还发现，检索和生成难以推广到未来的对话，因为时间测试集划分显示，验证集与测试集之间存在显著的性能差距。总而言之，我们认为，这些任务对当前可用的模型提出了重大挑战，并鼓励研究社区进一步探索该数据集，构建能够帮助加拿大统计局用户的对话模型。

局限性

任务和模型的局限性 检索任务中的表格源自 statcan.gc.ca，这意味着其内容主要涉及加拿大人口统计数据[2]，并由加拿大统计局员工进行专业编辑。此外，生成任务专门用于根据检索到的表格生成高度忠实的回复，因此不应将此任务直接用于非预期用途或非研究场景（例如部署虚拟助理），因为这可能导致幻觉，进而对利益相关方造成负面影响。此外，这些局限性也可能体现在我们训练的模型中，因此发布时我们会在模型卡中说明这些局限性（Mitchell 等人，2019年）。

环境影响 我们承认，第5节中的模型使用了能耗显著的硬件。我们特意选择了规模适中、可在一块 GPU 上复现的模型。此外，我们使用的硬件由可再生能源供电。

制品和计算实验 我们根据模型的预期用途，使用相应的软件库进行训练，并将依据原始许可证发布相关制品。实验的计算细节详见附录 A.4。

伦理声明

隐私和数据访问 如 第 1 节所述，我们为移除个人可识别信息（PII）付出了大量努力。不过，我们无法排除处理过程中遗漏某些细节的可能性。因此，任何希望使用数据的用户都必须通过机构数据存储库进行身份验证并接受使用条款；使用条款要求用户报告任何个人身份信息泄露事件，我们会通过更新数据集将其删除。此外，我们要求将任何衍生作品或修改内容与原始使用条款一并发布在同一数据存储库中，并保留或扩展原有许可证。

在线话语中的有害内容风险 加拿大统计局工作人员接受过培训，能够以专业方式与在线用户开展工作。此外，由于用户以匿名和虚拟方式访问 statcan.gc.ca，因此更有可能观察到有害的在线去抑制现象（Lapidot-Lefler andBarak，2012年），这可能导致用户话语中出现有害内容。因此，我们请数据集用户报告对话中的任何有害内容，我们将以审查个人身份信息泄露的相同方式进行审查。

[2]更多信息请见：https://www.statcan.gc.ca/en/subjects-start/population_和_人口统计

致谢

感谢我们在加拿大统计局的合作者为我们提供数据集、指导我们了解表格的技术细节，并分享对本项目的宝贵反馈。感谢 Sivan Milton 就对话行为和分析进行有益的讨论。

参考文献

VaibhavAdlakha、Shehzaad Dhuliawala、KaheerSule- man、Harm de Vries 和 Siva Reddy。2022年。Topi- OCQA：通过主题切换进行开放域对话式问题回答。计算语言学协会会刊。

Raviteja Anantha、Svitlana Vakulenko、Zhucheng Tu、Shayne Longpre、Stephen Pulman 和 Srinivas Chappidi。2021年。开放域问题回答通过问题改写变得更具对话性。载于 *2021*年北美分会计算语言学协会会议论文集：人类语言技术，第 520–534 页，线上。计算语言学协会。

PayalBajaj、DanielCampos、NickCraswell、Li Deng、Jianfeng Gao、Xiaodong Liu、Rangan Majumder、Andrew McNamara、Bhaskar Mitra、Tri Nguyen、Mir Rosenberg、Xia Song、Alina Stoica、Saurabh Tiwary 和 Tong Wang。2018年。MS MARCO：一个由人工生成的机器阅读理解数据集。*arXiv:1611。09268 [cs]*。ArXiv：1611.09268。

Satanjeev Banerjee 和 Alon Lavie。2005年。METEOR：一种机器翻译评价自动评价指标，与人工判断的相关性有所提高。载于*ACL*研讨会：机器翻译和／或摘要的内在与外在评价指标论文集，第65–72页，密歇根州安阿伯。计算语言学协会。

Paweł Budzianowski、Tsung-Hsien Wen、Bo-Hsiang Tseng、Iñigo Casanueva、Stefan Ultes、Osman Ra- madan 和 Milica Gaši´。2018年。MultiWOZ——一个用于面向任务的对话建模的大规模多领域Wizard-of-Oz数据集。载于 *2018*年自然语言处理实证方法会议论文集，第5016–5026页，比利时布鲁塞尔。计算语言学协会。

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 和 David Traum。2010 年。迈向对话行为标注的 ISO 标准。载于 第七届国际语言资源与评估会议（*LREC'10*）。

Harry Bunt、Volha Petukhova、Emer Gilmartin、Cather- ine Pelachaud、Alex Fang、Simon Keizer 和 Lau- rent Prevot。2020年。对话行为标注的 ISO 标准。载于第*12*届语言资源与评测会议（*LREC 2020*年），第549–558页。欧洲语言资源协会（ELRA）。

Derek Chen、Howard Chen、Yi Yang、Alexander Lin 和 Zhou Yu。2021年。基于动作的对话数据集：用于构建更深入的任务导向对话系统的语料库。载于*2021*年计算语言学协会北美分会会议论文集：人类语言技术，第3002–3017页，线上。计算语言学协会。

Tianqi Chen、Bing Xu、Chiyuan Zhang 和 Carlos Guestrin。2016年。以次线性内存成本训练深度网络。*arXiv:1604.06174 [cs]*。ArXiv：1604.06174。

Wenhu Chen、Ming-Wei Chang、Eva Sc hlinger、William Yang Wang 和 William W. Cohen。2020年。表格与文本上的开放域问答。收录于国际学习表征会议论文集。

Eunsol Choi、He He、Mohit Iyyer、Mark Yatskar、Wen-tau Yih、Yejin Choi、Percy Liang 和 Luke Zettlemoyer。2018年。QuAC：情境中的问答。收录于*2018*年自然语言处理实证方法会议论文集，第2174–2184页，比利时布鲁塞尔。计算语言学协会。

Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 KristinaToutanova。2019年。 BERT：预训练用于语言理解的深度双向 Transformer。载于 *2019*年计算语言学协会北美分会会议论文集：人类语言技术，第*1*卷（长篇与短篇论文），第4171–4186页，明尼阿波利斯，明尼苏达州。计算语言学协会。

Martin d'Hoffschmidt、Wacim Belblidia、Quentin Heinrich、Tom Brendlé 和 Maxime Vidal。2020年。FQuAD：法语问答数据集。收录于计算语言学协会论文集：*EMNLP 2020*年论文选，第1193–1208页，线上。计算语言学协会。

Emily Dinan、Stephen Roller、Kurt Shuster、Angela Fan、Michael Auli 和 Jason Weston。2018年。维基百科巫师：知识驱动的对话代理。载于 国际学习表征会议。

Etalab Lab-IA。2021年。在法语问答上训练的 DPR CamemBERT。

Song Feng、Siva Sankalp Patel、Hui Wan 和 Sachindra Joshi。2021年。MultiDoc2Dial：基于多个文档的对话建模。载于 *2021*年自然语言处理实证方法会议论文集，第 6162–6176 页，线上会议，蓬塔卡纳，多米尼加共和国。计算语言学协会。

Song Feng、Hui Wan、Chulaka Gunasekara、Siva Patel、Sachindra Joshi 和 Luis Lastras。2020年。doc2dial：面向目标的文档基础对话数据集。载于 *2020*年自然语言处理实证方法会议（*EMNLP*）论文集，第 8118–8128 页，线上会议。计算语言学协会。

Timnit Gebru、Jamie H. Morgenstern、Briana Vec- chione、Jennifer Wortman Vaughan、Hanna M. Wallach、Hal Daumé 和 Kate Crawford。2021年。数据集说明书。*ACM* 通讯，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 和 Haifeng Wang。2018年。DuReader：来自真实应用场景的中文机器阅读理解数据集。载于问答机器阅读研讨会论文集，第37–46页，澳大利亚墨尔本。计算语言学协会。

Charles T. Hemphill、John J. Godfrey 和 George R. Doddington。1990 年。ATIS 口语系统试点语料库。载于语音与自然语言研讨会论文集——*HLT '90*，第 96–101 页，宾夕法尼亚州 Hidden Valley。计算语言学协会。

JonathanHerzig、ThomasMüller、SyrineKrichene 和 Julian Eisenschlos。2021年。开放域问答

通过稠密检索回答表格问题。收录于*2021*年计算语言学协会北美分会年会论文集：人类语言技术，第 512–519 页，线上。计算语言学协会。

Jonathan Herzig、Pawel Krzysztof Nowak、Thomas Müller、Francesco Piccinno 和 Julian Eisenschlos。2020年。TaPas：通过预训练进行弱监督表格解析。载于计算语言学协会第*58*届年度会议论文集，第4320–4333页，线上。计算语言学协会。

Armand Joulin、Edouard Grave、Piotr Bojanowski 和 Tomas Mikolov。2017年。高效文本分类技巧集。载于欧洲计算语言学协会分会第*15*届会议论文集：第*2*卷，短篇论文，第427–431页，西班牙瓦伦西亚。计算语言学协会。

Ali Kabbadj。2021年。法语SQuAD：用于问答的法语机器阅读。原始日期：2019-09-22T17:54:45Z。

VladimirKarpukhin、Barlas Oguz、Sewon Min、Patrick Lewis、Ledell Wu、Sergey Edunov、Danqi Chen 和 Wen-tau Yih。2020年。用于开放域问答的稠密段落检索。载于*2020*年自然语言处理实证方法会议（*EMNLP*）论文集，第6769–6781页，线上。计算语言学协会。

RachelKeraron、Guillaume Lancrenon、Mathilde Bras、Frédéric Allary、Gilles Moyse、Thomas Scialom、Edmundo-Pavel Soriano-Morales 和 Jacopo Sta- iano。2020年。PIAF项目：构建法语原生问答数据集。载于第十二届语言资源与评测会议论文集，第5481–5490页，马赛，法国。欧洲语言资源协会。

Tom Kwiatkowski、Jennimaria Palomaki、Olivia Red- field、Michael Collins、Ankur Parikh、Chris Al- berti、Danielle Epstein、Illia Polosukhin、Jacob De- vlin、Kenton Lee、Kristina Toutanova、Llion Jones、Matthew Kelcey、Ming-Wei Chang、Andrew M. Dai、Jakob Uszkoreit、Quoc Le 和 Slav Petrov。2019年。Natural Questions：问答研究基准。计算语言学协会会刊，7：453–466。

Noam Lapidot-Lefler 和 Azy Barak。2012年。匿名性、不可见性和缺乏眼神接触对线上有毒性去抑制效应的影响。《计算机与人类行为》，28(2):434–443。

