Retrieval-Augmented Neural Team Formation

Abstract. This study investigates the formation of expert teams that collectively possess a specified skill set. While traditional methods have employed graph search techniques to identify subgraphs that meet skill requirements or neural architectures to map skills to experts, we introduce a novel approach that emphasizes both cohesive team dynamics and comprehensive skill coverage. Our retrieval-augmented generation model is designed to optimize the probability of successful collaboration among team members. Extensive experiments demonstrate that our proposed method significantly outperforms existing state-of-the-art approaches, offering a more effective solution for expert team formation.

1 Introduction

In today's fast-paced and complex project environments, assembling effective teams has become increasingly critical across various sectors, including scientific research, engineering, and healthcare. Complex projects often demand a diverse set of skills and expertise that no single individual can provide. Consequently, the basic form of the team formation problem involves selecting a group of individuals from a pool of candidates who collectively possess the required skill set [11,6]. Candidates for team formation can be found on platforms such as LinkedIn, GitHub, and Google Scholar. However, selecting the right experts is difficult due to the vast number of potential candidates available. [15]. The team formation problem can be likened to the set covering problem [9], where the objective is to find the smallest subcollection of sets (in this case, experts) whose union covers a target set of skills [1,3]. Traditional solutions to expert selection often fall short in creating optimal team compositions, as they typically ignore key factors like collaboration dynamics and interpersonal relationships [7,10].

Recognizing the complexities of team formation, there is increasing interest in using machine learning to tackle this challenge. Neural network-based methods have been developed to model the intricate relationships between required skills and expert capabilities. [5,16]. Specifically, neural variational Bayesian models address the sparsity and uncertainty in expert-skill relationships, capturing probabilistic associations to enhance robustness under uncertain conditions [15]. Recurrent neural network-based methods have been applied to capture the sequential dynamics of team formation, enabling predictions of future team compositions that adapt to changing project needs [19,13]. Bayesian inference techniques also offer valuable insights by incorporating prior knowledge and managing uncertainty in a principled manner [2]. Despite these advancements, existing models often prioritize either skill matching or collaboration history, failing to effectively integrate both aspects in forming optimal teams. [16,18]

To address these challenges, we propose a novel approach on collaborative team formation that emphasizes not only comprehensive skill coverage but also the importance of past collaboration in enhancing team cohesion and productivity. By leveraging a Retrieval-Augmented Generation (RAG) [12,4] model, our approach retrieves historical team formation data and combines it with a generation module that predicts expert teams capable of fulfilling the required skills while maintaining effective collaboration dynamics. In summary, the contributions of this paper are as follows: **(1)** We have

proposed a custom RAG model for the Team Formation problem as a downstream task. The proposed model retrieves and utilizes items that are effective and informative for generating ideal teams [1]. **(2)** Our proposed method shows robustness in generating ideal teams thanks to its custom RAG architecture. Proposed teams are predicted using a custom generator model that is customized to only proposed experts for a given skill set. **(3)** Through extensive experiments, we demonstrated the effectiveness of our proposed method compared to state-of-the-art baselines from different categories of techniques.

2 Methodology

Given a set of experts $E = \{e_1, e_2, \ldots, e_n\}$ and a set of required skills $S = \{s_1, s_2, \ldots, s_m\}$, the objective of *collaborative team formation* involves selecting a subset of experts to form an *optimal team* that balances two criteria: (1) **Skill Coverage**: If each expert e_i possesses a subset of skills $S_{e_i} \subseteq S$, the objective of team formation is to identify a subset of experts $E_t \subseteq E$ that collectively covers all required skills S, i.e., $\bigcup_{e_i \in E_t} S_{e_i} = S$. (2) **Collaborative Effectiveness**: The team should prioritize experts who have a history of working together. Let C denote the collaboration matrix, where $C_{ij} = 1$ if experts e_i and e_j have collaborated previously, and $C_{ij} = 0$ otherwise. The objective is extended to maximize collaborative compatibility among team members by maximizing the sum of C_{ij} for all pairs $(e_i, e_j) \in E_t$, i.e., $\max_{E_t \subseteq E} \sum_{(e_i,e_j) \in E_t} C_{ij}$.

