A Appendix

A.1 Tabular Candidates Format for BM25 Retriever

In order to represent a structured table as passages, we flattened each table into passages by concatenating cell values along each row. If the flattened table exceeds 100 words, we split it into a separate passage, respecting row boundaries. The column headers are concatenated to each tabular passage. For example,

Country	Film title	Language	Director
Argentina	The Island	Spanish	Alejandro
Austria	Tales...	German	Maximilian

is flattened to be

> *[header] Country ; Film title ; Language ; Director [row] Argentina ; The Island ; Spanish ; Alejandro [row] Austria ; Tales from the Vienna Woods ; German ; Maximilian*

A.2 EM of the Unified Models

In this section, we train two unified models DUPERA and DUREPA– on the combined training data from all datasets, and then test them on each individual dataset. Similar to the individual-model setting in Section 4.3, we observe that having the ability to generate structural queries is always beneficial even for extractive questions like SQuAD and NQ. And for WikiSQL-type questions, the gain of SQL generation is significant.

A.3 Recalls on WikiSQL-both and NQ datasets

We present the recall@k for k = 1, 5, 10, 25, 50 and 100 on the OpenWikiSQL and OpenNQ datasets in Table 10 and 11.

A.4 More Examples that Can Only be Answered by DUREPA

We provide more examples of the SQuWiki questions that can be answered correctly by DUREPA but incorrectly by DUREPA–.

A.5 Zero-shot Performance on OpenNQ

In Table 12, we present some predictions on NQ questions under the zero-shot setting. The models used here are trained on MixSQuWiki dataset and tested directly on NQ questions without finetuning.

We observe that for all these questions, the generated SQLs are reasonable and are coherent to what the original questions are asking about. For the first question, even though the groundtruth set is not exactly equal to the executed results, they largely overlap. For the second question, the generated SQL query is indeed searching for colleges that Johnny Manziel has played for. The generated SQL can also be successfully executed and returns a long list of answers, which indeed contains the correct answer "Texas A&M". In this case, the error comes from an erroneous table. This is also the case for the third example. For the fourth to eighth examples, our model generates SQLs that execute to correct answers. Some of them requires generating multiple conditions connected by "AND". This demonstrates that the DUREPA model indeed learns useful generalizable semantics of SQL queries.

A.6 Upper bound analysis on OTT-QA

In this section, we provide an upper bound analysis for the OTT-QA questions. The main purpose of this experiment is to investigate the main bottleneck of DUREPA model on this multi-hop datasets.

Under the oracle setting, if the groundtruth supporting table is successfully retrieved by the retriever, we then add its hyperlinked supporting passages. Similarly, we add the linked supporting tables for the groundtruth passages if they are successfully retrieved. This is in order to mimic the functionality of the fusion-retriever, which is trained using the hyberlinks in the OTT-QA paper.

The results are presented in Table 13. We can see that if we have an "oracle" retriever, the end-to-end EM can be improved to 32.2, which is more than doubled compared to 15.8 EM under the normal setting. Therefore, the retriever is indeed the bottleneck of our method on multi-hop QA. Actually, this also holds true for the baseline models. The FR+CR model also significantly improves upon IR+CR model by only replacing the iterative retriever with the fusion-retriever, which is trained to link each table with its hyperlinked paragraphs.

A.7 Some Ambiguous WikiSQL Questions

In Table 14, we demonstrate that some errors our models make on WikiSQL questions are actually due to the ambiguity. These questions often have different possible answers depending on the contexts. Nevertheless, the SQL queries generated by DUREPA are reasonable and reflect what the original questions are asking for.

Model	Evidence Corpus Type	OpenSQuAD	OpenNQ	OTT-QA	OpenWikiSQL
FiD(T5-base)	Text-only	53.4	48.2	-	-
FiD(T5-large)	Text-only	**56.7**	51.4	-	-
IR+CR	Text+Table w/o SQL	-	-	14.4	-
FR+CR	Text+Table w/o SQL	-	-	**28.1**	-
Unified Model	Text+NQ Table w/o SQL	-	**54.6***	-	-
Our unified model					
DUREPA–	Text-only	56.2	41.6	14.5	10.1
DUREPA–	Table-only w/o SQL	1.6	12.9	4.0	28.3
DUREPA	Table-only with SQL	1.7	13.2	4.5	40.3
DUREPA–	Text+Table w/o SQL	56.2	42.3	15.1	28.5
DUREPA	Text+Table w/ SQL	**56.7**	**45.1**	**15.8**	**41.0**

Table 8: Comparison of the unified model to the state-of-the-art on open-domain QA datasets. The numbers reported are in EM metric. FiD(T5-base & T5-large) is reported from (Izacard and Grave, 2020), IR+C (Iterative Retrieval+Cross-block Reader) and FR+CR (Fusion Retrieval+Cross-block Reader) are from (Chen et al., 2020a), Unified Model is from (Oguz et al., 2020).