Harrison Lee、Raghav Gupta、Abhinav Rastogi、Yuan Cao、Bin Zhang 和 Yonghui Wu。2022年。SGD-X：模式引导的对话系统中鲁棒泛化的基准。载于 *AAAI*人工智能会议论文集。

Quentin Lhoest、Albert Villanova del Moral、Yacine Jernite、Abhishek Thakur、Patrick von Platen、Suraj Patil、Julien Chaumond、Mariama Drame、Julien Plu、Lewis Tunstall 等。2021年。Datasets：自然语言处理社区库。载于*2021*年自然语言处理实证方法会议论文集：系统演示，第175–184页。

Chin-Yew Lin。2004年。ROUGE：一种用于自动评估摘要的工具包。载于文本摘要的拓展，第74–81页，西班牙巴塞罗那。计算语言学协会。

Chin-Yew Lin 和 Franz Josef Och。2004年。使用最长公共子序列和跳跃二元组统计自动评估机器翻译质量。载于 第*42*届计算语言学协会年度会议论文集 - *ACL '04*，第605–es页，西班牙巴塞罗那。计算语言学协会。

Ilya Loshchilov 和 Frank Hutter。2017年。解耦权重衰减正则化。发表于 国际学习表示会议。

Ryan Lowe、Nissan Pow、Iulian Serban 和 Joelle Pineau。2015年。Ubuntu对话语料库：用于研究非结构化多轮对话系统的大型数据集。载于话语与对话特别兴趣小组第*16*届年会论文集，第285–294页，捷克共和国布拉格。计算语言学协会。

Marco Lui 和Timothy Baldwin。2011年。用于语言识别的跨领域特征选择。载于第五届自然语言处理国际联合会议论文集，第553–561页，泰国清迈。亚洲自然语言处理联合会。

Louis Martin、Benjamin Muller、Pedro Javier Ortiz Suárez、Yoann Dupont、Laurent Romary、Éric de la Clergerie、Djamé Seddah 和 Benoît Sagot。2020年。CamemBERT：一种美味的法语语言模型。载于第*58*届计算语言学协会年会论文集，第7203–7219页，线上。计算语言学协会。

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji 和 Timnit Gebru。2019年 模型卡：模型报告。载于公平性、问责制与透明度会议论文集，FAT* '19，第 220–229 页，纽约州纽约，美国。计算机协会。

Ankur Parikh、Xuezhi Wang、Sebastian Gehrmann、Manaal Faruqui、Bhuwan Dhingra、Diyi Yang 和 Dipanjan Das。2020年。ToTTo：一个受控的表格到文本生成数据集。收录于《*2020*年

自然语言处理实证方法会议（*EMNLP*），第1173–1186页，线上。计算语言学协会。

Chen Qu、Liu Yang、Cen Chen、Minghui Qiu、W. Bruce Croft 和 Mohit Iyyer。2020年。开放检索式对话问答。载于第 *43* 届国际 *ACM SIGIR* 信息检索研究与发展会议论文集，SIGIR '20，第 539–548 页，美国纽约州纽约市。美国计算机协会。

Colin Raffel、Noam Shazeer、Adam Roberts、Katherine Lee、Sharan Narang、Michael Matena、Yanqi Zhou、Wei Li 和 Peter J. Liu。2020年。探索迁移学习的极限：统一的文本到文本转换器。机器学习研究杂志，21(140)：1–67。

Abhinav Rastogi、Xiaoxue Zang、Srinivas Sunkara、Raghav Gupta 和 Pranav Khaitan。2020年。迈向可扩展的多领域对话智能体：模式引导对话数据集。载于*AAAI* 人工智能会议论文集，第 34 卷，第 8689–8696 页。

Siva Reddy、Danqi Chen 和 Christopher D. Manning。2019年。CoQA：对话式问题回答挑战。计算语言学协会会刊，7：249–266。

Radim Rehurek 和 Petr Sojka。2010年。面向大型语料库的主题建模软件框架。载于*LREC 2010* 自然语言处理框架新挑战研讨会论文集，第 45–50 页。

Stephen Robertson 和 Hugo Zaragoza。2009年。概率相关性框架：BM25 及其扩展。信息检索基础与趋势®，3(4)：333–389。

Yossi Rubner、Carlo Tomasi 和 Leonidas J. Guibas。2000年。地球移动距离作为图像检索的度量。国际计算机视觉期刊，40(2)：99–121。

Marzieh Saeidi、Max Bartolo、Patrick Lewis、Sameer Singh、Tim Rocktäschel、Mike Sheldon、Guillaume Bouchard 和 Sebastian Riedel。2018年。对对话式机器阅读中的自然语言规则的解释。载于*2018*年自然语言处理实证方法会议论文集，第 2087–2097 页，比利时布鲁塞尔。计算语言学协会。

Noam Shazeer 和 Mitchell Stern。2018年。Adafactor：具有次线性内存开销的自适应学习率。载于第*35*届国际机器学习会议论文集，第80卷，机器学习研究论文集，第4596–4604页。PMLR。

Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Łukasz Kaiser 和 Illia Polosukhin。2017年。Attention is Allyou Need。载于神经信息处理系统进展，第30卷。Curran Associates, Inc.

Ellen M. Voorhees。2001年。TREC 问答赛道。自然语言工程，7(4)：361–378。

Harm de Vries、Dzmitry Bahdanau 和 Christo- pher Manning。2020年。迈向对语言用户界面的生态效度研究。*arXiv:2007。14435 [cs]*。ArXiv：2007.14435。

Thomas Wolf、Lysandre Debut、Victor Sanh、J ulien Chaumond、Clement Delangue、Anthony Moi、Pier- ric Cistac、Tim Rault、Remi Louf、Morgan Funtow- icz、Joe Davison、Sam Shleifer、Patrick von Platen、Clara Ma、Yacine Jernite、Julien Plu、Canwen Xu、Teven Le Scao、Sylvain Gugger、Mariama Drame、Quentin Lhoest 和 Alexander Rush。2020年。Transformers：自然语言处理领域的先进技术。载于*2020*年自然语言处理实证方法会议论文集：系统演示，第38–45页，线上。计算语言学协会。

Linting Xue、Noah Constant、Adam Roberts、Mi- hir Kale、Rami Al-Rfou、Aditya Siddhant、Aditya Barua 和 Colin Raffel。2021年。mT5：一种大规模多语言预训练文本到文本Transformer。载于*2021*年计算语言学协会北美分会人类语言技术会议论文集，第483–498页，线上。计算语言学协会。

Yi Yang、Wen-tau Yih 和 Christopher Meek。2015年。WikiQA：开放域问答挑战数据集。载于*2015*年自然语言处理实证方法会议论文集，第2013–2018页，里斯本，葡萄牙。计算语言学协会。

Pengcheng Yin、Graham Neubig、Wen-tau Yih 和 Sebastian Riedel。2020年。TaBERT：文本与表格数据联合理解的预训练。载于计算语言学协会第*58*届年度会议论文集，第8413–8426页，线上。计算语言学协会。

张天翼、Varsha Kishore、Felix Wu、Kilian Q. Weinberger 和 Yoav Artzi。2019年。BERTScore：使用 BERT 评估文本生成。载于国际学习表征会议。

赵伟、Maxime Peyrard、刘飞、高阳、Chris- tian M. Meyer 和 Steffen Eger。2019年。MoverScore：使用上下文化嵌入和推土机距离评估文本生成。载于*2019*年自然语言处理实证方法会议暨第九届自然语言处理国际联合会议（*EMNLP-IJCNLP*）论文集，第563–578页，香港，中国。计算语言学协会。

A 附录

A.1 完整用户意图

为深入了解加拿大统计局在线聊天所提供的帮助类型，我们对对话进行了定性分析，并在下文列举主要用户意图的示例。

查找数据表聊天功能的主要意图是帮助用户找到特定的数据表。例如，一位用户想查找蒙特利尔某些地区 2012年至2016年的人口数量。这是我们工作的重点。

获取表格的元信息有些用户并非要查找数据表，而是想了解特定表格的元信息。例如，一位访客想知道下一次人口普查何时发布。另一位用户则想了解劳动力可用性（WFA）和劳动力市场可用性（LMA）的定义。

操作数据表有些用户希望以不同的格式或呈现方式获取数据表。例如，一位用户正在查看某个特定数据表，并询问能否查看年度值，而不是月度值。

获取用户界面方面的帮助有些用户在寻求用户界面方面的帮助。例如，一位用户想下载某个特定数据表，却找不到下载链接。

域外请求 我们发现，许多对话超出了加拿大统计局在线聊天的服务范围。例如，有用户询问将某种特定产品运往国外需要提供哪些文件。

第一种意图见 表1，后续意图见 表18。我们认为，所有意图都是对话研究中值得探索的方向，但我们重点关注表格检索意图，因为（i）这类对话数量众多；（ii）任务是否成功有明确的衡量标准，即是否检索到了正确的表格。在本文余下部分，我们研究智能体返回表格网址的对话。

A.2 补充表格说明

A.2.1 格式和大小

完整表格以长 格式存储，其中每列对应一个维度，除了最后一列对应价值。数量行数对应于各维度成员的笛卡尔积。这意味着，对于 n_i 个成员和 d 个维度，完整表格中总计有 $\prod_{i=0}^{d} n_i$ 行。当表格包含大量成员项时，完整表格可能会变得极其庞大。此外，statcan.gc.ca 上显示的表格通常是完整表格经过透视和筛选后的视图，这意味着某些成员会成为列，其他成员会成为行，还有许多成员会被直接省略。

A.2.2 详细规范

本节为 第3.3节 提供补充细节。

产品标识符（PID）分配给每个已发布数据表的唯一 8 至 10 位标识符。虽然智能体可能会分享其他类型的表格，但官方数据表始终会有 PID；因此，URL 中没有 PID 的任何表格都不纳入本任务。前两位数字（1-2）代表与该表格相关的主题代码（可在基本信息中找到），接下来的第 3-4 位代表产品类型，在本任务中，产品类型为表格，通常编码为“10”。第 4-8 位数字是唯一标识符，用于标识该主题下的相应表格。第 9-10 位数字表示表格视图，通常为“01”，代表默认视图；构建 URL 时需要这些数字，但它们是可选的，并且在发布的数据集中会被省略。

成员项 各个表格的标签，可通过透视操作表示为列索引或行索引。每个表格的成员项都与其他表格不同。

维度 互不重叠的成员项集合；每个成员项都必须属于某个维度。对于包含 d 个维度的表格，每个数据值都与这些维度的笛卡尔积中的一个 d 元组相关联，但该积中的并非每个元组都会有关联的数据值（如果未记录该数据值，或该元组无效）。

