← Go to **NeurIPS 2025 Conference** homepage

# GFMate: Empowering Graph Foundation Models with Pre-training-agnostic Test-time Prompt Tuning

10 Apr 2025 (modified: 29 May 2025) NeurIPS 2025 Conference Submission Conference, Senior Area Chairs, Area Chairs, Reviewers, Authors Revisions CC BY 4.0
**Keywords:** Graph Foundation Models, Test-time Prompt Tuning

**Abstract:**
Graph foundation models (GFMs) have shown strong potential in graph-based applications owing to its generalisation ability across diverse tasks and domains. Recent advancements in graph prompt tuning enhance GFMs by tuning auxiliary prompts in few-shot scenarios. Despite their progress, most prompt-enhanced GFM methods inject domain-specific information from source datasets into GFM pre-training process, which makes **these prompting techniques entangled with GFMs and lack generalisability**. Furthermore, these approaches only utilise the labelled samples in the target dataset, despite the fact that unlabelled test nodes are generally accessible during few-shot tuning. This typically **neglects the rich information contained in the test data, leaving the distribution shift between the training and testing samples unresolved**. In light of the above insights, this paper proposes **GFMate** to empower GFMs with a novel pre-training-agnostic test-time prompt tuning strategy. GFMate introduces a centroid prompt and a layer prompt only after pre-training to avoid entanglement with GFMs. In addition, a complementary learning objective is devised to tune the prompts in the few-shot scenario with both labelled and unlabelled target data, effectively mitigating the impact of the train-test distribution shift during GFMs downstream adaptation. Extensive experiments on 12 benchmark datasets across diverse domains demonstrate the superior performance and efficiency of GFMate, with improvements of up to **30.63%**. The code will be released upon acceptance.

**Checklist Confirmation:** I confirm that I have included a paper checklist in the paper PDF.
**Financial Support:**
**Reviewer Nomination:**
**Responsible Reviewing:** We acknowledge the responsible reviewing obligations as authors.
**Primary Area:** Deep learning (e.g., architectures, generative models, optimization for deep networks, foundation models, LLMs)
**LLM Usage:** Not used at all (you can then skip the rest)
**Declaration:** I confirm that the above information is accurate.
**Submission Number:** 1632

Filter by reply type... Filter by author... Search keywords... Sort: Newest First
*15 / 15 replies shown*

Add:

## Official Comment by Area Chair p1E9

Official Comment by Area Chair p1E9 05 Aug 2025, 22:49
Program Chairs, Reviewer 2PxR, Reviewer 7Hfg, Reviewer Yv5F, Reviewer ha7m, Senior Area Chairs, Area Chairs, Reviewers, Reviewers Submitted, Authors

**Comment:**
Dear reviewers,

Thank you for your effort in the reviews.

As the discussion period ends soon, please read authors' rebuttals and check if they have addressed your concern.

If authors have resolved your questions, do tell them so.

If authors have not resolved your questions, do tell them so too.

Thanks.

AC

Add:

## Official Comment by Authors

Official Comment by Authors 04 Aug 2025, 21:42
Program Chairs, Senior Area Chairs, Area Chairs, Authors

**Comment:**

Dear AC, SAC, and PC,

**We sincerely thank all the reviewers for their feedback and have carefully and thoroughly addressed all their concerns.** However, we have not yet received any responses from all the reviewers.

As the discussion phase reaches its end, we would appreciate your help in reaching out to the reviewers and encouraging them to engage in the discussion. We believe that timely and constructive discussion during the rebuttal period benefits both authors and reviewers.

**Thank you again for your support and dedication.**

Best regards,

Authors of Paper 1632

Add:

## Official Review of Submission1632 by Reviewer 2PxR

Official Review by Reviewer 2PxR 29 Jun 2025, 12:09 (modified: 24 Jul 2025, 21:31)
Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer 2PxR Revisions

**Summary:**

The authors propose GFMate for graph prompt tuning. GFMate empowers GFMs with centroid prompts and layer prompts. A graph complementary learning paradigm is introduced in GFMate to integrate information from unlabeled data. Extensive experiments are conducted to demonstrate the effectiveness of the proposed method.

**Strengths And Weaknesses:**

**Strengths**

- Experimental settings are clearly specified in experiments.
- The experiments are comprehensive and convincing.

**Weaknesses**

- The description about graph prompt tuning — *learnable prompt embeddings are incorporated into GFMs during pre-training* — is weird. To my knowledge, most existing graph prompting methods, such as GPPT, GPF, ProNoG, and GraphPrompt, introduce learnable prompts only after pre-training (i.e., during the test-time). They already follow the pipeline in Figure 1(b). The claimed challenge that *these prompt designs are pre-training-entangled and not easily generalisable to other GFMs* does not hold.
- Equation (9) should be the sum of negative log values.
- Preposition 1 is hard to follow. The authors should specify the meanings of the mentioned variables in Proposition 1, including , , and .
- The proof of Preposition 1 is questionable. The authors introduce several undefined variables e.g., , which is rather confusing.
- in Equation (3) should be deleted since Equation (3) does not use .

