A Appendix

Role of Model Reasoning. Table 1 presents the complete experimental results of the role of model reasoning. Across BM25, scaling the rewrite model and enabling "think" generally increase nDCG@10; Qwen3-8B(think) gives the largest gains among Qwen variants. Under SBERT, gains from "think" are smaller and sometimes negative for small models, suggesting dense encoders are less tolerant of verbose/latent-logic rewrites. In both retrievers, **ReDI** surpasses all size/think settings, indicating that structured decomposition + interpretation contributes more than raw model size or implicit chain-of-thought.

Table 1: nDCG@10 on BRIGHT with Qwen3 across different model sizes and reasoning models.

Param	Think Mode	Avg.All	Avg.SE	*StackExchange*							*Coding*		*Theorem-based*		
				Bio.	Earth.	Econ.	Psy.	Rob.	Stack.	Sus.	Leet.	Pony	AoPS	TheoQ.	TheoT.
Using BM25 Retriever															
-	-	14.5	17.2	18.9	27.2	14.9	12.5	13.6	18.4	15.0	**24.4**	**7.9**	**6.2**	10.4	4.9
0.6B	Disabled	15.3	19.7	22.0	31.9	17.0	23.9	13.9	16.6	12.7	14.4	3.4	1.1	15.5	11.3
0.6B	Enabled	15.7	19.3	26.7	26.2	15.3	21.6	13.5	14.6	17.0	20.6	6.3	2.3	15.8	8.0
4B	Disabled	19.0	24.9	28.8	39.3	10.8	27.3	19.6	21.0	16.7	19.4	3.7	2.4	19.3	19.5
4B	Enabled	20.1	25.7	32.6	40.5	18.4	28.3	17.3	22.5	**20.3**	18.8	3.8	3.0	17.6	18.0
8B	Disabled	20.7	26.3	32.8	43.3	18.2	**30.3**	19.6	22.4	17.3	19.4	3.7	2.4	19.3	19.5
8B	Enabled	**22.8**	**29.0**	**40.1**	**46.0**	**20.9**	29.6	**21.5**	**25.0**	20.2	16.9	6.4	3.1	**22.5**	**21.7**
Using SBERT Retriever															
-	-	14.9	15.6	15.1	20.4	**16.6**	**22.7**	8.2	11.0	15.3	**26.4**	7.0	**5.3**	20.0	10.8
0.6B	Disabled	10.7	12.6	12.4	19.0	11.8	19.1	**10.0**	6.8	9.1	12.3	0.7	3.5	14.3	9.1
0.6B	Enabled	10.3	12.2	13.5	18.2	13.9	3.3	9.1	9.1	18.3	9.5	4.0	3.6	15.7	5.1
4B	Disabled	13.7	14.4	**17.9**	18.3	13.5	20.4	7.6	10.9	12.4	15.8	3.5	5.4	19.1	19.7
4B	Enabled	14.3	15.0	15.1	20.8	12.0	**22.7**	5.4	12.3	16.5	18.0	6.1	1.8	23.0	17.6
8B	Disabled	15.3	16.6	17.5	**24.4**	14.4	21.9	6.5	10.8	**20.7**	18.9	10.8	1.3	22.7	13.9
8B	Enabled	**16.8**	**17.2**	15.7	23.0	15.7	22.1	9.3	**14.9**	19.8	18.2	**11.2**	3.4	**26.9**	**21.2**

Flexible vs. Fixed Decomposition Granularity. Table 2 presents the complete experimental results of the influence of the decomposition unit. With BM25, performance improves from 3⟶9 units and plateaus after 9–11, while very large unit counts show mild regressions—evidence of coverage–noise trade-off. With SBERT, trends are flatter, and peaks occur around 9–11, reflecting dense encoders' preference for fewer, stronger facets. In both cases, **ReDI** (learned, intent-adaptive granularity) exceeds any fixed setting, supporting adaptive decomposition over manual heuristics.