基本信息 由标题、日期范围、频率、维度、主题类别和调查组成的元数据项[3]，数据源自该调查。

[3]调查列表及详细信息链接可在此处找到：https://www.statcan.gc.ca/en/survey/list

脚注 当需要补充说明时，表格中通常会包含非结构化注释；这些注释通常与需要进一步解释的特定成员项相关联。

完整表格 对于所有具有 PID 的表格，完整表格（以 CSV 文件形式）、完整元数据及其基本信息都会作为数据发布的一部分提供，并可用于我们提出的任务。示例表格见第3.3节。

A.2.3 表格更新与归档

加拿大统计局的表格有时会定期更新，有时则可能只发布一次。当表格停止更新且信息过时时，会被标记为已归档。在某些情况下，归档表格可能会创建一个有重大更改（例如新增列）的新版本。如表格 14所示，表格的发布没有预先确定的时间表，但自 2000 年以来，平均发布数量一直在增加，并在 2017年出现高峰，这主要是由于发布了许多健康相关表格；2021年也出现高峰，这是由于发布了许多劳动力、科学和收入表格，如图 6所示。此外，图 5 显示，2019年到2021年间发布的大多数表格仍为最新（当前）状态，而此前发布的大多数表格都已归档。

A.3 补充统计数据

A.3.1 高频请求表格

最常见的表格汇总见 表21。其中，2个表格的数据来源于消费者价格指数调查（通常用于追踪通货膨胀），另外4个表格涵盖收入、人口统计、企业绩效和犯罪等更一般、更广泛的主题。这4个表格分别来自不同的调查。一个表格每月更新，另一个每半年更新一次，其余表格每年更新。最早的表格于2019年更新，当时对话开始被记录。

A.3.2 表格频率统计

在表格 3和 12中，我们可以计算出，英语对话中每张表格平均返回 5.68 次（标准差为 12.86），法语对话中平均返回 4.25 次（标准差为 9.73）。因此，最常被请求的表格所占比例远高于较少受欢迎的表格，而且表格被使用的次数方差很大。事实上，有 294 张表格只在任一数据划分中出现过一次（即 28.9%）。

A.3.3 细粒度对话统计

在 表13中，我们发现，消息数和轮次数围绕平均值波动很大，超过68%的对话持续2至7轮。在极端情况下，一段对话可持续多达28轮。此外，我们还注意到，尽管大多数消息约含32个词元，但最长的消息可多达1374个词元；在这些情况下，我们会看到智能体撰写大量文本，有时还会在回复用户时复制并粘贴大段内容（例如，从模板数据库中复制）。尽管这些内容通常会在一轮内通过多个连续消息发送，但智能体有时也会选择一次性全部发送。

A.4 实现细节

本节介绍了实现第5节中模型的细节。

实现 Transformer 模型 所有基于 Transformer 架构的模型（Vaswani 等人，2017年）均使用 Hugging Face 的库实现（Wolf 等人，2020年）。

BM25 为便于复现，我们使用 Gensim 实现了该模型（Rehurek 和 Sojka，2010年）。

DPR 和 TAPAS-NQ 我们使用了 DPR 的基础版本和 TAPAS-NQ 的大型版本。我们采用了在 Natural Questions（Kwiatkowski 等人，2019年）（NQ）的 32 万个问题上训练的 DPR 检查点。训练期间，使用 AdamW（Loshchilov 和 Hut- ter，2017年）以 10^{-5} 的学习率优化网络，权重衰减设为零。根据原始研究，网络训练了 30 个训练轮次，每批包含 64 个查询、正段落和困难负段落（后者通过 BM25 检索）。采用批内负采样来增加负例。为确保结果可复现，网络在单个 32GB GPU 上训练，并使用梯度检查点（Chen 等人，2016年）以减少内存占用。对话长度为 512 个词元，元数据词元长度分别为：标题 128 个、基本信息（定义见第3.3节）256 个，其余内容 512 个。

T5 我们使用 Adafactor 优化器（Shazeer 和 Stern，2018年），学习率为 0.001。我们使用批次大小为 16，并进行 8 步梯度累积和梯度检查点，以复现原始实现中 128 个样本的批次大小。模型在单个 32GB GPU 上训练了 10 个训练轮次。源序列和目标序列长度分别为 512 和 256 个词元；为确保截断后保留最新消息，源序列从右侧截断。我们遵循原始实现，使用束宽 4 和 0.6 的长度惩罚。

训练时对于英语数据集划分，DPR 的每个变体都可在 A100 GPU 上用 68 分钟完成训练。TAPAS-NQ 的每个大型变体都需要在 V100 GPU 上训练 15 小时。T5 的每个变体都可在 V100 GPU 上用 16 小时完成训练。复现所有 DPR 结果需要 8 小时，TAPAS 需要 90 小时，T5 需要 96 小时；法语划分所需时间则按比例更少。

A.5 法语子集建模

基本统计数据英语数据集划分中，每段对话的消息数从 2 条到 59 条不等，中位数为 11 条（分布见图7）。根据表12，平均每段对话有 3.9 轮次，但包含 12.3 条消息。平均而言，每条消息包含超过 30 个词元（使用 T5 分词器）。

语言划分为了确定每段对话所使用的语言，我们使用了两个常见的语言识别库：*langid.py*（Lui 和 Baldwin，2011年）以及一个针对语言检测进行微调的 fastText 网络（Joulin 等人，2017年）。将模型应用于每段对话后，我们只保留语言标签一致的对话（两者都判定为英语或都判定为法语）。

训练集和评估 法语子集的训练流程和评估完全遵循以下指定的任务第4节。

响应生成建模 我们没有使用 T5，而是采用了由 Xue 等人(2021年)提出的多语言 T5 模型，因为它能够自然地处理法语文本。

检索建模 我们使用了一个基于 CamemBERT 改编的 DPR 变体（Martin 等人，2020年），并使用三个法语问答数据集进行训练（Keraron 等人，2020年；d'Hoffschmidt 等人，2020年；Kabbadj，2021年），由 EtalabLab-IA (2021年)。

检索结果 在 表17中，我们观察到，与英语数据集划分不同，将成员项添加到基本信息或标题中会提升验证结果，但不会提升测试结果，这可能表明模型出现了过拟合。不过，我们注意到不同运行之间的方差很大，因此难以确定成员项是否有帮助。过拟合和高方差可能都是由训练集规模较小所致。此外，BM25 在任何元数据视图上的表现都极差，这也可能与数据集规模有关。

生成结果 在 表 33中，我们注意到所有指标均显著下降，标题准确率始终为 0%。这可能是因为法语划分显著更小，但复杂度仍与英语数据集划分相当，因此 mT5 难以对其进行建模。就标题准确率而言，我们在法语测试集划分中发现了 55 个目标文本中包含标题（即由智能体返回标题）的实例。然而，在其中 54 个案例中，增强版 mT5 返回了通用回复（例如，“请稍候，我正在进行搜索。”），而非预期的标题。这表明 mT5 无法判断何时适合返回标题，也无法生成非模板化回复。

A.6 T5 生成的回复

在本节中，我们从验证集中选取了几段对话，并考察 T5 以及结合 DPR 检索到的标题的 T5（T5+D）生成的消息。

常见和不常见的回复 在 表 23中，我们注意到，T5 和 T5+D 都能够生成常见的言语行为，例如“谢谢”和“请稍候……”，但面对陌生情境时却难以应对（例如，用户意外断开连接后，需要请求对方澄清）。

常见表格 在 表24中，在检索到的表格（R_i）标题中，第一个标题部分正确。T5 和 T5+D 都补全了标题，并输出了所需的标识符，与预期的智能体回复一致；该表格是训练集中最受欢迎的表格之一（见表 21）。

多个表格、日期选择 在 表 28中，我们注意到 T5 只返回了智能体返回的两个表格中的一个。另一方面，DPR 检索到了正确的表格，但 T5+D

未能选出日期正确的表格（它选中了2019年6月，而不是2019年12月），但除此之外选出的表格都相关。

解释的详尽程度 T5+D 还提供了一段解释，并链接到相关资源，包括加拿大统计局网站上的资源（非表格类）和外部资源。这是因5+D 在训练过程中记住了这些信息，只需将表格替换为更新后的日期（表27）。

罕见表格 在表格 25、 26、 29所示的对话中，检索到的表格在训练集中出现了 10 次，这比表24中检索到的表格少得多。对于 表25，智能体返回的表格并未被 DPR 检索到，导致 T5+D 返回了检索到的第一个标题。然而，在表26中，DPR 检索到了正确的标题（标题 # 4），但 T5+D 未能在生成的消息中使用该正确标题。至于表29，智能体通过指出“作为一项标准产品，我们有按行业划分的就业表格”提供了提示；DPR 正确识别了这一点，因为检索到的第二个标题与智能体回复中表格的 PID 完全匹配。然而，T5+D 再次未能选出正确标题，而是选择返回一条笼统的回复（“请稍候，我正在查找信息”）；T5 则生成了一个与其生成的标题不同的 PID（两者都不正确）。

训练期间未见过的表在训练集（见表11了解更多信息）中未出现的表格里，有七个在验证集或测试集中出现了三次或更多次（表32）。在表30中，我们看到一段对话，其中 DPR 检索到了正确的标题，且 T5+D 正确返回了该标题，而 T5 未能返回。另一方面，在表31中，尽管 DPR 也正确检索到了标题，T5+D 却未能返回该标题，这可能是因为智能体说了“不幸的是，……”，使其产生了误导。

A.7 统计检验

为支持第6节中的论述，我们进行了多个单尾 Welch t 检验，使用了以下数据中的均值和校正后的标准差：16。零假设是，在 3 次运行中，实验 A 的均值与实验 B 的均值不同。除非另有说明，我们使用测试集上的 R@1。

1. 结论：在标题中添加基本信息和成员项会导致 DPR 出现显著差异。 A 为仅使用标题的模型， B 使用基础 + 成员，我们的 p值为 0.014。

2. 声明：对于 DPR，使用成员项会导致结果大幅下降。设 A 为仅使用成员项的模型， B 使用标题 + 成员项，我们的 p值为 0.00086。

3. 声明：与完整表格相比，TAPAS-NQ 使用标题和成员项时表现更好。设 A 为使用完整表格的模型， B 使用标题 + 成员项，我们的 p值为 0.016。

4. 声明：在图4中，TAPAS、DPR 标题和 DPR 基础的验证集召回率@10 高于测试集 + 成员。其中 A 为验证分数， B 测试分数，p值分别为 $0。00197$， 2.18×10^{-5}，$0。00014$。

B 数据集卡片

本节介绍一张数据集卡片，其格式遵循Lhoest 等人 (2021年提出的方案，该方案的灵感来自Mitchell 等人 (2019年)和Gebru 等人(2021年)的工作。

摘要 StatCan 对话数据集由超过 20K+段对话组成，这些对话发生在加拿大统计局（StatCan）工作的智能体与访问 StatCan 网站并需要通过官方在线聊天系统获得支持的用户之间。