Index	AES **textual**	Reranked **textual**	AES **tabular**	Reranked **tabular**	Reranked **hybrid**
Top-1	13.10	18.69	51.70	50.24	50.28
Top-5	20.08	25.61	66.27	68.15	68.15
Top-10	22.54	28.84	70.93	74.09	74.10
Top-25	25.24	32.34	75.53	80.91	80.89
Top-50	29.66	35.39	80.54	84.78	84.63
Top-100	33.20	38.14	84.14	87.18	87.13
MAP	13.15	18.48	47.63	47.92	44.93
MRR	16.56	22.03	58.49	58.34	58.38

Table 9: Recalls on top-n textual, tabular or the hybrid candidates for OpenWikiSQL questions.

Index	BM25 **textual**	Reranker **textual**	BM25 **tabular**	Reranker **tabular**	Reranker **both**
Top-1	23.68	51.63	9.31	23.52	52.91
Top-5	47.12	71.00	20.42	34.82	72.33
Top-10	56.76	75.12	26.18	38.48	76.37
Top-25	67.73	78.42	34.43	42.24	80.03
Top-50	74.76	79.64	40.66	44.54	81.33
Top-100	80.25	80.25	45.54	45.54	82.22
MAP	19.32	43.87	9.98	22.15	42.55
MRR	34.54	59.95	14.79	28.81	61.31

Table 10: Recalls on top-n textual, tabular or the hybrid candidates for OpenNQ questions.

Question:	What was the control for the year with a Conservative Party result of 10 (+5)?
Groundtruth:	['labour hold']
Top-1 generation by sw:	`sql: SELECT Control FROM table_2-16041438-1 WHERE Conservative Party = "10 (+5)"`
Execution result:	['labour hold']
Top-1 generation by sw–	answer: no overall control
Question:	What was the date that the decision was Niittymaki, the attendance larger than 19,207, and the record 28–17–5?
Groundtruth:	['january 31']
Top-1 generation by sw:	`sql: SELECT Date FROM table_2-11902580-6 WHERE Decision = "niittymaki" AND Attendance >"19,207" AND Record = "28-17-5"`
Execution result:	['january 31']
Top-1 generation by sw–	answer: february 31

Table 11: Examples of the SQuWiki questions that are answered correctly by model DUREPA but incorrectly by model DUREPA–.