**Quality:** 2: fair
**Clarity:** 1: poor
**Significance:** 2: fair
**Originality:** 2: fair
**Questions:**

- What is the difference between in Equation (5) and in Equation (8)?
- Could the authors explain how to obtain the second row of Equation (14)?

- Graph classification is usually inductive, which means the unlabeled graphs are usually not available during prompt tuning. How should the proposed method handle this scenario?

**Limitations:**
I did not find any potential negative societal impact.

**Rating:** 2: Reject: For instance, a paper with technical flaws, weak evaluation, inadequate reproducibility and incompletely addressed ethical considerations.
**Confidence:** 4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.
**Ethical Concerns:** NO or VERY MINOR ethics concerns only
**Paper Formatting Concerns:**
I did not find any major formatting issues in this paper.

**Code Of Conduct Acknowledgement:** Yes
**Responsible Reviewing Acknowledgement:** Yes

Add:

## Rebuttal by Authors Rebuttal by Authors

27 Jul 2025, 16:32 (modified: 01 Aug 2025, 04:30)
Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors Revisions

**Rebuttal:**
**We sincerely thank you for your valuable question.** We have thoroughly addressed your concerns. We summarise them and **highlight** the key information for improved clarity and readability as follows:

> **Re W1: GFMs, test-time tuning, and claim**

We would like to respectfully clarify and disagree that:

1. **GFM definition**: (i) **The reviewer-mentioned GPPT, GPF, ProNoG, and GraphPrompt are not GFMs**, do **not support cross-dataset generalisation**, and are **supervised learning methods for a single domain dataset**. (ii) Our GFMate is GFM for **multiple datasets in a cross-domain setting**, where the pre-training and downstream target domain are different.
2. **Test-time prompt definition**: Test-time training methods use **unlabelled test data** to update the trained model. (i) **The reviewer-mentioned GPPT, GPF, ProNoG, and GraphPrompt are not test-time prompt tuning because they only use labelled test data** for model adaptation. Most existing/latest GFM prompt methods, GCoPE, MDGPT, SAMGPT, and MDGFM) are **not test-time tuning either**, because they rely solely on **labelled few-shot test data** for prompt fine-tuning. (ii) Our GFMate is test-time graph prompt tuning because **GFMate uses unlabelled test data** to tune the prompt when the model is being tested with real data in the real world.
3. Thus, **the reviewer-mentioned GPPT, GPF, ProNoG, and GraphPrompt do not belong to Figure 1** because Figure 1(a) is for GFMs without test-time prompt tuning, and Figure 1(b) is for our GFMate with test-time prompt tuning. But **GPPT, GPF, ProNoG, and GraphPrompt are not GFMs nor test-time** prompt tuning methods.
4. Thus, our claim **"these prompt designs are pre-training-entangled and not easily generalisable to other GFMs"** does hold.

> **Re W2: Equation 9 should be the sum of negative log values**

We would like to respectfully clarify and disagree that: **Equation 9 is correct, and the term should be the sum of positive log value**. Complementary label refers to the class that a sample may not belong to [16, 17 in paper]. Equation 9 aims to **minimise the similarity (which is the positive log value)** between the embeddings of unlabelled testing samples and their complementary labels, denoted as $\bar{y}$. They are dissimilar, so that their similarity should be minimised.

However, if we sum the negative log values as you mentioned, it will correspond to minimising the distance between the testing embeddings and the embeddings of their complementary labels. **This is contradictory to the intended optimisation objective, which is to maximise the separation between them.**

> **Re W3: Proposition 1 notation:**

We apologise for not clearly stating the definitions of the variables in Proposition 1 in the main text. In the current submission, we have deferred the definitions of these notations to Appendix D (Lines 663–682). **Specifically, the notation refers to a hypothesis; denotes the conventional loss function trained on standard (true) labels, whereas denotes the complementary loss function trained on complementary labels**, i.e., labels indicating the classes to which the sample does not belong. We will incorporate the formal definitions of these notations into the main text in our next revision.

Intuitively, the complementary loss function **penalises the model when it assigns high similarity or confidence to the complementary class**, thereby encouraging the model to separate the predicted embedding from incorrect classes. Consequently, in proposition 1, **measures the model's ability to avoid predicting the complementary labels on unlabelled test data.**

> **Re W4&Q2: Proof of Proposition 1 in Equation 14:**

We apologise for the confusion caused by the proof of Proposition 1 in Equation (14). The notation measures the population risk (or excess risk) of the hypothesis evaluated using the complementary-labelled test-time loss function , over unlabelled testing samples. The goal of Equation (14) is to **establish an upper bound on the excess risk by leveraging the Rademacher complexity framework [37 in paper].** The use of Rademacher complexity for risk minimisation analysis is well-established and widely adopted in fields such as weakly supervised learning [1] and noisy label learning [2]. We will add the formal definitions of these notations in the text in our next revision.

In the proof of Proposition 1 in Equation 14, the definitions of each variable are:

1. : The empirical risk (computed as the average loss on the training samples).
2. : The true risk of a hypothesis , which is the expected loss over the true data distribution.
3. : The true risk of the learned model .
4. : The empirical risk of the learned model .
5. : The true risk of the optimal hypothesis that minimises the true risk.
6. : The empirical risk of the optimal hypothesis .