任务使用 1.9 万条对话轮次子集构建两个任务：

1. 根据正在进行的对话自动检索相关表格。对于每段部分对话，任务是返回智能体所返回的最可能表格的标识符。使用 recall @k 指标进行评估。2. 在每轮自动生成恰当的智能体回复。对于每段部分对话，任务是返回智能体最可能的回复，其中包括相关表格的链接。使用以下四项指标进行评估，相关指标详见 第4.2节。

排行榜 排行榜和提交说明可在项目网页上找到。每项提交都将附带一个标签，标明是否：

- 这是自行报告的；
- 这些提交内容经过了外部评估；
- 推理过程按照所提供的指示进行了复现；
- 完整的训练过程得到了独立复现。

语言 对话以加拿大英语（en-CA）和加拿大法语（fr-CA）进行。

B.1 数据集结构

B.1.1 数据实例

对话完整的对话示例见表19。各用户意图的示例见表18，带有对话行为标注的两个对话见表22。在A.6中的案例研究里，我们展示了表格24、25、26、27、28、29、30、31中的部分对话。

表格表格的完整元数据见表20，可通过以下网址访问：doi.org/10.25318/3210035401-eng。表 21展示了最受欢迎表格的基本信息。

B.1.2 数据字段

完整数据集提供了一个包含以下字段的 CSV 文件：

- 对话：以 JSON 格式表示的部分对话（返回表格之前）。
- 对话_索引：唯一索引，用于在此任务之外标识该对话。
- 对话_处理后：将对话转换为易读的文本格式，移除时间戳等额外信息，将 URL 替换为特殊标签，并添加分隔标签（</s>）。

检索任务为每个数据划分提供包含以下字段的 CSV 文件：

- 对话
- 对话索引_
- 对话_已处理
- 目标文本_pid：智能体返回的表格的产品 ID
- 语言：实时聊天系统报告的语言，由于标注错误，该语言可能并不总是准确。
- ft_检测到的_语言：由 fastText 预测的语言。
- ft_检测_语言：fastText 输出的分数。
- lid_检测_语言：langid.py 预测的语言。
- lid_检测_概率：langid.py 输出的分数。

元数据 检索过程中使用的元数据以 CSV 文件形式提供，其中包含以下字段（5907 个表格各对应一个字段）：

- pid：表格的产品ID
- 标题：表格的标题

• 基本信息：_文本格式中的基本信息

• 成员信息：以文本形式列举的成员项目-_

• 脚注信息：以文本形式列举的脚注-_

• 完整_信息：基本信息、成员项目和脚注均包含在单个条目中

•x_和_y：上述两个项目的组合，例如 x=标题和 y=脚注。

• * 法语：以上所有内容也提供_法语版本，并通过后缀_fr 表示。

生成任务 每个数据划分均提供包含以下字段的 CSV 文件：

• 源文本：等同于对话。

•源文本_处理后：等同于处理后的对话_。

• 目标文本：智能体在对话结束后撰写的消息。

• 目标文本_处理后：智能体在对话结束后撰写的消息，其中的 URL 已替换为特殊标签。

•对话索引_

每个用于对话任务的 CSV 文件都提供了一个增强变体，其中包含上述所有字段以及以下字段：

• 源文本_增强：内容与对话相同，并附加由在基本信息上训练的最佳 DPR 变体检索出的前5个表的标题。

• 目标文本_增强_1：内容与源文本_增强相同，但仅使用第一个表格，而非前5个表。

对话 JSON 格式对话采用以下 JSON格式：[

... {'timestamp': '2019年3月13日 17:03:22', '参与者': '用户', '参与者_姓名': '<NAME>', 'text': "我来看看", 'urls': []}, {'timestamp': '2019年3月13日 17:04:12', '参与者': '操作员', '参与者_姓名': 'Kelly C', 'text': '建筑许可证...', 'urls': ['https:...']}, ...]

B.1.3 数据划分

检索数据划分包含以下数量的样本：

• 训练集：3782（英语）；869（法语）

• 验证集：799（英语）；201（法语）

• 测试集：870（英语）；141（法语）

它们对应于表3和表12中的查询数量，因为每个查询都会检索一个表格，而在一次对话中可能会发生多次检索。在这种情况下，部分对话会在对话中的不同轮次处截断。

生成数据划分的样本数量如下：

• 训练集：21582（英语）；3977（法语）

• 验证集：4464（英语）；861（法语）

• 测试集：4850（英语）；884（法语）

样本数量少于表3 和表12 中的消息数量，因为该任务的目标只是预测智能体将撰写的消息。与检索任务一样，部分对话将在对话中的不同轮次处截断。

B.2 数据集创建

B.2.1 筛选理由

第1节 详细阐述了数据集策划的动机。总结如下，我们列出了要点：

• 来自真实用户的数据：我们希望构建一个能够捕捉现实世界中存在的语言挑战的数据集

• 面向任务的对话：我们希望设计任务，具体目标是帮助在线聊天用户查找统计数据。

• 真实世界应用：我们的模型可直接应用于其他希望建立聊天系统的统计局，而我们的数据集将对具备以下条件的任何组织有所帮助：(1) 拥有聊天系统；(2) 拥有表格数据库。

• 多语言对话：我们希望构建能够处理英语以外语言的模型，因此也提供法语版本的任务。

B.2.2 源数据

对话 这些数据取自 statcan.gc.ca 上的实时聊天系统，并由加拿大统计局的开发团队进行了匿名化处理。对话发生于2019年3月至2021年3月期间。

表格这些表格可公开获取，下载时遵循了网络数据服务用户指南中的说明：statcan.gc.ca/en/developers/wds/user-guide。这些表格由加拿大统计局于2000年至2021年间发布。数据要么由加拿大统计局直接收集（例如通过人口普查或调查），要么汇编自现有源数据（例如私营部门组织和政府机构提供的数据），用于形成官方统计数据。现有源数据包括：

• 行政数据：由政府或私营部门在日常运营过程中收集，包括出生和死亡记录、税务、边境管制以及卫星数据。

• 微观数据链接：将现有信息关联起来，以创建新的数据。出于隐私保护目的，现有信息可能并非总是公开可用，因此，链接操作可以在保护公众保密性的同时，补充此前无法获得的新信息。

• 开放数据：以机器可读格式提供且可免费获取的数据，来源于各种渠道（例如 OpenStreetMap）。

• 网络抓取：由加拿大统计局从互联网抓取的数据（不包括个人信息以及“不会用于生成统计产出的任何信息”）。

B.2.3 注释

该数据集不包含通过实时聊天系统收集并纳入表格元数据的内容以外的任何其他注释。

B.2.4 个人与敏感信息 n

对话中的个人和敏感信息已通过程序移除，正式发布的表格仅包含能够保护参与者保密性的信息汇总。尽管移除流程非常先进，但仍存在非零概率，即某些信息可能被用于重构用户的个人档案。因此，研究人员访问这些数据时，必须先注册并同意使用条款；任何衍生作品都必须在同一平台上分享，并附带相同的条款。

B.3 使用数据时的考量 B.3.1 数据集的社会影响

该数据集旨在开发和评估模型，帮助知识工作者从数据源中查找相关表格。通过提供一种专用检索系统，其返回结果比通用系统更相关，从而提高知识工作者的工作效率。对于公共机构和统计局而言，这将使许多对某个特定社区相关统计数据感兴趣的在线聊天用户受益。

B.3.2 偏见讨论

如第 7 节所述，在线话语中始终存在有害内容风险，这意味着在线用户可能会在其消息中表现出负面偏见。不过，StatCan 工作人员的职责是以专业的方式与在线用户沟通。因此，研究人员不应使用此数据集来构建生成在线用户撰写的消息的模型，也不应在偏见可能对利益相关方造成负面影响的情境中使用基于该数据集训练的任何模型。

B.3.3 其他已知局限性

第 7 节介绍了其他已知局限性。

B.4 补充信息

B.4.1 数据集整理者

该数据集由本文作者根据加拿大统计局开发人员和智能体收集并处理的原始数据整理而成。

B.4.2 许可信息

这些对话采用自定义许可，有意访问该对话的研究人员需要先接受该许可。表格依据加拿大统计局开放许可发布：stat-can.gc.ca/en/reference/licence。

