1 Computation of Token-level Attention Score

In the Transformer architecture [3], the attention mechanism operates on three matrices: queries $\mathbf{Q} \in \mathbb{R}^{m\times d_k}$, keys, $\mathbf{K} \in \mathbb{R}^{n\times d_k}$, and values $\mathbf{V} \in \mathbb{R}^{n\times d_v}$. The attention function is defined as:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V}. \tag{1}$$

Here, each element $(QK^\top)_{ij} = q_i \cdot k_j$ measures the similarity between the i-th query and the j-th key. Dividing by $\sqrt{d_k}$ controls the magnitude of the dot products, preventing overly large values that could destabilize training. The softmax function is applied row-wise, normalizing the scores into a probability distribution whose weights sum to one. These attention weights are then used to compute a weighted sum of the value vectors V, producing the output representation.

At the token level, the attention weight

$$\mathbf{A}_{ij} = \text{softmax}\left(\frac{q_i \cdot k_j}{\sqrt{d_k}}\right)$$

represents how much token i attends to token j. These weights are computed dynamically for each input, enabling the model to capture semantic dependencies between tokens.

2 Proofs of Theorem 4.1 and Proposition 4.2

Proof of Theorem 4.1. According to previous studies of multi-objective optimization [1], the problem of balancing competing objectives can be expressed in the weighted sum form:

$$\max_{\mathcal{D}_{\text{rerank}} \subseteq \mathcal{D}_{\text{retrieve}}, |\mathcal{D}_{\text{rerank}}| \leq K} \text{Cover}(\mathcal{D}_{\text{rerank}}) - \lambda \cdot \text{Noise}(\mathcal{D}_{\text{rerank}}),$$

where λ is a trade-off coefficient between the two objectives.

Step 1: Coverage function. From Equation (7), the coverage of the selected set is:

$$\text{Cover}(\mathcal{D}_{\text{rerank}}) = \sum_{i=1}^{|\mathcal{E}|} \mathbf{e}_i \cdot \mathbb{P}(e_i \text{ covered by } \mathcal{D}_{\text{rerank}}).$$

For each information requirement e_i, the probability it is not covered by any selected document is:

$$\prod_{d \in \mathcal{D}_{\text{rerank}}} \left(1 - \mathbf{W}_{d,i}\right).$$

Thus, the probability it is covered is:

$$1 - \prod_{d \in \mathcal{D}_{\text{rerank}}} \left(1 - \mathbf{W}_{d,i}\right).$$

Introducing the binary selection variable $\mathbf{x}_d \in \{0, 1\}$, where $\mathbf{x}_d = 1$ if document d is selected, the above product becomes:

$$\prod_{d=1}^{K_1} \left(1 - \mathbf{W}_{d,i}\mathbf{x}_d\right).$$

Therefore, the coverage term is:

$$\sum_{i=1}^{|\mathcal{E}|} \mathbf{e}_i \left[1 - \prod_{d=1}^{K_1} \left(1 - \mathbf{W}_{d,i}\mathbf{x}_d\right)\right].$$

Step 2: Noise function. From Equation (8), noise is defined as:

$$\text{Noise}(\mathcal{D}_{\text{rerank}}) = \sum_{d \in \mathcal{D}_{\text{rerank}}} \text{DisUtil}(d|q).$$

Here, disutility for a document is defined as:

$$\text{DisUtil}(d|q) = 1 - \max_i \mathbf{W}_{d,i}\mathbf{e}_i.$$

Using $\mathbf{x}_d$ to indicate selection, the noise term becomes:

$$\sum_{j=1}^{K_1} \left(1 - \max(\mathbf{W}_j \odot \mathbf{e})\right) \mathbf{x}_j,$$

where $\odot$ denotes element-wise multiplication.

Step 3: Binary optimization formulation. Combining the coverage and noise terms, and introducing the budget constraint $\sum_i \mathbf{x}_i \leq K$, we obtain:

$$\begin{aligned}&\max_{\mathbf{x}} \sum_{i=1}^{|\mathcal{E}|} \mathbf{e}_i \left[1 - \prod_{d=1}^{K_1} \left(1 - \mathbf{W}_{d,i}\mathbf{x}_d\right)\right] - \lambda \sum_{j=1}^{K_1} \left(1 - \max(\mathbf{W}_j \odot \mathbf{e})\right) \mathbf{x}_j,\\ &\text{s.t.} \sum_i \mathbf{x}_i \leq K, \quad \mathbf{x}_i \in \{0,1\}, \ \forall i.\end{aligned}$$

This exactly matches the statement in Equation (9), proving that the denoised coverage-aware document selection problem can indeed be formulated as a 0−1 integer (binary) multi-objective optimization problem. □

Proof of Proposition 4.2. We prove the claim by showing each per-evidence term is monotone and submodular, and then use closure properties of submodular functions. Let's define the multi-objective objective function in Problem (6) as $g(S) = G(S) - \lambda \cdot R(S)$.