Hyperparameter Sensitivity. Table 3 presents the complete experimental results of the parameter influence in sparse retrieval. BM25 attains its best average around $k_3 \approx 0.4$, with clear degradation as k_3 increases (i.e., approaching linear qtf). Small k_3 strengthens repeated key terms in short units (beneficial for decomposition), whereas large k_3 reduces this advantage and hurts long-document domains. The curve is smooth with a broad optimum $k_3 \in [0, 0.8]$, indicating stable tuning.

Table 4 presents the complete experimental results of the influence of interpretation weight in dense retrieval. SBERT peaks near description weight $\lambda = 0.5$, confirming that balancing the original sub-query and its interpretation yields the best alignment with document embeddings. Over-weighting interpretations $\lambda \geq 0.8$ or the sub-query $\lambda \leq 0.2$ both reduce effectiveness, especially on theory-heavy sets.

Fine-tuning Paradigm. Table 5 presents the complete experimental results of different fine-tuning paradigms. Two-stage training outperforms joint training for both BM25 and SBERT, with the largest gains on StackExchange and Theorem-based. Decoupling decomposition/interpretation learning from retrieval scoring likely reduces optimization interference and improves stability.

Fusion Methods. Table 6 presents the complete experimental results of different fusion methods. Simple *Sum* is consistently best. *Max* discards complementary evidence and underperforms; *RRF* helps some tails but lags on average; naive *Concat* notably hurts dense retrieval (embedding dilution). These results support additive, facet-wise evidence aggregation for complex intents.

Reasoning-based Retriever of ReasonIR-8B.. Table 7 presents the complete experimental results of ReasonIR-8B. With ReasonIR-8B, systems built on single long-form rewrites lead on StackExchange and Theorem-based, while **ReDI** is competitive on Coding. This highlights a retriever-adaptive trade-off: specialized high-capacity dense encoders favor globally coherent rewrites; decomposition brings larger benefits with lexical or lighter dense retrievers and on modular tasks.

Table 2: nDCG@10 on BRIGHT with different nums of sub-query + interpretation unit.

Num	Avg.All	Avg.SE	*StackExchange*							*Coding*		*Theorem-based*		
			Bio.	Earth.	Econ.	Psy.	Rob.	Stack.	Sus.	Leet.	Pony	AoPS	TheoQ.	TheoT.
Using BM25 Retriever														
3	26.67	33.28	44.83	50.82	25.25	36.20	22.43	26.79	26.63	21.81	7.99	5.07	25.65	26.57
5	27.14	34.16	44.60	51.27	27.26	37.66	21.62	28.45	28.28	21.82	7.42	4.77	24.50	28.01
7	27.53	34.70	46.07	49.12	26.79	38.13	22.03	31.41	29.35	23.35	7.91	4.94	24.49	26.76
9	27.68	35.17	47.10	50.22	24.76	41.00	23.28	30.64	29.20	23.16	5.49	4.92	24.78	27.67
11	27.53	34.75	47.57	48.52	26.80	39.91	22.93	29.05	28.49	23.25	6.82	5.07	24.90	27.06
13	27.46	34.87	47.75	48.76	26.52	41.40	23.30	29.09	27.26	22.42	6.45	4.58	25.17	26.86
15	26.99	34.34	48.17	48.72	24.88	40.53	22.14	29.59	26.33	21.95	5.45	4.15	24.83	27.18
Flex.	30.82	38.25	48.96	53.52	28.65	43.36	27.50	36.33	29.42	25.32	9.30	5.98	31.47	30.02
Using SBERT Retriever														
3	18.98	20.35	21.48	26.91	18.83	26.24	11.13	15.59	22.30	21.65	13.41	3.47	25.97	20.81
5	19.95	20.83	22.14	30.44	17.67	25.55	9.87	16.28	23.83	20.69	15.38	3.63	30.85	23.06
7	19.99	21.05	22.95	30.90	17.62	26.68	10.05	17.48	21.66	21.13	14.48	3.77	30.51	22.67
9	20.61	21.17	23.03	30.31	17.13	26.65	10.06	16.89	24.12	21.69	16.22	3.94	30.01	27.28
11	20.54	21.21	22.68	30.98	18.37	26.70	9.30	17.20	23.22	21.70	15.51	3.85	29.47	27.54
13	20.18	21.04	24.09	29.60	17.85	26.46	9.52	15.91	23.83	21.35	16.07	3.87	29.48	24.11
15	19.36	20.31	21.13	27.97	18.38	26.20	9.29	16.48	22.72	20.30	14.58	3.79	27.57	23.90
Flex.	22.80	23.70	25.00	32.30	20.80	28.00	13.80	20.20	25.60	25.20	17.10	6.20	33.20	25.80