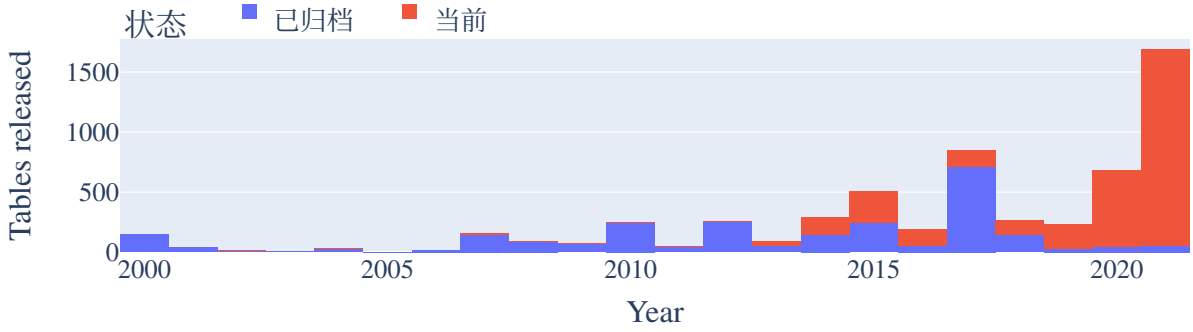

图 5：所有可用表格的发布日期，来源：statcan.gc.ca

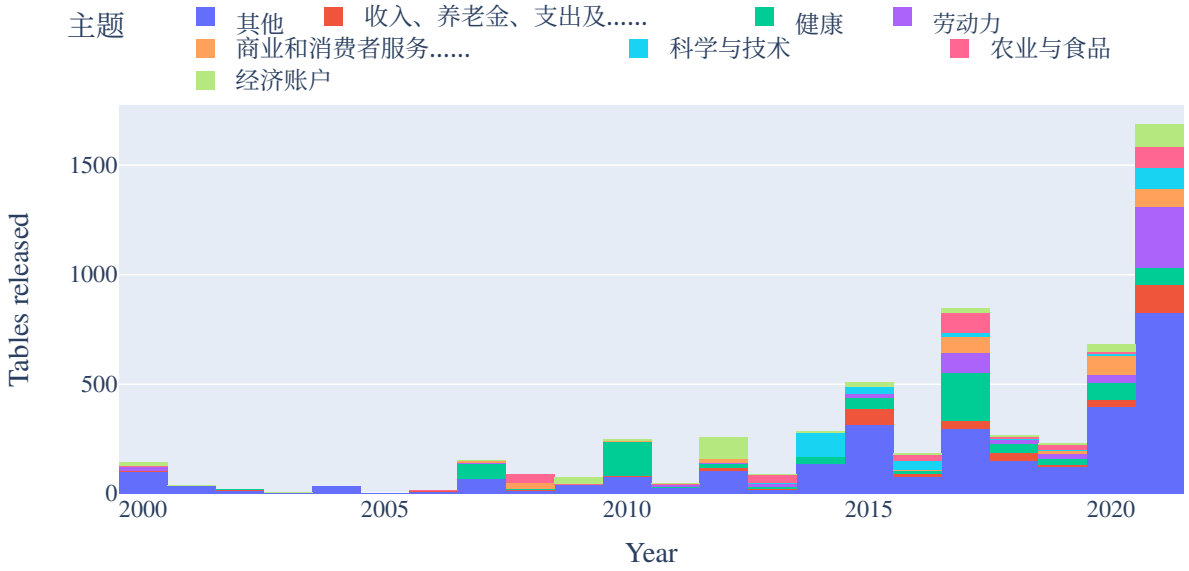

图 6：按主题划分的所有表格的发布日期。为便于阅读，仅显示排名前 8 的主题。

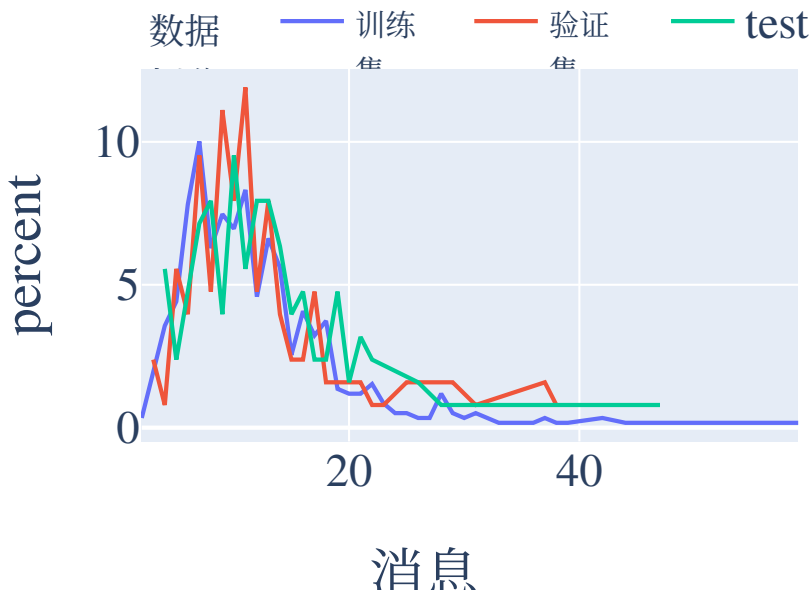

图7：法语任务数据集划分中每段对话的消息直方图。英语数据集划分见图 1.

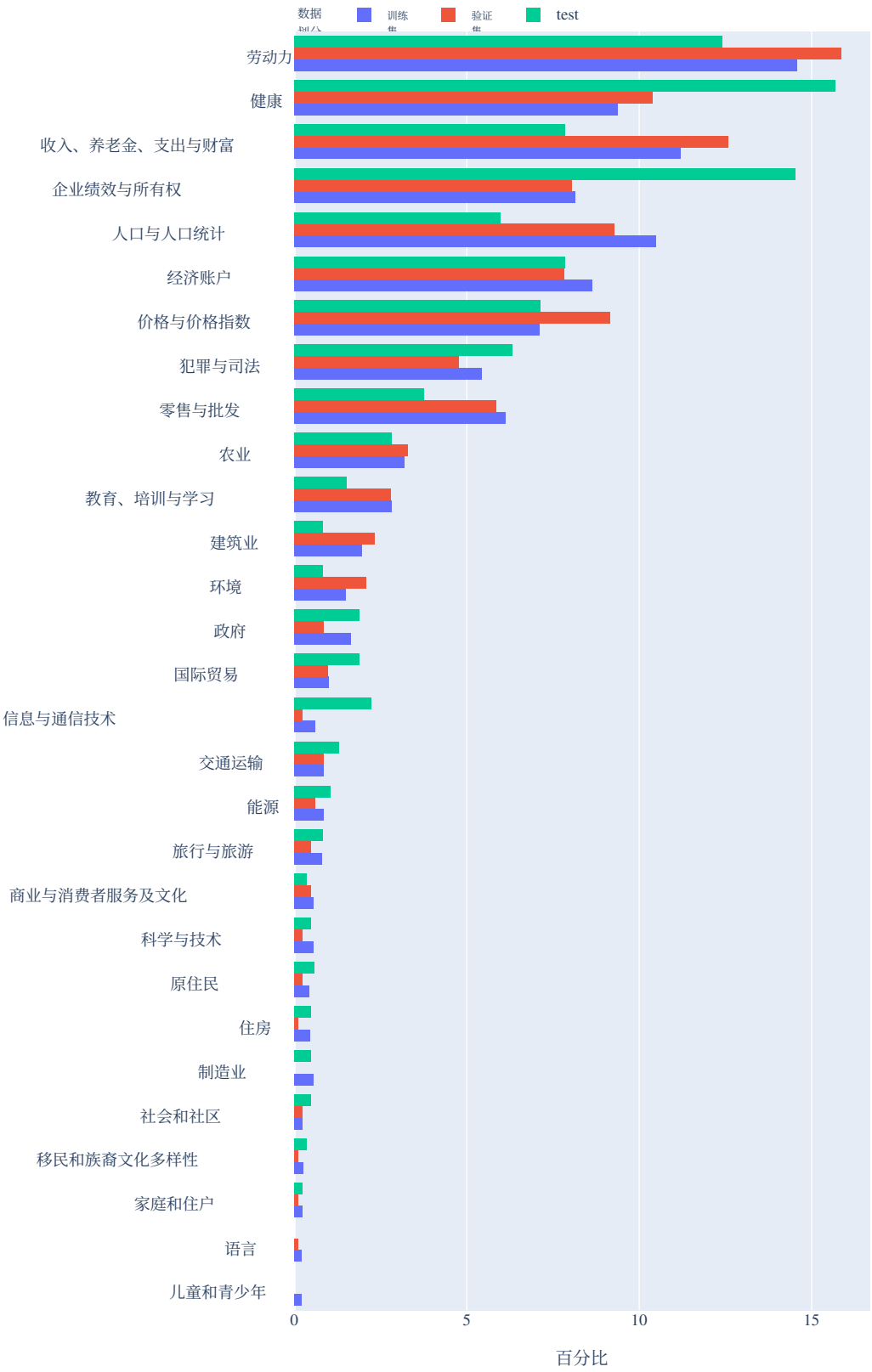

图8：检索任务中表格主题类别的分布，按数据划分着色。

	训练集	验证集	Test	总体
训练集	0	470	535	0
验证集	41	0	199	0
Test	145	238	0	0
总体	181	610	571	0

(a) 差异

	训练集	验证集	Test	总体
训练集	778	308	243	778
验证集	308	349	150	349
Test	243	150	388	388
总体	778	349	388	959

（b）重叠

	训练集	验证集	Test	总体
训练集	0	154	184	0
验证集	18	0	64	0
Test	44	60	0	0
总体	61	197	201	0

（c）差异（法语）

	训练集	验证集	Test	总体
训练集	224	70	40	224
验证集	70	88	24	88
Test	40	24	84	84
总体	224	88	84	285

(d) 重叠（法语）

表11：各数据划分之间存在差异的表格数量（a、c）和重叠的表格数量（b、d）（两个任务均使用的子集）。差异按行减列计算。汇总结果见表3。

数据集	训练集	验证集	Test	All
# 对话	562	122	109	793
# 轮次	2147	472	439	3058
# 消息	6807	1492	1389	9688
# 查询	869	201	141	1211
# 每条消息的词元数	47.25	49.86	40.14	46.63
# 每段对话的轮次	3.82	3.87	4.03	3.86
# 每段对话的消息数	12.11	12.23	12.74	12.22
# 每段对话的查询数	1.55	1.65	1.31	1.53
# 表格	224	88	84	285
# 新增表格	0	18	44	61
# 每个表格的维度数	3.4	3.5	3.5	3.4
# 每个表格的成员数	205.8	249.9	164.1	190
# 每个表格的注释数	20.3	26.3	23.4	21

表12：法语划分的检索与生成任务统计数据。请参见表3，了解英语数据集划分。

数据集	训练集	验证集	Test	All
每条消息的最大令牌数	1243	857	912	1243
每段对话的最大轮数	28	18	21	28
每段对话的最大消息数	78	58	73	78
每段对话的最大问题数	15	11	11	15
每条消息的标准词元数	1	1	1	1
每段对话的标准轮次	2.58	2.36	2.62	2.56
每段对话的标准消息数	8.53	7.55	8.98	8.47
每段对话的标准问题数	1.27	1.24	1.18	1.25
每条消息的最大词元数	1374	1373	352	1374
每段对话的最大轮数	15	10	15	15
每段对话的最大消息数	59	38	47	59
每段对话的最大问题数	15	15	5	15
每条消息词元数的标准差	1	2	2	1
每段对话轮次数的标准差	2.21	2.08	2.1	2.17
每段对话消息数的标准差	7.02	6.74	6.61	6.91
每段对话问题数的标准差	1.84	2.18	0.77	1.79

表13：对话层级的更多统计数据（最大值和标准差；上方：英语，下方：法语），参见 表3。

发布年份	# 已归档	# 最新
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

表格14：加拿大统计局表格多年来的发布情况

合并后的言语行为	数量	原始言语行为
回答	50	回答 ，确认 ，否认
请求	31	请求、接受提议、拒绝提议
时间管理	28	拖延、暂停
告知	63	告知、同意、不同意、更正
信息寻求问题	25	命题问题、确认问题、集合问题、选择问题、直接问题
承诺	18	承诺、接受请求、拒绝请求
自动反馈	25	正向自动反馈，负向自动反馈
提议	11	提议
指示	18	指示
澄清问题	16	澄清问题
问候	28	初始问候, 回应问候
自我介绍	17	初始自我介绍 ，回应自我介绍
致谢	47	致谢
接受致谢	15	接受致谢