(1) Per-evidence term. Fix an evidence e_i. Define

$$h_i(S) \ = \ 1 - \prod_{d \in S} (1 - \mathbf{W}_{d,i}), \qquad S \subseteq \mathcal{D}.$$

For any $S \subseteq T \subseteq \mathcal{D}$ and any $d \in \mathcal{D} \setminus T$ the marginal gain of adding d is

$$\Delta_d(S) \ = \ h_i(S \cup \{d\}) - h_i(S) = \prod_{t \in S}(1 - \mathbf{W}_{t,i}) - \prod_{t \in S}(1 - \mathbf{W}_{t,i})(1 - \mathbf{W}_{d,i}) = \mathbf{W}_{d,i} \prod_{t \in S}(1 - \mathbf{W}_{t,i}).$$

Similarly,

$$\Delta_d(T) = \mathbf{W}_{d,i} \prod_{t \in T}(1 - \mathbf{W}_{t,i}).$$

Since $0 \leq 1 - \mathbf{W}_{t,i} \leq 1$ for all t, we have

$$\prod_{t \in S}(1 - \mathbf{W}_{t,i}) \geq \prod_{t \in T}(1 - \mathbf{W}_{t,i}),$$

and therefore $\Delta_d(S) \geq \Delta_d(T)$. This is exactly the diminishing returns property, so $h_i(\cdot)$ is submodular. Moreover $\Delta_d(S) = \mathbf{W}_{d,i} \prod_{t \in S}(1 - \mathbf{W}_{t,i}) \geq 0$, so $h_i(\cdot)$ is monotone non-decreasing.

(2) Sum preserves submodularity and monotonicity for Coverage Objective. The coverage objective $G(S) = \sum_i \mathbf{e}_i h_i(S)$ is a nonnegative linear combination of the functions h_i. Nonnegative linear combinations of submodular (resp. monotone) functions remain submodular (resp. monotone) [2]. Hence $G(\cdot)$ is monotone and submodular.

(3) Per-document term. Fix a document d. Define its disutility score as

$$\text{DisUtil}(d) \ = \ 1 - \max_{e \in \mathcal{E}} \mathbb{P}(d \mid e)\mathbb{P}(e \mid q).$$

Now consider the noise function over a set of documents

$$S \subseteq \mathcal{D} : R(S) \ = \ \sum_{d \in S} \text{DisUtil}(d).$$

Notice that each per-document term $\mathbf{c}_d \cdot \text{DisUtil}(d)$ depends only on d and the evidence set $\mathcal{E}$, but not on the other documents in S. Hence, each term contributes additively and independently.

(4) Submodularity and monotonicity of Noise Objective. For any $S \subseteq T \subseteq \mathcal{D}$ and any document d$\in \mathcal{D} \setminus T$, the marginal gain of adding d is

$$\Delta_d(S) \ = \ R(S \cup \{d\}) - R(S) \ = \text{DisUtil}(d).$$

Similarly,

$$\Delta_d(T) = \text{DisUtil}(d).$$

Since the marginal contribution of d is identical regardless of the context set, the diminishing returns property holds trivially. This means $R(\cdot)$ is modular (a special case of submodular) [2].

Moreover, as $\text{DisUtil}(d) \geq 0$, we have $\Delta_d(S) \geq 0$, implying monotonicity. Therefore, the noise objective

$$R(S) = \sum_{d \in S} \left(1 - \max_{e \in \mathcal{E}} \mathbb{P}(d \mid e)\mathbb{P}(e \mid q)\right)$$

is a monotone modular function, and hence also submodular.

(3) NP-hardness. Maximizing our proposed objective $g(\cdot)$ under a cardinality constraint is generally NP-hard. A cardinality constraint means that the solution set S is restricted to contain at most K documents (i.e., $|S| \leq K$), which reflects the practical setting of re-ranking where we can only select a limited number of documents for the generator.

This complexity includes the well-known Max-k-Cover problem as a special case. Specifically, if each document d either fully covers a particular information requirement $e_i(p_{d,i} = 1)$ or does not cover it at all $(p_{d,i} = 0)$, then the coverage function $h_i(S)$ becomes an indicator function that equals 1 if e_i is covered by at least one document in S, and 0 otherwise. Under this setting, maximizing $g(S)$ is equivalent to selecting K documents to cover as many weighted information requirements as possible, which is exactly the Max-k-Cover problem, which is known to be NP-hard [2].

Therefore, even in the more general case where $p_{d,i}$ takes continuous values (representing partial coverage), maximizing $g(\cdot)$ remains NP-hard. This justifies the need for an efficient approximation algorithm, such as our proposed greedy approach. □

References

[1] Kalyanmoy Deb, Karthik Sindhya, and Jussi Hakanen. 2016. Multi-objective optimization. In *Decision sciences*. CRC Press, 161–200.

[2] Rishabh Krishnan Iyer. 2015. *Submodular optimization and machine learning: Theoretical results, unifying and scalable algorithms, and applications*. Ph. D. Dissertation.

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).