**Then, we provide the steps for the second line in Equation 14:**

1. Firstly, the second line **replace the minimum over all hypotheses (the last term) with the risk of the optimal hypothesis .**
2. Secondly, **because (the third term) is the empirical risk of the learned model , which is the global minimizer of the empirical risk, it holds that for all (including )**, , leading to the final inequality result in the second row.

**We greatly appreciate your concern and will explicitly include the definition of each variable in the revised version.**

> **Re W5: About Equation 3:**

Thank you for your insightful suggestion. here is a variable obtained together with . We agree with you that they can be removed from Equation 3, since has already been defined in line 145. We will delete them in the revised version.

> **Re Q1: Notation for and :**

We apologise for the lack of clarity regarding the formal definitions of and . Specifically, **is introduced in Line 154 to denote the initial layer-wise prediction, while is defined in Line 191 as the complementary label used in the test-time loss.**

Intuitively, (Line 154) corresponds to the class with the highest cosine similarity to the test sample embedding, whereas (Line 191) denotes the class with the lowest cosine similarity at the pivot layer with the lowest mean entropy.

We will revise the manuscript to ensure these notations are clearly and consistently defined in the main text.

> **Re Q3: Inductive setting and unlabelled graphs during test time:**

We would like to clarify that whether it's the **inductive/transductive setting** in graph training is **not related to the test-time prompt tuning**.

1. In **inductive/transductive setting**, they correspond to whether the model can see testing nodes during training.
2. In **test-time prompt tuning**, the model is trained already and deployed for testing. And the test-time prompt tuning task is to further **update the trained model when real testing nodes arrive**.

Therefore, it is inappropriate to classify the test-time prompt tuning into inductive/transductive because **they are not at the same stage.** And for test-time prompt tuning, test data will always be available when the model is being tested, **as the ultimate goal of the test-time method is to predict the test data.**

> **Thank you again, and if you have any further questions, we would be more than happy to discuss them with you.**

[1] Theoretical Comparisons of Positive-Unlabeled Learning against Positive-Negative Learning, NeurIPS 2016.

[2] Learning with Noisy Labels, NeurIPS 2013.

Add:

## Official Review of Submission1632 by Reviewer ha7m

Official Review by Reviewer ha7m 24 Jun 2025, 16:21 (modified: 24 Jul 2025, 21:31)
Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer ha7m Revisions

**Summary:**
In the paper, the authors proposed a new graph prompt tuning method named GFMate, which can be applied to various GFM methods without specific adaptation. The method trains two prompt vectors, one is the centroid prompt, which is used to adjust the center of each class; the other is the layer-wise importance prompt. The author also introduces the Layer-wise Entropy-based Augmentation and test-time graph complementary learning to augment the learning process. The method shows both efficiency and high accuracy, achieving SOTA results across different benchmarks.

**Strengths And Weaknesses:**

### Strengths

1. The experimental results are really promising, with solid ablation studies to verify the effectiveness of the method from various angles.
2. The proposed method is fairly general and can be applied to various GFM frameworks.

### Weaknesses

1. The writing of the paper can be improved significantly. The current flow and notation are confused and make it hard for the reader to understand the method at first glance.
2. Overall, I am feeling that the paper is more likely a combination of different tricks that work effectively empirically. However, there still lacks some deeper insight about why the method is proposed in this way and why it works really well. More specifically,

- What's the difference between the proposed centroid prompt and the existing graph prompt methods that do addition or multiplication on the feature/embedding side?
- What's the purpose of introducing a multi-layer ensemble in the prediction? I notice the author did an ablation study on the layer-wise importance prompt (). I am not sure how the author did that, but I am wondering if the author could add the ablation study with the following variance: (a) remove  but keep the multi-layer ensemble; (b) use only the last layer; (c) use the pivot layer.
- How does the layer-wise entropy-based augmentation differ from existing works that generate pseudo-labels with label propagation or other ways?
- How effective is the test-time graph in complementary learning? I didn't find the ablation study on that. Please remind me if I missed it.

3. The proposed method can only work on few-shot scenarios but cannot be applied to more challenging zero-shot settings.

**Quality:** 3: good
**Clarity:** 2: fair
**Significance:** 3: good
**Originality:** 3: good

**Questions:**

1. I can get the intuition behind Equation 9. But it looks like a negative loss. Do you want to minimize it? Further, Equation 11 sums over Equations 9 (negative) and 10 (positive), which is confusing. How is it actually implemented? Please correct me if I missed something.

**Limitations:**

1. The performance of the proposed method is significantly better than the baseline. However, no code is provided in the submission for validation or checking, which decreases the reliability.

**Rating:** 4: Borderline accept: Technically solid paper where reasons to accept outweigh reasons to reject, e.g., limited evaluation. Please use sparingly.
**Confidence:** 3: You are fairly confident in your assessment. It is possible that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work. Math/other details were not carefully checked.
**Ethical Concerns:** NO or VERY MINOR ethics concerns only
**Paper Formatting Concerns:**
Looks good to me.