表 15：验证集对话中 100 个轮次里出现的合并后和原始言语行为。此表总结于 表4。

模型编码	模型编码	验证集 R@1	验证集 R@10	验证集 R@20	Test R@1	Test R@10	Test R@20
DPR	基础	37.8 ± 1.82	73.1 ± 1.02	80.8 ± 0.51	14.7 ± 1.13	45.0 ± 0.24	55.0 ± 1.67
	基础 + 脚注	40.0 ± 0.64	73.3 ± 0.85	81.3 ± 1.19	13.9 ± 0.37	44.4 ± 2.73	54.2 ± 2.65
	基础 + 成员	40.0 ± 0.44	74.5 ± 1.38	81.6 ± 0.55	15.7 ± 0.7	46.2 ± 0.44	56.3 ± 0.58
	成员	28.0 ± 0.47	62.1 ± 1.07	71.4 ± 1.06	10.7 ± 0.88	35.0 ± 0.57	46.3 ± 0.83
	标题	37.8 ± 0.21	70.9 ± 0.67	78.2 ± 1.18	13.9 ± 0.59	43.8 ± 1.22	53.4 ± 1.42
	标题 + 脚注	39.5 ± 0.69	73.2 ± 0.81	80.8 ± 0.47	13.3 ± 0.7	43.6 ± 0.46	53.6 ± 1.5
	标题 + 成员	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	表格	16.9 ± 4.26	45.9 ± 5.46	55.5 ± 3.13	3.9 ± 0.65	17.2 ± 2.83	24.6 ± 3.45
	标题	21.5 ± 0.33	54.7 ± 3.39	63.7 ± 1.38	4.7 ± 0.53	20.2 ± 0.85	28.7 ± 1.98
	标题 + 成员	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	表格	25.3 ± 3.46	60.0 ± 2.43	70.0 ± 2.89	5.3 ± 0.8	22.8 ± 2.32	32.0 ± 2.48
	标题	26.5 ± 1.85	64.4 ± 1.95	71.6 ± 2.32	6.3 ± 0.64	27.4 ± 1.84	36.1 ± 2.53
	标题 + 成员	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	基础	1.0	7.5	10.3	1.0	7.2	9.8
	基础 + 脚注	0.9	2.4	3.1	0.9	2.8	3.9
	基础 + 成员	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
	标题	2.0	10.3	13.8	2.5	10.9	14.0
	标题 + 脚注	0.8	2.3	3.1	0.7	2.9	4.0
	标题 + 成员	0.0	1.4	2.4	0.3	2.3	3.8

表16：英语数据划分的完整检索结果。报告的数值为截至 k的召回率百分比。DPR 和 TAPAS 运行了 3 次并取平均值（标准差列于 ±之后）。表8中的选定结果。

模型编码		验证集 R@1	验证集 R@10	验证集 R@20	Test R@1	Test R@10	Test R@20
DPR	基础	19.7 ± 1.52	57.4 ± 2.55	67.0 ± 5.01	12.5 ± 1.78	40.0 ± 1.08	48.5 ± 4.27
	基础 + 脚注	23.2 ± 3.8	55.4 ± 1.88	66.8 ± 1.04	13.9 ± 1.48	35.9 ± 0.82	49.9 ± 3.91
	基础 + 成员	22.7 ± 4.35	57.9 ± 2.45	67.7 ± 3.48	12.8 ± 2.13	38.1 ± 2.17	47.0 ± 1.78
	标题	23.7 ± 2.92	58.5 ± 2.24	69.8 ± 2.35	14.7 ± 2.28	35.7 ± 3.2	48.2 ± 2.84
	标题 + 脚注	23.2 ± 2.74	54.7 ± 3.59	64.7 ± 2.28	14.4 ± 1.48	36.6 ± 2.69	45.4 ± 2.13
	标题 + 成员	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	基础	0.0	0.0	0.0	0.0	0.0	0.0
	基础 + 脚注	0.0	0.0	0.0	0.0	0.0	0.0
	基础 + 成员	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
	标题	0.0	0.5	0.5	0.0	0.0	0.7
	标题 + 脚注	0.0	0.0	0.0	0.0	0.0	0.0
	标题 + 成员	0.0	0.5	1.0	0.0	0.7	0.7

表17：法语数据划分的完整检索结果。报告的值为 $k \in \{1, 10, 20\}$ 时的召回率百分比。DPR 运行了 3 次并取平均值（标准差列于 ± 之后）。

	获取表格的元信息
U:	你好，我想问一下新一轮人口普查的日期。
A:	你好，我叫Graham，请问有什么可以帮您?
U:	你好，Graham
U:	我目前正在查看：加拿大人口普查和全国住户调查
U:	我看到上一次人口普查是在2011年和2016年进行的。
U:	2020年接下来的月份里，我是否应该期待开展新一轮人口普查?
A:	不，下一次人口普查将在2021年5月举行。
A:	每 5 年一次
U:	好的，太好了，谢谢。
A:	不客气！祝你今天愉快。
	用户界面帮助
U:	你好，Graham——有没有办法查看年度总值，而不是每月数据?
U:	适用于 - https://www150.statcan.gc.ca/t1/tbl1/en/cv!recreate.action?...
A:	请稍等——我看看是否有提供年度频率的类似表格
U:	太好了，谢谢你
A:	很遗憾，没有。关于建筑投资，目前唯一有效的表格是我发给你的每月表格——还有一张非活动的季度表格《建筑施工投资》（https://www150.statcan.gc.ca/n1/en/surveys/5014）
A:	也可以下载 .CSV 格式的月度数据，在电子表格中将各月的数据相加，以获得年度总量
A:	或者，如果您愿意，也可以向我们的分析师提出定制请求（可能需要付费）。请向我们发送请求，清楚说明所需的变量、地理区域、时间范围和数据频率（即年度、每月或每季度）请发送至 infostats@canada.ca (mailto:infostats@canada.ca)，并注明案件编号：986005，以便我们查阅此次在线聊天的文字记录。您的请求必须包含您的姓名、地址、邮政编码和电话号码（请请勿在在线聊天中分享您的个人信息。）智能体将与您联系并提供报价，且如适用，请继续付款。您也可以拨打 1-800-263-1136 提交请求。
U:	没关系！非常感谢你 :)
U:	非常有帮助
U:	:)!
U:	祝你今天愉快！
A:	很高兴能帮上忙！你也一样。
	用户界面帮助
U:	您好，您知道在哪里可以下载第32周期综合社会调查的数据集吗?
A:	请稍候，我来帮您查找此信息。
U:	谢谢
U:	（我指的是这个： https://www23.statcan.gc.ca/imdb/p2SV.pl?... ）
A:	综合社会调查：照护与接受照护，公众使用微观数据文件（https://www150.statcan.gc.ca/...）
U:	抱歉，我找不到第32周期（2018年）的下载链接。
A:	你得点击“更多信息”，这样就会跳转到订单表。
U:	哦哦哦，我明白了
U:	谢谢你告诉我
U:	谢谢你的帮助！
A:	不客气，很高兴能帮上忙！
	域外请求
U:	我想知道如何获取一位家庭成员的尸检报告副本。医生办公室说他们没有收到尸检报告。
A:	加拿大统计局不签发出生、死亡或结婚证书。民事登记属于省级管辖权。因此，您必须联系所在省或地区的政府部门、登记处或生命统计办公室出生、死亡或婚姻事件发生的地区。要获取您所在省或地区的联系信息，请致电加拿大服务部，电话号码为 1-800-O-Canada（1-800-622-6232），或访问以下页面：生命统计数据联系信息（https://www.statcan.gc.ca/eng/health...）。
A:	另见：验尸官报告（https://spvm.qc.ca/en/...）

表18：用户意图示例，见A.1节。

U: 你好，我想获取有关加拿大GDP增长、不列颠哥伦比亚省住房开工量以及加拿大石油价格/桶的季度数据。
A: 你好，我叫 Kelly C.。请稍等片刻，我来搜索一下。

U: 谢谢！
A: 有关GDP增长率，请参阅以下链接：按行业划分的基本价格国内生产总值（GDP），每月，增长率
A: 我会继续搜索其余内容。请稍等片刻。
A: 你所说的“不列颠哥伦比亚省住房开工量”是什么意思?

U: 我需要按季度研究不列颠哥伦比亚省的所有新屋开工数据。
A: 定义“starts”。

U: 单位口径
U: 新屋开工数是指在特定月份内开始建设的新住宅建设项目数量。
A: 好的，请稍等片刻。
A: 我还在搜索。抱歉让您久等了。

U: 没关系。
A: 我会有关于新颁发的建筑许可证的月度数据。

U: 我来看看。
A: 建筑许可证，按建筑结构类型和工作类型划分（https://www150.statcan.gc.ca/t1/...）
A: 可以试着调整变量，找到你想要的内容。

U: 谢谢
A: 要查看/操作数据表中的变量： • 点击“添加/移除数据” • 在每个选项卡中至少选择一个变量，然后点击“应用”，即可查看自定义表格。要下载数据： •点击“下载” • 选择三种输出格式中的一种，然后点击超链接并保存表格。以下视频也可能对您有所帮助：如何使用数据表（https://www.statcan.gc.ca/eng/sc/video/howto）。
A: 我来查一下石油价格。请稍等。

U: 你们也有加拿大原油（西德克萨斯中质原油）的价格/桶数据吗?
U: 好的，谢谢！
A: 您要找的是原油的零售价格吗?
A: 如果是，我找到了一些较小地理区域的数据。
A: 按地理区域划分的汽油和燃料油月平均零售价格（https://www150.statcan.gc.ca/t1/...）

U: yes
A: 这些地理区域的数据够用吗?
A: 还是说您只想查找加拿大的数据?

用户：嗯
U: 我需要一些更适用于整个加拿大的数据。
A: 这个怎么样？食品和其他精选产品的月平均零售价格（https://www150.statcan.gc.ca/t1/...）
A: 还有什么我能为您效劳的吗?