Question:	cast of law & order special victim unit
Groundtruth:	['kelli giddish', 'richard belzer', 'stephanie march', 'diane neal', 'ice-t', 'danny pino', 'dann florek', 'tamara tunie', 'michaela mcmanus', 'mariska hargitay', 'adam beach', 'b. d. wong', 'christopher meloni', 'raúl esparza', 'michelle hurd', 'peter scanavino']
Top-1 generation:	`sql: SELECT Cast FROM table_22266670-7 WHERE Program = "law & order: special victims unit"`
Execution result:	['chris orbach', 'donnetta lavinia grays', 'nicholas gonzalez', 'robert john burke', 'kathryn erbe', 'jerry orbach', 'joel de la fuente', 'john schuck', 'william h. burns', 'dean winters', 'connie nielsen', 'adam beach', 'adam baldwin', 'jesse l. martin']
Question:	who did johnny manziel play college football for
Groundtruth:	answer: texas a&m
Top-1 generation:	`sql: SELECT College(s) played for FROM table_3401335-11 WHERE Player = "johnny manziel"`
Execution result:	['montana', 'ucla', 'texas a&m', 'california', 'baylor', 'purdue', 'clemson', 'michigan', 'southern california', 'princeton', 'california lutheran', 'cornell', 'oregon', 'columbia', 'yale , rutgers', 'trinity college', 'tennessee', 'texas', 'iowa', 'kentucky state', 'army', 'ohio state', 'mississippi', 'lsu', 'penn state', 'boston college', 'yale']
Question:	who is the original singer of you re going to love me
Groundtruth:	answer: jennifer holliday
Top-1 generation:	`sql: SELECT Original artist FROM table_30996994-1 WHERE Song (original artist) = "you re going to love me"`
Execution result:	['whitney houston', 'diana ross', 'dreamgirls', 'blu cantrell', 'the beatles', 'faith hill', 'jennifer holliday', 'monica']
Question:	who wrote the song what child is this
Groundtruth:	answer: william chatterton dix
Top-1 generation:	`sql: SELECT Author FROM table_36169771-1 WHERE Song = "what child is this?"`
Execution result:	['william chatterton dix']
Question:	who sings the pokemon theme song season 1
Groundtruth:	answer: jason paige
Top-1 generation:	`sql: SELECT Vocalist FROM table_2144389-13 WHERE Title = "pokémon theme" AND Episodes used 1 = "pokémon theme"`
Execution result:	['jason paige']
Question:	when did david akers kick the 63 yard field goal
Groundtruth:	answer: september 9, 2012
Top-1 generation:	`sql: SELECT Date FROM table_8378967-1 WHERE Distance = "63 yards" AND Kicker = "david akers"`
Execution result:	['september 9, 2012']
Question:	what album is sacrifice by elton john on
Groundtruth:	answer: sleeping with the past.
Top-1 generation:	`sql: SELECT Album FROM table_4105885-1 WHERE Artist = "elton john" AND Song = "sacrifice"`
Execution result:	['sleeping with the past']
Question:	who played raquel in only fools and horses
Groundtruth:	answer: tessa peake-jones
Top-1 generation:	`sql: SELECT Actor FROM table_6994109-1 WHERE Role = "raquel" AND Film/Show = "only fools and horses"`
Execution result:	['tessa peake-jones']

Table 12: Some predictions on NQ questions under zero-shot learning setting.

Model	**Evidence Corpus Type**	**EM (normal setting)**	**EM (oracle-retriever setting)**
IR + CR	Text+Table w/o SQL	14.4	–
FR + CR	Text+Table w/o SQL	28.1	–
DUREPA–	Text-only	14.5	31.9
DUREPA–	Table-only w/o SQL	4.1	3.6
DUREPA	Table-only with SQL	4.7	4.8
DUREPA–	Text+Table w/o SQL	15.0	28.5
DUREPA	Text+Table with SQL	15.8	32.2

Table 13: Upper bound results on OTT-QA dataset. The results show that the main bottleneck of DUREPA methods on OTT-QA dataset is the retriever.

Question:	What is the record when the opponent is washington redskins?
Groundtruth:	[”0–3”]
Groundtruth SQL:	`SELECT Record FROM table_1-18847692-2 WHERE Opponent = Ẅashington Redskins¨`
***Top-1 generation by* DUREPA:**	`SELECT Record FROM table_2-15581223-3 WHERE Opponent = ẅashington redskins¨`
Execution result:	[”1–0”]
Ambiguity:	There are many records when the opponent is Washington Redskins.
Question:	With the nickname the swans, what is the home ground?
Groundtruth:	[”lilac hill park”]
Groundtruth SQL:	`SELECT Home ground(s) FROM table_1-18752986-1 WHERE Nickname = S̈wans¨`
***Top-1 generation by* DUREPA:**	`SELECT Home ground(s) FROM table_2-17982112-1 WHERE Nickname = s̈wans¨`
Execution result:	[””]
Ambiguity:	There are more than one home grounds with nickname the swans.
Question:	Who had the fastest lap in the Belgian Grand Prix
Groundtruth:	[”rubens barrichello”]
Groundtruth SQL:	`SELECT Fastest Lap FROM table_1-1132600-3 WHERE Grand Prix = B̈elgian Grand Prix¨`
***Top-1 generation by* DUREPA:**	`SELECT Fastest Lap FROM table_1-1140077-2 WHERE Race = B̈elgian Grand Prix¨`
Execution result:	[”carlos reutemann”]
Ambiguity:	The question does not specify which year. Many answers are possible.

Table 14: Examples of ambiguous questions in OpenWikiSQL.