**Code Of Conduct Acknowledgement:** Yes
**Responsible Reviewing Acknowledgement:** Yes

Add:

## Rebuttal by Authors Rebuttal by Authors

27 Jul 2025, 18:04 (modified: 01 Aug 2025, 04:30)
Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors Revisions

**Rebuttal:**
**We sincerely appreciate the time you've spent on our paper and the constructive questions you've provided.** We have thoroughly addressed your question and concerns. We summarise them and **highlight** the key information for improved clarity and readability as follows:

> **Re W1: Presentation issues**:

We apologise for the unclearness in the manuscript. We will further improve the writing and would greatly appreciate if you could kindly point out any specific parts that are unclear, and we will do our best to revise them accordingly.

> **Re W2: Combination of different tricks that work effectively empirically:**

We would like to respectfully clarify that all modules are clearly motivated **logically, theoretically and empirically**. They are not ad-hoc or simple tricks. We summarise our response to each point of the question in weakness 2 as follows:

> **Re W2 (1): The proposed centroid prompt VS. addition or multiplication prompt on the feature/embedding side:**

1. Our proposed **centroid prompt works in the (a) output space of (b) only the few-shot labelled samples.** Existing prompts operate on the **(c) input space or embedding space of (d) all input data samples.**
2. To obtain cross-domain generalisation, **existing GFM prompt methods**, such as SAMGPT, **require a separate prompt for each domain/dataset even since pretraining**. In contrast, our GFMate only needs a centroid prompt over the **label space of few-shot samples**. GFMate is significantly more lightweight and efficient.
3. From Point 2, **GFMate is pre-training-agnostic, with no assumptions about the pre-training** of the GFM. While **existing GFM prompt methods are tightly coupled with specific pre-training strategies**. For example, SAMGPT is entangled to GFMs by graph contrastive learning, and MDGPT is tied to link prediction pre-training. This decoupling of model and prompts makes **centroid prompt more generalisable across different GFMs**, as further validated by experiments in Section 5.4, where we demonstrate consistent effectiveness across GFMs pre-trained with diverse objectives.

> **Re W2 (2): Motivation for layer prompt:**

The design of the layer prompt is well motivated by the layer-wise performance of a fixed pre-trained GFM may vary across different unseen target domain graphs. We provided motivation in Figure 2 and logic descriptions between Lines 143 and 147.

Logically, in the **cross-domain** GFM setting, **a fixed pre-trained GFM may encounter previously unseen graph datasets with distinct characteristics and structural patterns** (such as homophilous versus heterophilous graphs), which may impact the neighbourhood aggregation process of a pre-trained GFMs. **These differences can naturally lead to variations in model performance across GFM layers, depending on the target dataset.**

To verify this hypothesis, we conduct an empirical study using a fixed pre-trained GFM (pre-trained on Arxiv and tested on different target datasets). As shown in Figure 2, the layer-wise performance varies significantly across different target datasets.

**This observation motivates the design of a learnable layer prompt**, which allows our method to better adapt to the target domain by learning the importance of each layer during test-time. **Therefore, the introduction of the layer prompt is both necessary and well motivated.**

In the **ablation study**, we evaluate a variant of GFMate without the layer prompt, which instead uses the **mean prediction across all layers. Following your suggestion, we provide further experimental results as shown in the complete ablation table in the answer to your next question (W2 (4)).** It is clear that the one-shot mean accuracy for GFMate using only the last layer is significantly lower than those of **GFMate with the pivot layer** and even lower than GFMate **without the layer prompt**. This highlights the effectiveness of the ensemble strategy and the importance of the learnt layer prompt in our framework.

> **Re W2 (3) Difference from pseudo-label methods:**

The proposed layer-wise entropy-based augmentation is **fundamentally different** from existing pseudo-label methods based on label propagation in the following aspects:

1. **Motivation Difference**. The key difference between our layer-wise entropy-based augmentation and pseudo-label propagation methods lies in our design tailored specifically for the GFM setting. As discussed in the previous question, **the layer-wise performance of a fixed pre-trained GFM model can vary significantly across different target domain datasets.** Therefore, directly using the last-layer embedding as previous method to compute similarity-based pseudo-labels for GFMs may lead to inaccurate predictions.

**Motivated to address this issue, we propose an entropy-based augmentation strategy that selects the most confident layer**, i.e., the one with the lowest prediction entropy, for each sample to generate the initial predictions. Based on these predictions, we then select the **top-k most confident unlabelled samples from the most confident layer** to augment the few-shot labelled set. The effectiveness of this augmentation is validated in our ablation study, which compares GFMate with and without the proposed augmentation, demonstrating its benefits in the GFM setting.

2. **Setting Difference**. Prior methods, such as PTA [1], are designed for semi-supervised learning on graphs, where the **full label set is available during training.** In contrast, GFMate is designed for a **more challenging few-shot setting, where only extremely limited labelled samples are provided.**
3. **Effectiveness**. We further compare GFMate with a pseudo-label-based prompt tuning baseline, which directly uses the last-layer predictions as pseudo labels and tunes the prompt on these pseudo-labelled testing samples. As shown in **Table 10 in the appendix**, GFMate consistently outperforms pseudo-label tuning methods across different ratios, demonstrating the effectiveness of the proposed method compared to conventional pseudo-label-based approaches.