U: 我想应该就这些了
U: I
U: 我会试着使用，谢谢！
A: 没问题。祝你一切顺利！

表19：取自对话数据集的完整示例（样例#42）。URL 已更新为链接至默认视图（而非筛选后的视图）。参见表1了解截断后的对话。

标题：枫树产品的产量与价值
日期范围：1924-01-01 至 2020-01-01 维度：地理、枫树产品 主题：农业 调查：枫树产品 频率：年度
地理： 标识符：1，父级：无 ，姓名：加拿大 标识符：2，父级：1 ，姓名：新斯科舍省 标识符：3，父级：1 ，姓名：新不伦瑞克省 标识符：4，父级：1，姓名：魁北克 标识符：5，父级：1 ，姓名：安大略省 枫树产品： 标识符：9，父级：无，姓名：以糖浆计的枫树产品，总计 标识符：10，父级：无，姓名：枫树产品总值 标识符：1，父级：无，姓名：枫糖产量 标识符：2，父级：无，名称：枫糖价值 标识符：3，父级：无，名称：枫糖浆产量 标识符：4，父级：无，名称：枫糖浆价值 标识符：5，父级：无，名称：枫糖太妃产量 标识符：6，父级：无，名称：枫糖太妃价值 标识符：7，父级：无，名称：枫糖黄油产量 标识符: 8, 父级: 无, 姓名: 枫糖黄油价值
标识符: 0, 注释: 换算系数：1加仑糖浆等于10.0磅枫糖。1加仑糖浆的重量为 13.24760 磅。一加仑糖浆相当于 10.4 磅太妃糖。枫糖太妃由魁北克省和新斯科舍省上报。 仅限新斯科舍省，并分别于 1965 年和 1983 年开始报告。枫糖太妃转化为糖浆的方式各不相同 具有该年份糖浆的密度。 标识符：0，注：商业生产和价值数据不包括库存。 标识符：2，注：由新斯科舍省园艺行业专家编制的报价。 标识符：3，注：从 1986 年开始，新不伦瑞克省和安大略省的调查数据按抽样次数加权。 在最近完成的农业普查年份的春季，对枫树进行的采集。 标识符：4，注：由魁北克统计研究所、Groupe AGÉCO 和 Fédération des 生成的报价。 魁北克的枫糖浆生产商。 标识符：5，注：自1986年起，新不伦瑞克省和安大略省的调查数据根据水龙头数量进行加权 在最近一次已完成的农业普查年份的春季从枫树上采集。 标识符：9，注：“这些产品由加拿大北美产品分类系统（NAPCS）表示 2017年类别 115136
枫糖浆和其他枫树产品。枫糖黄油、枫糖太妃和枫糖的数据经换算后 为枫糖浆当量，并归入此类别。” 标识符：10，注：“这些产品由加拿大北美产品分类系统（NAPCS）表示 2017年类别 115136 – 枫糖浆和其他枫树产品。为枫糖黄油、枫糖太妃和 枫糖会换算为等量枫糖浆，并计入此类别。” 标识符：0，备注：“这些产品由加拿大北美产品分类系统（NAPCS）代表。 2017年 类别 115136 – 枫糖浆和其他枫树产品。收集的枫糖黄油、枫糖太妃和 枫糖折算为等量枫糖浆，并计入此类别。”

表20：表格的完整元数据见第3.3节。添加单线以界定范围：标题、基本信息、项目名称和脚注。添加双线以界定DPR和TAPAS模型的截断限制（512个词元）。基本信息呈现在表5中。

PID	**18100004（数量 = 170）**
标题	消费者价格指数，每月，未进行季节性调整调整
日期	1914-01-01 至 2021-08-01
Dim.	地理、产品和产品组
主题	价格和价格指数
调查	消费者价格指数
频率	每月
PID	**11100222（数量 = 125）**
标题	家庭支出，加拿大、地区及省
日期	2010年1月1日至2019年1月1日
Dim.	地理、统计、家庭支出、汇总级类别
主题	收入、养老金、支出和财富
调查	家庭支出调查
频率	年度
PID	**17100005（数量 = 123）**
标题	7月1日的人口估计，按年龄及sex
日期	1971年1月1日至2021年1月1日
Dim.	地理、性别、年龄组
主题	人口与人口统计
调查	年度人口统计 估计 加拿大，省和地区
频率	年度
PID	**33100214（数量 = 117）**
标题	加拿大企业数量（含雇员）2019年6月
日期	2019年1月1日至2019年1月1日
Dim.	地理、就业规模、北美行业分类系统（NAICS）
主题	企业绩效与所有权
调查	企业名录
频率	半年
PID	**35100177（数量 = 104）**
标题	基于事件的犯罪统计，按详细违规行为划分，加拿大、省、地区及 Census 苏斯大都市区
日期	1998-01-01 至 2020-01-01
Dim.	地理、违法行为、统计数据
主题	犯罪与司法
调查	统一犯罪报告调查
频率	年度
PID	**18100005（数量 = 91）**
标题	消费者价格指数、年度平均值，非季节性调整
日期	1914-01-01 至 2020-01-01
Dim.	地理、产品和产品组
主题	价格和价格指数
调查	消费者价格指数
频率	年度

表 21：检索任务中所有数据划分里最常检索到的表格。每个表格的基本信息均已提供。分析见第 A.3.1 节。

U：嗨，我想问一下，你们有没有电子游戏销售额或高中辍学率的统计数据？信息搜寻问题；请求；A：你好，我叫Sylvain，请问有什么可以帮您？提议；U：你好，我想查询高中辍学率和青少年抑郁率，但似乎找不到超过一年前的数据。请求；A：高中辍学数据由省教育部汇编。如需获取这些信息，请联系你所在省份的教育部。告知；承诺；A：我们的网站上有情绪障碍（抑郁症）数据，见数据表 13-10-0096-18 (https://www150.statcan.gc.ca/...)。回答；A：要查看/调整数据表中的变量： • 点击“添加/移除数据” •在每个选项卡中至少选择一个变量，然后点击“应用”，即可查看自定义表格。要下载数据： • 点击“下载” • 选择三种输出格式之一，然后点击超链接并保存表格。以下视频也可能对你有所帮助：如何使用数据表 (https://www.statcan.gc.ca/eng/sc/video/...)。告知；请求；U：团队好，你们能帮我查询特许银行的公共存款数据吗？信息搜寻问题；请求；A：请稍等。时间管理；U：当然。自动反馈；A：你可以在我们的网站上查到这些信息，见数据表 10-10-0116-01 (https://www150.statcan.gc.ca/t1/...)。回答；A：要查看/调整数据表中的变量： • 点击“添加/移除数据” •在每个选项卡中至少选择一个变量，然后点击“应用”，即可查看自定义表格。要下载数据： •点击“下载” •选择三种输出格式之一，然后点击超链接并保存表格。以下视频也可能对你有所帮助：如何使用数据表 (https://www.statcan.gc.ca/eng/sc/video/...)。告知；指示；A：请注意，这不是加拿大统计局的数据表。该信息由加拿大银行提供。告知；U：非常感谢！那么，我该如何查找公共存款？或者说，应该看哪一项才能找到公共存款？澄清问题；A：在“特许银行存款”类别下。如需了解更多信息，请访问加拿大银行网站并使用其在线请求表 (https://www.bankofcanada.ca/...)。回答；告知；

表22：对话分析示例，见 节-第 3.2 节，总结见表 15。 The言语行为位于蓝色每条消息的末尾。其余注释见补充材料。例子。示例见表4。

U:	**感谢您提供信息，不过我已经做过了。您之前的网站有一个"manipulate"功能，可以计算百分比变化。我现在看不到那个按钮了，想问问您是否还有其他已准备好的表格，因为那个功能已经没有了。**
A:	**您好，我叫 Jimmy N.，请问有什么可以帮您?**
T5:	您好，我叫 Jimmy N.，请问有什么可以帮您?
T5+D:	您好，我叫 Jimmy N.，请问有什么可以帮您?
U:	**我刚才在和 Sylvain 交谈，您能看到记录吗?**
A:	**好的，请稍候，我来核实一下。**
T5:	好的，请稍候，我来核实一下。
T5+D:	好的，请稍候，我来核实一下。
A:	**确认一下，您要找的是所有项目吗?**
T5:	您可以通过以下超链接获取此信息：表：18-10-0004-13 消费者价格指数按产品组、每月、百分比变化、未经季节性调整、加拿大、省、怀特霍斯，耶洛奈夫和伊魁特（）
T5+D:	您可以通过以下超链接获取此信息：表：18-10-0004-01 消费者价格指数，按产品组、每月，[...] 以及伊魁特（）

表 23：对话 #4890。每条消息之间以水平线分隔。原始对话（**粗体**部分）仅包含用户（U）和智能体（A）发送的消息。T5 以及结合前 5 个 DPR 检索到的标题的 T5（T5+D）生成的回复，会预测相应智能体的消息（以**粗体**显示）。

	[...]
U:	**理想情况下，我想要这些项目的百分比变化**
R_1:	消费者价格指数，每月，未经季节性调整
R_2:	每月调整后的价格指数，暂定
R_3:	消费者价格指数，每月，季节性调整
R_4:	工业品价格指数按主要产品组划分的历史（实时）发布，每月，已停用
R_5:	加拿大银行消费者价格指数（*CPI*）统计数据，替代指标，未经调整和季节性调整加拿大
A:	**您可以通过以下超链接获取此信息：表：18-10-0004-13 消费者价格指数按产品组划分，每月，百分比变化，未经季节性调整，加拿大各省，怀特霍斯、耶洛奈夫和伊魁特（https://www150.statcan.gc.ca/t1/tbl1/...）**
T5:	您可以通过以下超链接获取此信息：表：18-10-0004-13 消费者价格指数 [...] 和伊魁特（）
T5+D:	您可以通过以下超链接获取此信息：表：18-10-0004-13 消费者价格指数 [...] 和伊魁特（）

表24：对话 #4890 的第二部分，紧接在表 23之后。 R_i 表示由 iDPR 检索出的第5+个表格，该表格仅由 T5+D。

U：这真不错！A：你好，我叫Martine，请问有什么可以帮您？U：我有个问题——如何查询2019年联邦政府通过就业保险/老年保障等社会项目向加拿大人支出的金额？A：我来查查看.....U：谢谢R_i： 联邦科学技术外部支出，科学类型、执行部门、支付类型和地理R_i： 历史：对其他级别政府的当期转移，*1968*年国民账户体系（*SNA*），年度，*1926—1986*年R_i： 私人非营利组织外包的研究与开发支出R_i： 资本支出，资金来源，按地理划分R_i： 历史：政府对个人的转移支付，*1968*年国民账户体系（*SNA*），年度，*1926—1986*年**A**： **收入、支出和预算余额——一般政府（× 1,000,000）（https://www150.statcan.gc.ca/t1/...3610047701**）T5：数据表（https://www150.statcan.gc.ca/n1/en/type/data?text=联邦+债务）T5+D：联邦科学技术外部支出，科学类型、执行部门、支付类型和地理（）

表24：对话#8960。

用户：我如何查找与基洛纳人口估计、人口统计、人口预测估计以及按年龄和性别划分的人口相关的信息？助手：你好，今天我能为你提供什么帮助？ 用户：我如何查找与基洛纳人口估计、人口统计、人口预测估计以及按年龄和性别划分的人口相关的信息？ 助手：让我看看能找到哪些主题相关的信息。关于预测，我们的表格只提供省级数据，无法细分到更低层级。 <额外_标识符_0>(<额外_标识符_0>)R_1： 加拿大、省和地区*0*至*29*岁人口估计和预测，按年龄组划分R_2： 截至*7*月*1*日的预测人口，按预测情景、年龄和性别划分R_3： *7*月*1*日人口估计，按年龄和性别划分R_4： 预测人口增长的组成部分，按预测情景划分R_5： 季度人口估计**助手**： **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

表26：对话 #21533。

U：你好，你知道加拿大大约有多少家手工蒸馏酒生产商在营业吗？A：你好，我叫 Sylvain，请问有什么可以帮您？U：我想了解一下您是否掌握加拿大手工蒸馏酒生产商的相关信息。具体来说，加拿大有多少家手工蒸馏酒生产商在营业？A：行业根据北美行业分类系统（NAICS）进行分类。要查找某个行业的北美行业分类代码，请参阅此页面（http://www23.statcan.gc.ca/imdb/...1181553）。您可以先从20个行业门类中选择一个，然后逐步查找，直到获得六位数的代码。 [...] 您可以在以下数据表中按北美行业分类代码查看企业经营地点数量：表 33-10-0105-01《加拿大企业数量：有雇员，2018年12月》（https://www150.statcan.gc.ca/t1/...3310010501） [...] 我还建议您查阅加拿大中小企业的财务数据；这些组织的年度收入在3万至2000万加元之间。此信息按公司所属行业、省、地区和法律地位（注册成立或未注册成立）提供。这些数据无法在我们的网站上获取，但可以在加拿大创新、科学与经济发展部的网站上，通过其财务绩效数据（http://www.ic.gc.ca/...)[...]