Figure 1 illustrates the architecture designed to estimate the function $f : S \rightarrow E$ which maps the powerset of skills to powerset of experts. This is achieved through a Retrieval-Augmented Generation (RAG) model that combines retrieval and generation modules to enhance the formation of collaborative teams. The RAG architecture leverages both pre-trained dense vector representations and historical team formations. Given the set of required skills S as input, the model begins by transforming S into a dense vector v_s. The retriever module uses this vector representation to search for relevant past teams, aligning the retrieval process with similar skill requirements from historical data. These retrieved results serve as the context for the generator module, which predicts the optimal expert team sequence $E_t = \{e_1, e_2, \ldots, e_k\} \subset E$. The generator fine-tunes its output by integrating the retrieved team formations, enhancing the generated team's relevance and ensuring the proposed team meets the required skill set with coherence and historical collaboration patterns. This combination of retrieval and generation components enables the model to balance skill coverage and collaborative effectiveness. The architecture's components are explained in detail as follows.

Encoder The encoder is responsible for transforming the input skill set S into a dense embedding that can be used for both retrieval and generation. Specifically, we employ a custom T5-based encoder tailored for the team formation problem, which generates skill embeddings v_s to capture the semantic relationships within S. To enhance the encoder's performance, we incorporate a contrastive loss function. This novel addition improves the model's ability to differentiate skill sets by bringing embeddings of semantically similar teams closer in the vector space, while increasing the separation between embeddings of dissimilar teams.

Retriever Given the input skill embedding v_s, the retriever retrieves the top-R relevant teams, denoted by $\mathcal{T} = \{(S_i, E_i)\}_{i=1}^{R}$, where the pair (S_i, E_i) includes the skill set of

[1] Our code and dataset are publicly available at `https://tinyurl.com/43frmv6p`

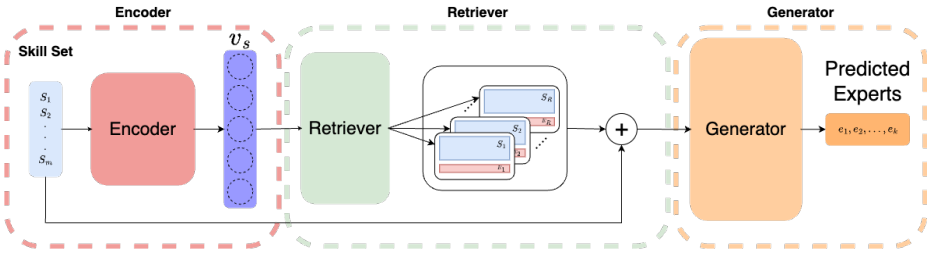


Fig. 1: RAG Architecture for Collaborative Team Formation.

the i-th team, and E_i is the corresponding expert team, from a pre-built FAISS index of historical teams. The relevance is determined by calculating the similarity between the input skill set S and each historical skill setS_i using *L2 distance*.

$$\text{dist}(S, (S_i, E_i)) = \|v_s - v_{s_i}\|_2 \tag{1}$$

where v_{s_i} represents the skill embedding of the i-th historical team retrieved from the index. The skill expert pair with the smallest distances are considered the most relevant.

Generator The generator takes the required skill set S and the retrieved historical teams $\mathcal{T} = \{(S_i, E_i)\}_{i=1}^{R}$ and to generate the expert team E_t. The generator is based on T5 sequence-to-sequence architecture. The probability of generating E_t is modeled as:

$$P(E_t|S) = \sum_{i=1}^{R} P(E_t|S, (S_i, E_i))P((S_i, E_i)|S) \tag{2}$$

where $P((S_i, E_i)|S)$ is the retrieval probability computed from the similarity score, and $P(E_t|S, (S_i, E_i))$ is the probability of generating the team E_t given the skill set S and historical retrieved relevant teams $\mathcal{T}$.