> **Re W2 (4) Effectiveness of TGCL:**

Sorry for mistakenly missing this part in the main text of the submission. We conducted the ablation study on the test-time complementary learning by removing the proposed test-time loss from GFMate. We provide the complete ablation table below:

| Dataset | GFMate (Last Layer Only) | GFMate w/o Layer Prompt (Mean) | GFMate w/o TGCL | GFMate w/o Centroid Prompt | GFMate w/o Aug | GFMate (Full) |
|---|---|---|---|---|---|---|
| Texas | 52.08 ± 7.93 | 58.76 ± 8.42 | 69.47 ± 7.25 | 66.02 ± 5.59 | 75.60 ± 6.43 | **77.30 ± 7.37** |
| Cornell | 43.36 ± 9.25 | 48.87 ± 7.69 | 72.59 ± 7.68 | 64.94 ± 6.09 | 74.39 ± 8.84 | **79.66 ± 8.11** |
| Citeseer | 44.25 ± 13.37 | 45.25 ± 10.04 | 52.92 ± 10.34 | 41.28 ± 12.94 | 53.77 ± 13.08 | **55.27 ± 11.97** |
| Cora | 49.92 ± 8.92 | 58.53 ± 6.28 | 54.57 ± 4.31 | 48.96 ± 5.17 | 57.98 ± 8.72 | **59.58 ± 5.08** |

These results demonstrate the effectiveness of the proposed test-time complementary learning. We will include this full ablation result in the revised version.

> **Re W3: Zero-shot setting:**

Thank you for your insightful observation. As in prior GNN-based GFM works (e.g., GCOPE, MDGPT, SAMGPT, MDGFM), the **zero-shot setting is generally not applicable to GNN-based GFMs in this setting**. We fully acknowledge the advantage of LLM-based GFMs in supporting zero-shot scenarios, as discussed in our Related Work section.

However, **LLM-based GFMs are typically limited to graphs with textual attributes and cannot be applied to more general types of graphs** where node features lie in continuous, non-textual feature spaces. In contrast, GNN-based GFMs are capable of handling such general graphs. Therefore, GFMate is designed to enhance GNN-based GFMs, considering their broader applicability and generalisability to real-world graph data without textual information.

> **Re Q1: Clarification for Equation 9 and Equation 10**:

Yes, we indeed minimise the test-time loss and the few-shot loss by implementing a convex combination of the two losses (Equation 11). **Complementary label refers to the class that a sample may not belong to [16, 17 in paper].** The objective of the test-time complementary loss in Equation.9 is to **encourage separation** between the embeddings of unlabelled testing samples and their complementary labels . Minimising this loss **reduces the similarity between each sample and its corresponding complementary class as expected.**

It is important to note that the few-shot loss (Equation 10) and the test-time loss (Equation 9) are **independent and orthogonal**: Equation 9 is optimised on the **few-shot labelled set**, while Equation 10 is optimised on the **complementary-labelled testing set.** Jointly minimising both losses is neither contradictory nor problematic, as they are optimised on **disjoint data and serve complementary purposes.** This joint optimisation is also unrelated to the sign (positive or negative) of the loss values, as the optimisation direction is **consistent with our objective**: to encourage separation between testing samples and their complementary classes, while promoting similarity between few-shot labelled samples and their true classes.

We will provide a clearer explanation of the intuition behind the test-time learning process in the revised version.

> **Re Limitation: Code availability**:

Thank you for your concern. As the **anonymous link is prohibited at this stage**, we are unable to provide the code or demo code link, and we will release the **full code for reproducibility** upon the acceptance of the paper, as stated in the paper.

> **If you have any further questions, we would be more than happy to discuss them with you.** We would be deeply grateful if you could champion our paper.

[1] On the Equivalence of Decoupled Graph Convolution Network and Label Propagation. WWW 2021.

Add:

*Replying to Rebuttal by Authors*

## Mandatory Acknowledgement by Reviewer ha7m

Mandatory Acknowledgement by Reviewer ha7m 05 Aug 2025, 03:36
Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

**Mandatory Acknowledgement:** I have read the author rebuttal and considered all raised points., I have engaged in discussions and responded to authors., I have filled in the "Final Justification" text box and updated "Rating" accordingly (before Aug 13) that will become visible to authors once decisions are released., I understand that Area Chairs will be able to flag up Insufficient Reviews during the Reviewer-AC Discussions and shortly after to catch any irresponsible, insufficient or problematic behavior. Area Chairs will be also able to flag up during Metareview grossly irresponsible reviewers (including but not limited to possibly LLM-generated reviews)., I understand my Review and my conduct are subject to Responsible Reviewing initiative, including the desk rejection of my co-authored papers for grossly irresponsible behaviors. https://blog.neurips.cc/2025/05/02/responsible-reviewing-initiative-for-neurips-2025/

Add:

*Replying to Rebuttal by Authors*

## thanks

Official Comment by Reviewer ha7m 05 Aug 2025, 03:47
Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

**Comment:**
Thanks, authors, for the detailed response. Most of them make sense to me.

I am still quite confused about equation 9-11. Based on the authors' response, Equation 9 is indeed a negative loss and Equation 10 is a positive one. Please confirm with me about it. If it is true, that's really weird to me based on my personal understanding. First, the numerical instability of the negative loss as it close to infinity makes it (almost) impossible to work. Even if it can be jointly optimized with Equation 11, the negative loss will approach infinity while the positive loss approaches zero, and the gradient will go to the negative loss, which is not desired.