表27：对话 #1628

U：您好。我想知道加拿大有多少家小型、中型和大型公司。谢谢。A：您好，我叫 Olivier C。请稍候，我来为您查找所需信息。R_1： *2019*年*12*月加拿大企业数量（无雇员）R_2： *2019*年*12*月加拿大企业数量（有雇员）R_3： *2019*年*6*月加拿大企业数量（无雇员）R_4： *2019*年*6*月加拿大企业数量（有雇员）R_5： *2017*年*12*月加拿大企业数量（无雇员）**A**： **感谢您的等候。点击以下超链接，即可查看按企业规模划分的企业数量信息：2019年12月加拿大企业数量（有雇员）（https://www150.statcan.gc.ca/t1/...pid=3310022201）2019年12月加拿大企业数量（无雇员）（https://www150.statcan.gc.ca/t1/...pid=3310022301）**T5：感谢您的等候。点击以下超链接，即可查看您需要的信息：2019年6月加拿大企业数量（有雇员）（ ）T5+D：感谢您的等候。行业按照北美行业分类系统（NAICS）进行分类。要查找某个行业的北美行业分类代码，请参阅此页面（http://www23.statcan.gc.ca/imdb/...1181553）。您可以从选择20个行业门类之一开始搜索，直至获得六位数的北美行业分类代码。以下数据表列出了按北美行业分类代码划分的企业经营地点数量：表 33-10-0214-01 2019年6月加拿大企业数量（有雇员）（ ）表 33-10-0215-01 2019年6月加拿大企业数量（无雇员）（ ）。我还建议您查阅加拿大中小企业的财务数据；这些组织的年度收入范围为30,000至2,000万美元。该信息按行业、省、地区以及公司的法律地位（注册成立或未注册成立）提供。我们的网站上没有这些数据，但可在加拿大创新、科学与经济发展部的网站上免费获取，具体请查阅其财务绩效数据（http://www.ic

表 28：对话 #8500。

U：您是否有关于制造业企业劳动力的数据，以及这些劳动力在直接劳动和间接劳动之间如何分配？A：您好，我叫 Jimmy N.。能否请您说明一下，您所说的直接劳动和间接劳动具体指什么？U：用于生产产品的直接劳动，行政管理、监督等间接劳动。A：您请求的数据不是标准产品。不过，可能可以通过定制订单获取（可能需要付费）。 = ... = 您想提出此请求吗？A：作为标准产品，我们有按行业划分的就业数据表格。U：我只是想查看标准产品。您能告诉我在哪里可以找到，或者它叫什么吗？ [：] 按行业划分的劳动力特征，每月，季节性调整，最近5个月R_1：按行业划分的就业，每月，季节性调整和未调整，以及趋势周期，最近*5*个月R_2：按行业划分的就业，年度R_3：按行业划分的就业，三个月移动平均，未进行季节性调整，省和经济区域，非活跃R_4：按行业划分的通常工作时数，年度R_5A：**您可以通过以下超链接获取这些信息：表格：14-10-0355-02 按行业划分的就业，每月，季节性调整（千人）表格：14-10-0023-01 按行业划分的劳动力特征，年度（千人）2015年、2016年人口普查中，全年全职工作并申报就业收入的就业劳动力（**https://www150.statcan.gc.ca/t1/...1410035502）**T5：表格：14-10-0296-01 按行业划分的劳动力特征，年度（千人）（）T**5+D：请稍候，我来查找相关信息。

表29：对话编号19568。

U：我正在浏览网页 https://www150.statcan.gc.ca/n1/daily-quotidien/201224/dq201224beng.htm，请问在哪里可以找到往年份的同一数据集，例如2019年、2018年、2017年等？ A：您好，请问您指的是哪个表格：https://www150.statcan.gc.ca/n1/daily-quotidien/201224/dq201224b-cansim-eng.htm U：2020年1月至10月的暂定死亡人数和超额死亡率——按周列出的调整后的死亡人数、预期死亡人数和超额死亡率估计值。 A：这就是该报告，但其中有5个不同的表格：https://www150.statcan.gc.ca/n1/daily-quotidien/201224/dq201224b-cansim-eng.htm 请让我为您核实一下。请稍候。要查看/操作数据表中的可用变量： • 点击“添加/移除数据” • [.....] 在每个选项卡中至少选择一个变量，然后点击“应用”，即可查看您自定义的数据表。数据可用于2016年至2020年。 U：参考期截至2020年1月，我似乎无法获取往年的数据。 A：点击日程图标，然后选择所需年份。（2016年—2020年） U：我已经这样做了，但所有年份都显示为灰色，意味着它们不会加载。 A：请让我核实一下。请稍候。 U：screenshot.lzscR_1：按年龄组和性别分列的每周死亡人数R_2：每周死亡人数的先前发布版本和修订R_3：按周、年龄组和性别分列的调整后的死亡人数、预期死亡人数和超额死亡率估计值R_4：按周列出的死亡人数，已停用R_5：按性别和详细年龄组分列的每月劳动力特征，未进行季节性调整。 A：您是否已经查看过替代表13-10-0785-**01** [.....]https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid**=1310081001**- 在此表中，您可以追溯到2016年。 T5：您可以按照以下步骤操作该表格： • 点击“添加/移除数据” • 点击复选框旁边的“ + ”图标，以展开变量列表（如适用） • 在每个选项卡中至少选择一个变量，然后点击“应用”，即可查看您自定义的数据表。要下载数据： • 点击“下载” •选择三种输出格式中的一种，然后点击超链接并保存表格。以下视频可能也会对您有所帮助：如何使用数据表（https://www.statcan.gc.ca/eng/sc/howto）。T5+D：请查阅以下数据表：按年龄组和性别分列的每周死亡人数（ ）

表30：对话#15376。

U:	你好
A:	您好，我的姓名是 Jimmy N.，请问有什么可以帮您?
U:	[.....] 我们看到了《温哥华太阳报》上的这篇文章，其中提到了加拿大统计局有关企业倒闭的数据。 [.....] 是否有办法收集乡镇的本地化数据，或者至少收集菲沙河谷的数据?
A:	很遗憾，这些信息仅按全国和省/地区层级提供：而对于某些人口普查大都市区
R_i:	加拿大、省和地区的人口普查企业开业与停业试验性估计 大都市地区，季节性调整
R_i:	企业进入和退出的季度估计
R_i:	在至少有一名员工被裁员的企业中，因.....被裁员并重新雇用的劳动力百分比 *COVID-19*，按企业特征划分
R_i:	加拿大企业数量，含雇员、人口普查大都市区和人口普查分区，*12*月 *2020*
R_i:	按行业划分的选定的*12*个人口普查大都市区制造业销售额
A:	**表格： 33-10-0270-01 - 实验性的 估计 for 企业 企业开业 and 企业关闭 for 加拿大， 省和地区， 人口普查 大都市 区域， 季节性而言 已调整** **（https://www150.statcan.gc.ca/t1/...3310027001）**
T5:	不过，也可以通过定制订单获取（可能需要付费）。如需获取免费费用估算 如需此项服务，请向我们发送请求，并明确说明变量、地理区域和时间范围 和所需的数据频率，并确保提供完整的联系信息（机构名称- 组织、地址、邮政编码和电话号码）请发送至以下电子邮箱地址：infostats@canada.ca (mailto:infostats@canada.ca) 或拨打电话 1-800-263-1136。我们的智能体周一至周五为您服务。 （节假日除外），东部时间上午 7:30 至晚上 7:30。智能体会与您联系以核实您的 信息、提供报价，并在适用时进行付款。
T5+D:	不过，您可以通过定制订单获取该信息（可能需要付费）。如需免费 [...] 智能体将 与您联系以核实信息、提供报价，并在适用的情况下继续办理付款。 请勿通过在线聊天分享任何个人信息。您想提出此请求吗?

表31：对话 #11869。

PID	标题	训练集	验证集	Test	总计
13100810	按周划分的选定分组死亡原因	0	0	6	6
33100270	加拿大企业开业与停业的试验性估计， 省和地区、人口普查大都市区，季节性调整	0	1	3	4
13100763	加拿大1至17岁儿童和青少年的健康特征 2019年儿童与青少年健康调查	0	0	4	4
35100060	凶杀案受害者和被控犯有凶杀罪者人数，按原住民身份划分 身份、年龄组和性别	0	1	2	3
32100042	使用.....的籽粒玉米和大豆估算面积、产量和产量，采用 转基因种子、魁北克和安大略省，采用公制和英制单位。	0	1	2	3
14100084	移民按性别和年龄组划分的劳动力特征，三个月 未按季节性调整的移动平均值	0	0	3	3
13100143	死亡人数，按死因分类，第五章：精神和行为障碍（F00 至 F99）	0	0	3	3

表32：未出现在训练集、但至少在验证集或测试集中出现3次的表格列表。

语言	数据划分	数据增强	METEOR	ROUGE-L	BERTScore	MoverScore	标题准确率
英语	Test	None	23.35	30.65	86.04	59.82	6.96
		排名第一的标题	24.07	30.76	86.11	60.23	7.99
		Top-5 标题	24.41	30.88	86.17	60.31	10.82
	验证集	None	31.95	43.31	88.31	65.50	30.39
		排名第一的标题	33.11	44.22	88.60	66.14	33.66
		Top-5 标题	33.38	44.20	88.55	65.94	38.89
法语	Test	None	5.98	7.21	61.74	51.88	0.00
		排名第一的标题	9.71	11.14	65.86	54.81	0.00
		Top-5 标题	8.96	10.75	64.76	54.38	0.00
	验证集	None	6.84	8.85	60.97	52.21	0.00
		排名第一的标题	8.09	9.03	64.22	53.62	0.00
		Top-5 标题	7.37	8.75	62.55	52.97	0.00

表 33：完整响应生成结果。所选结果见 表10。