Table 3: Performance of sparse retriever (BM25) under different k_3.

k_3	Avg.All	Avg.SE	*StackExchange*							*Coding*		*Theorem-based*		
			Bio.	Earth.	Econ.	Psy.	Rob.	Stack.	Sus.	Leet.	Pony	AoPS	TheoQ.	TheoT.
0	30.55	38.09	49.08	54.07	28.30	43.17	26.77	36.33	28.86	25.06	8.89	5.82	31.06	29.16
0.2	30.75	38.20	48.93	53.77	28.16	43.52	27.38	36.49	29.13	25.24	9.16	5.99	31.57	29.64
0.4	30.82	38.25	48.96	53.52	28.65	43.36	27.50	36.33	29.42	25.32	9.30	5.98	31.47	30.02
0.8	30.73	38.06	48.51	53.55	29.43	42.43	27.18	36.27	29.05	25.87	9.15	5.96	31.36	30.03
2	29.89	36.80	46.81	51.65	28.11	41.03	27.21	35.31	27.49	25.34	8.85	6.41	30.55	29.90
5	29.19	36.01	45.94	49.85	28.51	40.18	26.88	34.03	26.72	24.44	8.18	6.40	29.68	29.47
10	28.60	35.30	44.63	48.62	28.54	38.76	26.30	33.60	26.64	24.33	8.11	6.16	29.06	28.43
20	28.31	35.08	44.18	48.31	28.72	38.63	26.16	33.38	26.13	23.71	8.18	5.74	28.93	27.69
50	28.14	34.82	43.81	48.06	28.76	38.01	25.62	33.48	26.03	23.62	8.00	5.66	29.03	27.56
80	28.09	34.79	43.61	48.05	28.81	37.99	25.55	33.47	26.02	23.51	7.95	5.62	28.96	27.56
100	28.06	34.77	43.60	48.05	28.73	37.97	25.55	33.47	26.03	23.51	7.95	5.62	28.96	27.26

Table 4: Performance of dense retriever (SBERT) at varying interpretation weights.

$interp._w$	Avg. all	Avg. SE	*StackExchange*							*Coding*		*Theorem-based*		
			Bio.	Earth.	Econ.	Psy.	Rob.	Stack.	Sus.	Leet.	Pony	AoPS	TheoQ.	TheoT.
0.9	19.79	21.42	22.34	30.86	18.19	25.81	10.05	18.35	24.33	18.86	13.17	3.79	29.02	22.71
0.8	20.72	22.36	23.17	31.93	19.16	27.05	10.73	19.13	25.38	19.96	13.93	5.10	29.99	23.07
0.7	21.53	22.73	24.17	32.70	19.04	27.46	11.29	19.64	24.81	22.56	14.77	6.40	31.25	24.24
0.6	22.27	23.40	25.07	32.60	20.12	28.11	13.10	19.86	24.92	24.28	15.68	6.18	32.09	25.23
0.5	**22.76**	23.67	24.97	32.30	20.77	28.00	13.79	20.24	25.59	25.22	17.11	6.19	33.22	25.79
0.4	22.40	23.19	24.80	31.41	19.91	27.71	13.93	19.66	24.88	25.05	17.81	5.04	32.99	25.61
0.3	22.04	22.76	24.13	30.70	19.65	27.03	13.48	19.86	24.48	25.25	17.70	4.72	32.66	24.80
0.2	21.52	21.93	24.12	29.81	18.63	25.72	12.75	18.98	23.53	25.13	18.01	4.35	32.48	24.78
0.1	20.99	21.31	24.00	28.71	17.84	24.45	12.50	18.57	23.07	24.35	18.20	4.11	31.66	24.41