Add:

## Official Comment by Authors Official Comment by Authors

05 Aug 2025, 14:20 (modified: 05 Aug 2025, 14:27)
Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer ha7m
Revisions

**Comment:**
**Thank you for your response and for recognising our reply.**

We **confirm** that Equation 9 is negative and Equation 10 is positive.

We **fully agree** with your concern that a negative loss having **a possible negative infinity may hinder the stability** of the loss optimisation. This is **already considered** in the current framework that,

1. We introduce a **scaling factor** in Equation 11 to control the magnitude and balance the loss contribution during training of the two terms.
2. During the training, the term that pushes away the target node from the complementary label **will not make the log prob be optimised to close to negative infinity**. This is because (1) the optimisation term in Equation 10 will take turns to choose target few-shot labelled nodes and their ground truth label to **maintain the representation space semantics**; (2) and the optimisation objective in Equation 9 will also be changed along the model training to **choose different complementary labels** for test samples; (3) although the bound of Equation 10 is in negative infinite, the model cannot be optimised to this range, which is seen during our **training log that the loss in Equation 10 will converge in a small negative number** (e.g., around -6.45 for the Texas one-shot setting).

Yet we fully appreciate your suggestion. We have further **implemented and experimented with a bounded positive loss** for our test-time complementary loss in Equation 9. The loss is defined as (layer-wise computation and averaging across samples are omitted for brevity):

where  is the softmax probability between each complementary-labelled testing sample and its complementary class (i.e., the least similar class). Such loss is lower-bounded by 0 and is positive, with the **same objective: to minimise it in order to separate the test nodes from their complementary classes.** We provide the experimental results for each dataset using this positive loss below:

| Methods | Texas | Chameleon | Squirrel | Arxiv-year | Cora | Citeseer | Photo |
|---|---|---|---|---|---|---|---|
| GCOPE | 64.76 ± 14.84 | 30.58 ± 7.44 | 22.16 ± 5.77 | 17.98 ± 5.51 | 39.06 ± 12.52 | 42.26 ± 14.19 | 55.69 ± 4.68 |
| MDGPT | 59.76 ± 12.44 | 28.04 ± 4.28 | 24.41 ± 7.01 | OOM | 44.52 ± 11.39 | 41.98 ± 12.24 | 54.96 ± 10.25 |
| SAMGPT | 66.79 ± 10.77 | 38.12 ± 8.90 | 25.75 ± 6.29 | OOM | 52.83 ± 12.04 | 47.76 ± 10.55 | 56.33 ± 9.04 |
| MDGFM | -- | 28.36 ± 3.65 | 24.30 ± 3.26 | -- | 44.83 ± 7.41 | 42.18 ± 6.41 | -- |
| RiemannGFM | 58.60 ± 15.27 | 29.68 ± 9.95 | 20.13 ± 8.58 | OOM | 37.91 ± 16.13 | 38.02 ± 9.58 | 49.69 ± 13.32 |
| **GFMate-Eq9Pos** | **77.30 ± 7.37** | **47.35 ± 6.15** | **26.89 ± 6.42** | **30.24 ± 3.06** | **59.58 ± 5.08** | **55.27 ± 11.97** | **58.85 ± 2.18** |
| **GFMate-Eq9Neg** | **76.63 ± 7.81** | **47.25 ± 6.11** | **27.02 ± 6.22** | **30.19 ± 3.65** | **59.68 ± 5.37** | **56.25 ± 13.33** | **58.85 ± 2.17** |

It is evident that GFMate, utilising both the original negative test loss and the newly implemented positive test loss, achieves comparable and consistent performance, significantly outperforming existing GFM methods. **We will update the test loss formulation in the revised version.**

**Thank you again for your valuable suggestion and insightful feedback.**

Add:

## Official Review of Submission1632 by Reviewer Yv5F

Official Review by Reviewer Yv5F 11 Jun 2025, 22:02 (modified: 24 Jul 2025, 21:31)
Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer Yv5F Revisions

**Summary:**
The paper proposes GFMate, a novel framework for enhancing Graph Foundation Models (GFMs) through pre-training-agnostic test-time prompt tuning. Traditional GFMs often suffer from entanglement with domain-specific prompts during pre-training, limiting their generalizability. GFMate introduces centroid and layer prompts only at test-time, avoiding this entanglement and improving adaptability across tasks. Additionally, GFMate proposes a complementary learning objective that leverages both labeled and unlabeled data to mitigate train-test distribution shifts. Extensive experiments on 12 benchmark datasets demonstrate notable performance improvements of up to 30.63%.

**Strengths And Weaknesses:**
Novel Approach: GFMate offers a fresh perspective by focusing on test-time prompt tuning rather than pre-training entangled prompts, enhancing adaptability and generalization.