Model Training Our model training process consists of two key stages: first, pre-training the encoder with contrastive learning to create robust skill set representations, and second, integrating the pre-trained encoder into the RAG architecture to fine-tune the generator for optimized team formation. To train the encoder, we use the contrastive loss as follows:

$$\mathcal{L}_{\text{contrastive}} = \sum_{i=1}^{N_{\text{pos}}} \text{CosineLoss}(v_s, v_{\text{pos}}^i, 1) + \sum_{j=1}^{N_{\text{neg}}} \text{CosineLoss}(v_s, v_{\text{neg}}^j, -1) \tag{3}$$

where N_{pos} and N_{neg} denote the number of positive and negative samples, respectively. v_s is the embedding of the input skill set, v_{pos}^i represents the i-th positive embedding, and v_{neg}^j represents the j-th negative embedding. The loss function optimizes the encoder by reducing the cosine distance to positive samples (labeled as 1) and increasing the distance from negative samples (labeled as -1), enhancing its ability to distinguish between relevant and irrelevant teams. Algorithm 1 outlines the encoder training process.

The generator is trained by minimizing the negative log-likelihood, defined as:

$$\mathcal{L}_{\text{generator}} = -\log P(E_t^*|S, \mathcal{T}) \tag{4}$$

Algorithm 1: Encoder Training with Contrastive Learning

Input: Dataset $\mathcal{T} = \{(S_i, E_i)\}_{i=1}^{N_{\text{train}}}$, encoder $\mathcal{E}$
Output: Trained Encoder $\mathcal{E}$

foreach $(S_i, E_i) \in \mathcal{T}$ **do**
  Sort $\{(S_j, E_j) \mid j \neq i\}$ by similarity of E_j to E_i
  Select top R pairs as positive samples
end
repeat
  Sample $b \subset \mathcal{T}$
  $L_b \leftarrow 0$
  foreach $(S_i, E_i) \in b$ **do**
    $v_s \leftarrow \mathcal{E}(S_i)$
    Retrieve v_{pos} and select v_{neg} by random sampling of $E_j \neq E_i$
    $L_{pos} \leftarrow \sum \text{CosineLoss}(v_s, v_{pos}, 1)$
    $L_{neg} \leftarrow \sum \text{CosineLoss}(v_s, v_{neg}, -1)$
    $L_b \leftarrow L_b + L_{pos} + L_{neg}$
  end
  $L_b \leftarrow L_b / (|b| \cdot 2R)$
  Backpropagate L_b and update $\mathcal{E}$
until *convergence*
return $\mathcal{E}$

where E_t^* is the ground truth expert team for the input skill set S, and $\mathcal{T}$ is the list of the top-R retrieved teams. This loss function encourages the model to generate expert teams that best match the skill requirements specified by the input query, considering both the input skill set and the retrieved historical teams.

Model Inference During inference, our model predicts the expert team E_t for a given set of required skills S as summarized in Algorithm 2. First, the trained encoder transforms the input skill set S into a dense embedding vector v_s by applying mean pooling over the last hidden layer of the encoder's output. Next, the retriever uses this embedding to find the top-R relevant historical teams $\mathcal{T} = \{(S_i, E_i)\}_{i=1}^{R}$ by minimizing the L2 distance. Finally, the generator predicts the expert team E_t based on the input skill set S and the retrieved historical teams $\mathcal{T}$.

3 Experiments

Datasets We have adopted datasets that have been used by earlier work such as *Rad et al.* [15], *Lapas et al.* [11] and *Zihayat et al.* [20], DBLP[2] and Dota2[3]. DBLP tracks co-authorship networks among researchers, capturing their publications, expertise, and historical collaborations. The Dota2 dataset is derived from the gaming environment, where each game records player heroes and configurations. In the DBLP dataset, our task is to suggest authors as experts to cover the skills needed for a publication. In the Dota2 dataset, we aim to recommend players as experts who can win a game based on the

[2] `https://www.aminer.org/citation`
[3] `https://www.kaggle.com/datasets/devinanzelmo/dota-2-matches`

Algorithm 2: Inference Procedure using RAG Model

Input: $D_{\text{test}} = \{(S_i, E_i)\}_{i=1}^{N_{\text{test}}}$; Trained_Encoder; Trained_Generator; Custom_Tokenizer; $\mathcal{T} = \{(S_i, E_i)\}_{i=1}^{N_{\text{train}}}$; R

foreach $(S_i, E_i) \in D_{test}$ **do**

- $S_i^{\text{tok}} \leftarrow \text{Tokenizer}(S_i)$
- $\mathbf{H}_i \leftarrow \text{Encoder}(S_i^{\text{tok}})$
- $h_i \leftarrow \frac{1}{N} \sum_{j=1}^{N} \mathbf{H}_{ij}$
- $\{(s_{ir}, e_{ir})\}_{r=1}^{R} \leftarrow \text{Retriever}(h_i, \mathcal{T}, R)$
- $\hat{E}_i \leftarrow \text{Generator}\left(S_i, \{(s_{ir}, e_{ir})\}_{r=1}^{R}\right)$
- Evaluate metrics comparing $\hat{E}_i$ and E_i^{tok}

end

Table 1: Dataset statistics.

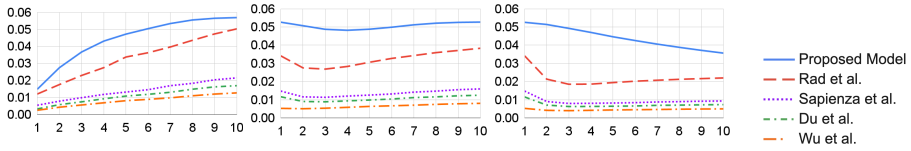
Dataset	# teams	# Unique experts	# Avg. experts/team	# Unique skills	# Avg. skills/team	# Avg. skills/expert
DBLP	10,675	10,831	4.04	2000	14.55	58.09
Dota2	640	2,727	5	3,057	31.46	36.86

configurations and opponent players. These tasks enable us to explore team formation strategies that maximize complementary skills and cohesive past performances. Detailed statistics for both datasets can be found in Table 1.

Metrics Following earlier work [15,11,20], we have adopted 3 retrieval metrics: (1) Recall, (2) MAP and (3) NDCG. To measure efficacy, we employ a 10-fold cross-validation strategy, with scores based on exact matches. The model is considered successful only if it proposes the exact team of experts expected for a given set of skills.

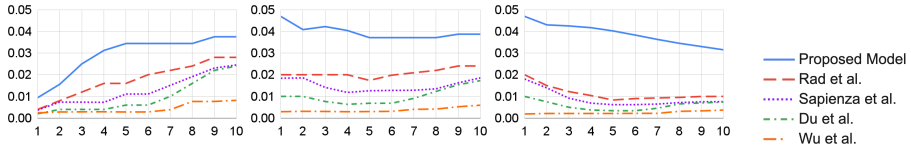
Baselines We have included state-of-the-art techniques from diverse approaches: **(1)** `Rad et al.` [15]:This paper is currently state-of-the-art and represents the neural variational Bayesian-based group of methods. **(2)** `Sapienza et al.` [17]: This paper utilizes autoencoder neural network architecture to learn a mapping function from skills to experts. **(3)** `Wu et al.` [19]: This paper uses recurrent neural network (RNN) architecture based on LSTM to learn past collaborations in the past. **(4)** `Du et al.` [2]: In this paper, the authors proposed a Bayesian Group Ranking (BGR) method to optimize the weights of a Bayesian inference model.

Findings The evaluation results for DBLP and Dota2 datasets are shown in Figure 2 and Figure 3 respectively. Comparing methods using ranking metrics, we make several observations: (1) in general, methods that are using encoder-decoder neural networks, i.e. `Rad et al.` [15], `Sapienza et al.` [17] and our proposed method are outperforming others. This can be a result of multiple facts: Sparcity is a known problem in Team Formation datasets [8,15]. Often, experts are seen collaborating on a limited number of teams, which is extremely low compared to the total number of collaborations in the dataset. Therefore models face a challenging situation where they need to learn potential



Recall NDCG MAP

Fig. 2: DBLP dataset performance results for Recall, NDGC and MAP metrics.



Recall NDCG MAP

Fig. 3: Dota2 dataset performance results for Recall, NDGC and MAP metrics.

collaboration from a few observations. This can especially be seen in the Dota2 dataset in Figure 3. Other reasons can be neural networks' abilities to capture meaningful connections from past collaborations due to their memorization and generalization abilities [14]; (2) by studying two other methods, i.e. `Wu et al.` [19] and `Du et al.` [2], we make an observation about the impact of memorization and generalization. `Wu et al.` uses recurrent neural networks to capture collaborations in the past; this makes this method a prime example of a memorization-focused method. In contrast, `Du et al.` uses Bayesian inference in proposing experts, which makes it representative of the generalization-focused method. While both methods lean on one side of the memorization-generalization trade-off, it can be seen that `Du et al.` demonstrates better performance. This proves while in the Team Formation problem, it is important to meet both memorization and generalization aspects, it is more crucial to ensure models have good inferences that result in strong generalization power. Our proposed method addresses memorization by using a custom retrieval function to extract past collaborations. Moreover, for the sake of generalization, our method uses a transformer-based neural network as a generator to propose teams of experts; (3) based on all metrics scores in Figures 2 and 3, we observe a stable superior performance of our proposed method compared to state-of-the-art over the full range of top-k cut-offs. This means our proposed method not-only is not only able to find more relevant experts for given skill sets but also keeps proposing experts that are relevant and have seen in past collaborations by increasing the top-k threshold.

4 Concluding Remarks

This paper introduces a retrieval-augmented generation model that effectively integrates historical collaboration data with required skill sets, enabling the selection of expert teams with both the necessary competencies and a proven track record of effective teamwork. Extensive experiments demonstrate that our method significantly outperforms existing state-of-the-art techniques, underscoring the importance of balancing individual expertise with interpersonal dynamics in forming successful teams.

References

1. Berktaş, N., Yaman, H.: A Branch-and-Bound Algorithm for Team Formation on Social Networks. INFORMS Journal on Computing **33**(3), 1162–1176 (July 2021). https://doi.org/10.1287/ijoc.2020.1000, https://ideas.repec.org/a/inm/orijoc/v33y2021i3p1162-1176.html
2. Du, Y., Meng, X., Zhang, Y., Lv, P.: GERF: A group event recommendation framework based on learning-to-rank. IEEE Trans. Knowl. Data Eng. **32**(4), 674–687 (2020). https://doi.org/10.1109/TKDE.2019.2893361, https://doi.org/10.1109/TKDE.2019.2893361
3. Elashmawi, W., Fouad, A., Tawhid, M.: An improved particle swarm optimization with a new swap operator for team formation problem. Journal of Industrial Engineering International **15** (07 2018). https://doi.org/10.1007/s40092-018-0282-6
4. Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Guo, Q., Wang, M., Wang, H.: Retrieval-augmented generation for large language models: A survey. ArXiv **abs/2312.10997** (2023), https://api.semanticscholar.org/CorpusID:266359151
5. Hamidi Rad, R., Bagheri, E., Kargar, M., Srivastava, D., Szlichta, J.: Retrieving skill-based teams from collaboration networks. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 2015–2019. SIGIR '21, Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3404835.3463105, https://doi.org/10.1145/3404835.3463105
6. Juárez, J., Brizuela, C.A.: A multi-objective formulation of the team formation problem in social networks: preliminary results. In: Proceedings of the Genetic and Evolutionary Computation Conference. p. 261–268. GECCO '18, Association for Computing Machinery, New York, NY, USA (2018). https://doi.org/10.1145/3205455.3205634, https://doi.org/10.1145/3205455.3205634
7. Kargar, M., An, A.: Discovering top-k teams of experts with/without a leader in social networks. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management. p. 985–994. CIKM '11, Association for Computing Machinery, New York, NY, USA (2011). https://doi.org/10.1145/2063576.2063718, https://doi.org/10.1145/2063576.2063718
8. Kargar, M., Golab, L., Srivastava, D., Szlichta, J., Zihayat, M.: Effective keyword search over weighted graphs. IEEE Transactions on Knowledge and Data Engineering **34**(2), 601–616 (2022). https://doi.org/10.1109/TKDE.2020.2985376
9. Karp, R.M.: Reducibility among Combinatorial Problems, pp. 85–103. Springer US, Boston, MA (1972). https://doi.org/10.1007/978-1-4684-2001-2_9, https://doi.org/10.1007/978-1-4684-2001-2_9
10. Kouvatis, I., Semertzidis, K., Zerva, M., Pitoura, E., Tsaparas, P.: Forming compatible teams in signed networks. ArXiv **abs/2001.03128** (2020), https://api.semanticscholar.org/CorpusID:210116773
11. Lappas, T., Liu, K., Terzi, E.: Finding a team of experts in social networks. In: IV, J.F.E., Fogelman-Soulié, F., Flach, P.A., Zaki, M.J. (eds.) Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, France, June 28 - July 1, 2009. pp. 467–476. ACM (2009). https://doi.org/10.1145/1557019.1557074, https://doi.org/10.1145/1557019.1557074
12. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.t., Rocktäschel, T., Riedel, S., Kiela, D.: Retrieval-augmented generation for knowledge-intensive nlp tasks. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems. vol. 33, pp. 9459–9474. Curran Associates, Inc. (2020), https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf

13. Mikolov, T., Karafiát, M., Burget, L., Cernocký, J., Khudanpur, S.: Recurrent neural network based language model. In: Proceedings of the 11th Annual Conference of the International Speech Communication Association. vol. 2, pp. 1045–1048 (09 2010). https://doi.org/10.21437/Interspeech.2010-343
14. Rabin, M.R.I., Hussain, A., Alipour, M.A., Hellendoorn, V.J.: Memorization and generalization in neural code intelligence models. Inf. Softw. Technol. **153**, 107066 (2023). https://doi.org/10.1016/J.INFSOF.2022.107066, https://doi.org/10.1016/j.infsof.2022.107066
15. Rad, R.H., Fani, H., Bagheri, E., Kargar, M., Srivastava, D., Szlichta, J.: A variational neural architecture for skill-based team formation. ACM Trans. Inf. Syst. **42**(1), 7:1–7:28 (2024). https://doi.org/10.1145/3589762, https://doi.org/10.1145/3589762
16. Rad, R.H., Seyedsalehi, S., Kargar, M., Zihayat, M., Bagheri, E.: A neural approach to forming coherent teams in collaboration networks. In: International Conference on Extending Database Technology (2022), https://api.semanticscholar.org/CorpusID:247863576
17. Sapienza, A., Goyal, P., Ferrara, E.: Deep neural networks for optimal team composition. Frontiers Big Data **2**, 14 (2019). https://doi.org/10.3389/fdata.2019.00014, https://doi.org/10.3389/fdata.2019.00014
18. Vinella, F.L., Hu, J., Lykourentzou, I., Masthoff, J.: Crowdsourcing team formation with worker-centered modeling. Frontiers in Artificial Intelligence **5** (2022), https://api.semanticscholar.org/CorpusID:249068234
19. Wu, C., Ahmed, A., Beutel, A., Smola, A.J., Jing, H.: Recurrent recommender networks. In: de Rijke, M., Shokouhi, M., Tomkins, A., Zhang, M. (eds.) Proceedings of the Tenth ACM International Conference on Web Search and Data Mining, WSDM 2017, Cambridge, United Kingdom, February 6-10, 2017. pp. 495–503. ACM (2017). https://doi.org/10.1145/3018661.3018689, https://doi.org/10.1145/3018661.3018689
20. Zihayat, M., An, A., Golab, L., Kargar, M., Szlichta, J.: Authority-based team discovery in social networks. In: Markl, V., Orlando, S., Mitschang, B., Andritsos, P., Sattler, K., Breß, S. (eds.) Proceedings of the 20th International Conference on Extending Database Technology, EDBT 2017, Venice, Italy, March 21-24, 2017. pp. 498–501. OpenProceedings.org (2017). https://doi.org/10.5441/002/edbt.2017.54, https://doi.org/10.5441/002/edbt.2017.54