Table 5: nDCG@10 on BRIGHT: Joint vs. Two-Stage Training

Method	Avg. all	Avg. SE	*StackExchange*							*Coding*		*Theorem-based*		
			Bio.	Earth.	Econ.	Psy.	Rob.	Stack.	Sus.	Leet.	Pony	AoPS	TheoQ.	TheoT.
Using BM25 Retriever														
Joint	28.3	35.4	46.7	49.9	27.2	38.4	24.5	33.8	27.0	23.6	6.9	5.5	28.3	28.1
Two-Stage	30.8	38.3	49.0	53.5	28.7	43.4	27.5	36.3	29.4	25.3	9.3	6.0	31.5	30.0
Using SBERT Retriever														
Joint	20.8	21.8	22.7	29.9	18.9	25.7	12.0	18.9	24.8	24.0	14.8	5.1	30.2	22.1
Two-Stage	22.8	23.7	25.0	32.3	20.8	28.0	13.8	20.2	25.6	25.2	17.1	6.2	33.2	25.8

Table 6: nDCG@10 on BRIGHT with different retrieval fusion methods.

Method	Avg.All	Avg.SE	*StackExchange*							*Coding*		*Theorem-based*		
			Bio.	Earth.	Econ.	Psy.	Rob.	Stack.	Sus.	Leet.	Pony	AoPS	TheoQ.	TheoT.
Using BM25 Retriever														
Sum	30.8	38.3	49.0	53.5	28.7	43.4	27.5	36.3	29.4	25.3	9.3	6.0	31.5	30.0
Max	24.8	29.6	37.6	40.6	22.0	32.5	20.8	31.3	22.5	20.1	11.3	3.5	27.2	28.2
RRF ($\text{rrf}_k = 5$)	28.3	35.2	45.2	49.0	27.9	39.0	26.1	31.6	27.6	23.6	9.0	4.8	25.7	30.2
Concat	29.2	36.8	50.4	53.7	26.0	40.2	26.6	33.2	27.8	21.1	8.9	6.4	28.4	27.4
Using SBERT Retriever														
Sum	22.8	23.7	25.0	32.3	20.8	28.0	13.8	20.2	25.6	25.2	17.1	6.2	33.2	25.8
Max	19.1	18.9	23.2	24.1	15.7	24.1	8.4	14.7	22.3	17.7	22.8	3.5	28.7	24.2
RRF ($\text{rrf}_k = 5$)	21.5	22.5	24.8	31.1	19.2	27.3	11.8	18.5	25.1	22.6	15.7	4.6	32.1	24.8
Concat	12.4	12.4	11.9	22.5	9.4	19.3	5.1	8.3	9.9	17.0	8.8	2.1	17.8	16.8

Table 7: nDCG@10 on BRIGHT Benchmark with ReasonIR-8B retriever.

QU Model	Params	Avg. all	Avg. SE	*StackExchange*							*Coding*		*Theorem-based*		
				Bio.	Earth.	Econ.	Psy.	Rob.	Stack.	Sus.	Leet.	Pony	AoPS	TheoQ.	TheoT.
GPT-4	-	28.8	31.2	42.6	40.9	30.0	36.7	19.7	23.2	25.3	30.2	19.6	7.9	33.7	37.9
TongSearch	7B	31.8	34.2	45.3	44.7	30.9	40.2	20.4	27.2	30.7	26.4	22.9	10.3	38.0	37.7
DIVER	14B	32.2	35.9	47.8	44.6	32.2	43.7	24.6	30.8	27.6	31.2	11.4	9.1	40.7	37.5
ReDI	8B	30.7	32.7	41.8	43.9	26.5	37.4	21.3	32.2	25.5	35.7	21.3	8.6	39.0	34.7