Efficient Utilization of Test Data: The complementary learning objective efficiently incorporates unlabeled test nodes, addressing distribution shifts and improving prediction accuracy.

Significant Performance Gains: The paper evidences strong empirical results with improvements up to 30.63% across diverse datasets.

Plug-and-play Design: As GFMate is pre-training-agnostic, it can be easily integrated into various existing GFMs with minimal adjustments.

Limited Scope: The method is specifically designed for GNN-based GFMs and may not be applicable to LLM-based GFMs, reducing its universality across different backbone architectures.

Complexity in Implementation: While providing considerable efficiency gains, the setup for entropy-based augmentation and complementary learning objectives may introduce complexity.

Further Validation Needed: Although the paper provides extensive experiments, more validation on real-world domains or large-scale datasets could strengthen claims of generalizability.

**Quality:** 2: fair
**Clarity:** 2: fair
**Significance:** 1: poor
**Originality:** 2: fair
**Questions:**
Can the framework be adapted or extended to LLM-based GFMs, or is it strictly limited to GNN-based architectures?

How does GFMate perform with datasets that include both graph-structured data and textual attributes?

What are the computational overheads associated with entropy-based augmentation and the complementary learning process?

**Limitations:**
Applicability to Real-world Scenarios: The framework needs to be validated beyond academic datasets to prove its effectiveness in industry settings where data variability and noise are prevalent.

Scalability Concerns: The efficiency gain claims need more examination regarding scalability to much larger datasets or real-time applications.

**Rating:** 3: Borderline reject: Technically solid paper where reasons to reject, e.g., limited evaluation, outweigh reasons to accept, e.g., good evaluation. Please use sparingly.
**Confidence:** 3: You are fairly confident in your assessment. It is possible that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work. Math/other details were not carefully checked.
**Ethical Concerns:** NO or VERY MINOR ethics concerns only
**Paper Formatting Concerns:**
None

**Code Of Conduct Acknowledgement:** Yes
**Responsible Reviewing Acknowledgement:** Yes

Add:

## Confidential Request to Investigate Possible LLM-Generated Review from Reviewer Yv5F

Author AC Confidential Comment by Authors
27 Jul 2025, 18:17 (modified: 01 Aug 2025, 09:54)
Program Chairs, Senior Area Chairs, Area Chairs, Authors Revisions

**Comment:**
**Dear PC, AC and SAC,**

We respectfully bring to your attention on our serious concerns regarding the authenticity of the review submitted by Reviewer Yv5F, which we believe exhibits **strong indications of being purely generated by a large language model (LLM). We outline the supporting evidence below and kindly rEquationuest further investigations in accordance with NeurIPS ethical reviewing guidelines.**

**1. Striking similarity to the paper's own content**

The review's summary, strengths, and weaknesses closely mirror the manuscript's abstract and limitation section. The summary appears to paraphrase our abstract, while the key limitation, that the method only applies to GNN-based GFMs, **is directly quoted from our own "Limitations" section, with no independent analysis or reviewer insight.**

**2. Inconsistency between review content and scores**

Despite the generally positive tone throughout the review, the assigned scores are unusually low, particularly for Significance and Overall Recommendation. **No concrete flaws or technical concerns are cited to justify such a low assessment. This inconsistency may point to formulaic scoring detached from actual evaluation, which further raises concerns.**

**3. Verification via multiple LLMs**

We conducted independent verification using publicly accessible LLMs. In each case, the generated review **closely resembles** Reviewer Yv5F's submission in both structure and content. **(We obey the rules of not using external links/PDFs in rebuttal. But in this non-rebuttal and confidential comment, we provide external evidence. If it is inappropriate, we will remove them.)**

- **ChatGPT**: The generated review follows **the same structure as Reviewer Yv5F's: novelty, efficiency, followed by limitations.** Notably, the limitation regarding GNN-only applicability is reproduced verbatim, despite being explicitly stated in our own manuscript in the limitation section. The full transcript is available at: https://chatgpt.com/s/t_68849589a324819182ecc7db1e5fc248
- **Gemini**: Prompted to review the same paper, Gemini independently identified the same limitation (method applies only to GNN-based GFMs), again suggesting a direct reflection of the content already stated in the paper rather than any novel critique. The proof can be viewed at https://g.co/gemini/share/33f52dd7b579
- **DeepSeek**: The generated review highlights the same two points raised by Reviewer Yv5F, nearly verbatim:
- "Scope: The method is limited to GNN-based GFMs..."
- "Computational Overhead..."

These align almost exactly with the reviewer's:

- "Limited Scope: The method is specifically designed for GNN-based GFMs..."
- "Complexity in Implementation..."

**4. Publicly accessible anonymised verification**

To ensure transparency, we have summarised all supporting materials—including prompts and LLM responses—into an anonymous GitHub repository: https://anonymous.4open.science/r/Review-5955

**Given the substantial overlap in structure, phrasing, and content between Reviewer Yv5F's comments and the responses generated by multiple LLMs, we believe there is a high likelihood that the review may not reflect an original, human-authored assessment. Such reviews may undermine the peer review process and compromise fairness across submissions.**

**We fully recognise the need to be cautious in making such claims and raise this matter with the utmost respect for the integrity of the review process. We kindly ask the PC, AC and SAC to review this case and consider an appropriate course of action in line with NeurIPS reviewing standards.**

Thank you very much for your time and attention to this matter.

Sincerely,

Authors.

Add:

## Rebuttal by Authors Rebuttal by Authors

27 Jul 2025, 19:20 (modified: 01 Aug 2025, 04:30)
Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors Revisions

**Rebuttal:**

**Thank you for your question. We have thoroughly addressed your concerns and highlighted the key information for improved clarity and readability as follows:**

**Re W1&Q1 GNN-based GFMs:**

As stated in Section 2, GFMate specifically targets GNN-based GFMs, as they can be applied to **a broader range of graphs beyond those with textual attributes.** In contrast, LLM-based GFMs are restricted to text-attributed graphs. As a result, **GNN-based GFMs offer better generalisability** to diverse graph data and thus have broader application potential. **This motivates our focus on GNN-based GFMs in the design of GFMate.**

**Re Q2 For text-attributed graphs:**

Same as our response to Q1, GFMate follows the setting of GNN-based GFMs, which are designed to operate on general types of graphs without requiring textual attributes.

Importantly, **GFMate has already been evaluated on text-attributed graphs, as you mentioned**, such as Cora, Citeseer, and Arxiv, where the textual information has been pre-processed (e.g., using bag-of-words or other encoding techniques) to transform it into the feature space, consistent with the data preprocessing technique in our baseline methods. **Therefore, GFMate is capable of handling graphs with both textual and structural information.**

**Re W2&Q3 Computational overhead**:

We already provide the detailed **complexity analysis** and **experiment on inference time/space utilisation** for the entropy-based augmentation and the complementary learning process in Section 4.7, Section 5.2 and Appendix E. The complexity of augmentation is , where is the number of nodes, is the number of edges, is the number of GNN layers, is the number of classes, and is the hidden dimension. The first two terms account for the -layer message passing, and the last term arises from computing layer-wise similarities and entropies for all nodes and classes. TGCL and ensemble prediction each have a complexity of , as they compute similarities across classes for nodes at layers using -dimensional embeddings. **

We have also provided a **detailed analysis of the computational cost in Section 5.2 and Appendix E.** The experimental results demonstrate that GFMate is highly efficient in terms of both convergence time and GPU memory usage. Specifically, the maximum improvements in time and memory efficiency reach 98.24% and 97.18%, respectively, compared to the state-of-the-art method, SAMGPT. With such significant gains in efficiency, GFMate stands out as the most effective approach among existing GFM-based methods.

**Re Limitation 1: Noisy case:**

We **already provide the robustness analysis** to examine the effectiveness of GFMate **under noisy conditions** during both the pre-training and test-time stages of GFMs, as detailed in **Section E.3 of the appendix. The results further verify that GFMate maintains its effectiveness on test-time noisy graphs and can effectively address domain shift, i.e., varying pre-training and testing dataset domains, in the GFM setting.**

**Re Limitation 2: Scalability:**

**GFMate demonstrates strong scalability when applied to datasets of varying sizes.** As shown in Tables 4 and 5 in the appendix. Moreover, GFMate demonstrates superior space efficiency, as reported in Section 5.2 and Appendix E.1. This advantage arises from our lightweight design: the prompts in GFMate are integrated only with class centroids and multi-layer predictions, rather than with every input node or domain. In contrast, existing methods often require learnable prompts for each source domain and each target node [56, 53], resulting in increased memory usage.

**Furthermore, the substantially reduced number of tunable parameters in GFMate contributes to lower memory utilisation.** The maximum improvements in memory efficiency reach 97.18%, compared to the state-of-the-art method, SAMGPT. **With such notable efficiency gains, GFMate emerges as the most space-efficient method among existing GFM-based approaches and demonstrates great scalability.**

**If you have any further questions, we would be more than happy to discuss them with you.**

Add:

*Replying to Rebuttal by Authors*

## Reply to Authors

Official Comment by Reviewer Yv5F 05 Aug 2025, 17:46
Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

**Comment:**

I have read the rebuttal by the authors. However, I still have one question: does the model proposed have a theoretical promise?

Add:

*Replying to Reply to Authors*

## Official Comment by Authors

Official Comment by Authors
05 Aug 2025, 19:40 Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

**Comment:**

Thank you for your question.

**We have already provided a theoretical promise for the proposed test-time graph complementary learning method in Proposition 1 (page 6, line 201)** with the detailed proof given in Appendix D. This analysis establishes an excess risk bound for the proposed complementary loss, thereby guaranteeing the effectiveness of our method.

If you have any questions, please feel free to let us know.

Add:

*Replying to Confidential Request to Investigate Possible LLM-Generated Review from Reviewer Yv5F*

## Author AC Confidential Comment by Area Chair p1E9

Author AC Confidential Comment by Area Chair p1E9 05 Aug 2025, 22:48
Program Chairs, Senior Area Chairs, Area Chairs, Authors

**Comment:**

Thank you for the information. We will take all factors into consideration.

Add: