# Learn-To-Learn on Arbitrary Textual Conditioning: A Hypernetwork-Driven Meta-Gated LLM

Luo Ji, Qi Qin, Ningyuan Xi, Teng Chen, Qingqing Gu, Hongyan Li

**Geely AI Lab**
(Luo.Ji1@geely.com)

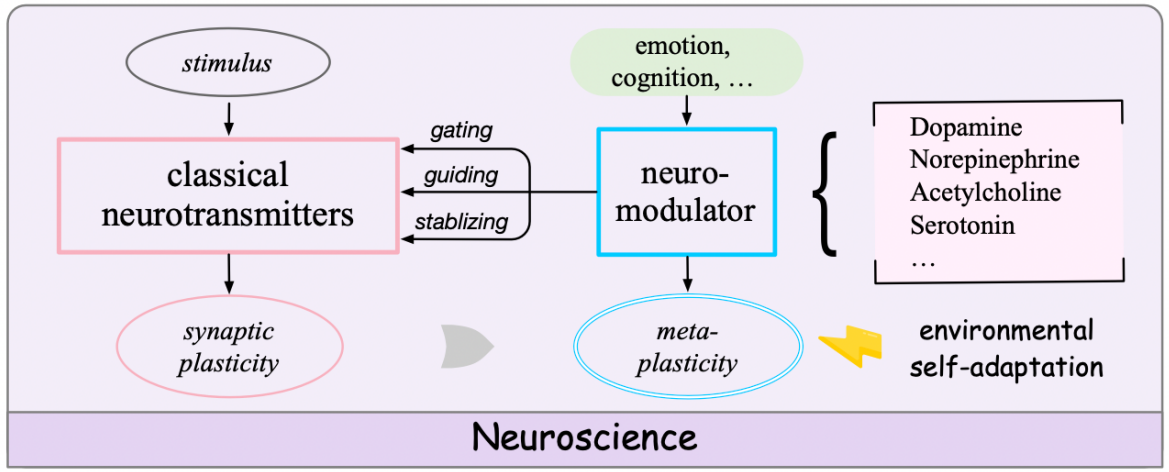
GEELY

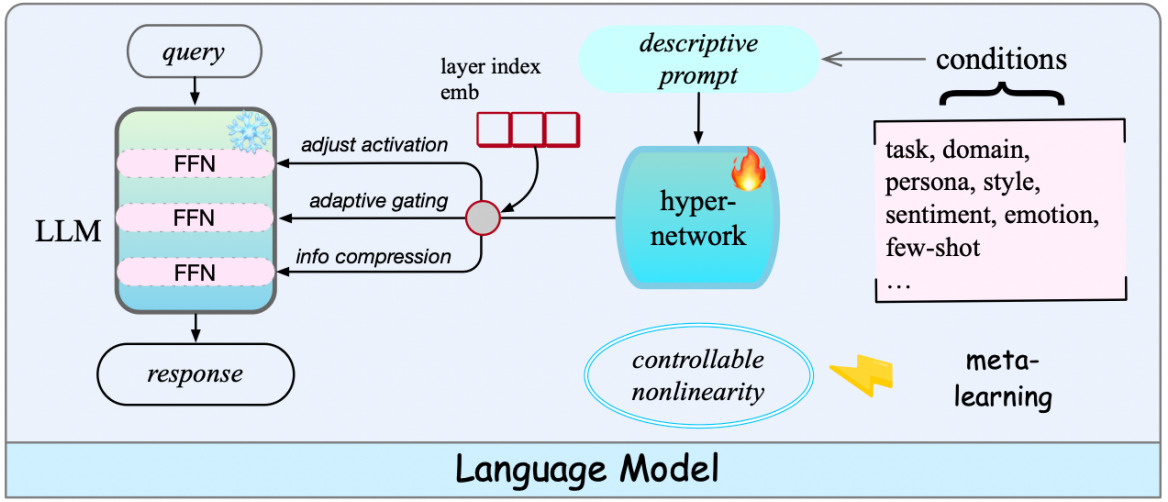
https://github.com/AaronJi/MeGan

ICML
International Conference
On Machine Learning

# Motivations

- Meta-learning (`*learn-to-learn*') requires AI to adapt its capability across subtle changes in conditions
- LLMs remain relatively deficient in the seamless, dynamic adaptation, e.g., `*catastrophic forgetting*' or `*alignment tax*'
- Human nervous system has a two-level architecture:
  - Neurotransmitters => Synaptic Plasticity
  - Neuromodulators => Meta-plasticity
  - Adaptive gain control without the metabolic cost of physical rewiring
- Analogously, we want to have:
  - Backbone parameters => general capability
  - Meta-parameters => meta-capability
  - Use a hypernetwork to adaptively adjust meta-parameters, without updating backbone parameters
- Potential benefits:
  - Explicit text-based meta-signals (input hypernetwork)
  - Reduce general capability degradation when adapting to specific tasks
  - Low training cost (small hypernetwork)



# Different Meta-Paradigms

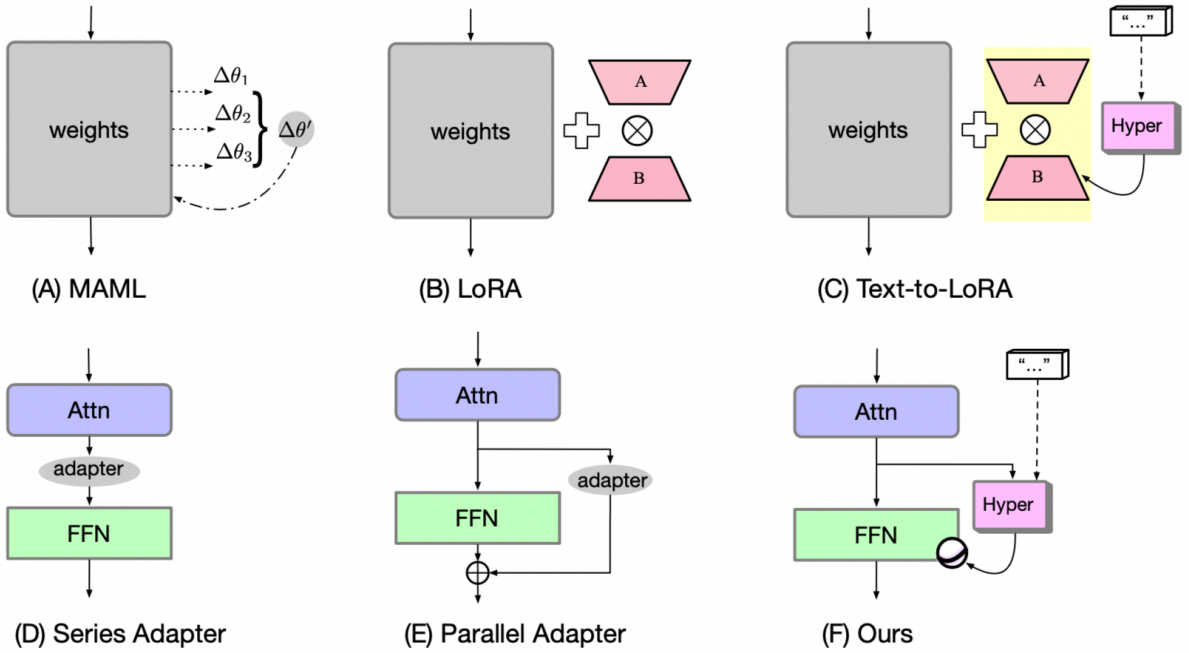



*Table 1.* Methodology summaries of the baselines and MeGan. *: dynamic low-rank reparameterization, which is similar to LoRA.

| Method | Structure customization | Meta tasks adaptation | Meta tasks update | Target tasks adaptation | Target tasks adjustment | Original param kept |
|---|---|---|---|---|---|---|
| LoRA (Hu et al., 2022) | ✗ | ✗ | LoRA | zero-shot | ✗ | ✓ |
| SFT | ✗ | ✗ | FT | zero-shot | ✗ | ✗ |
| CMAML (Song et al., 2020) | pruning | gradient | meta-update | zero-shot | ✗ | ✗ |
| ***Meta-in-context*** | | | | | | |
| MetaICL (Min et al., 2022) | ✗ | in-context | FT | few-shot | ✗ | ✗ |
| meta-icl (Coda-Forno et al., 2023) | ✗ | in-context | ✗ | few-shot | ✗ | ✓ |
| MAML-en-LLM(Sinha et al., 2024) | ✗ | in-context | meta-update | few-shot | ✗ | ✗ |
| ***Meta-on-LoRA*** | | | | | | |
| MLtD (Hou et al., 2022b) | struc. control | PEFT* | meta-update | zero-shot | PEFT* | ✗ |
| HydraLoRA (Tian et al., 2024) | ✗ | asym. LoRA | MoE routing | merged experts | LoRA | ✓ |
| Meta-LoRA (Li et al., 2025) | ✗ | grad. similarity | reweighting | zero-shot | LoRA | ✓ |
| Text-to-LoRA (Charakorn et al., 2025) | hypernetwork | task-wise text | FT / Recon | task-wise text | LoRA | ✓ |
| SHINE (Liu et al., 2026) | hypernetwork | in-context | PT / FT | in-context | LoRA | ✓ |
| ***Meta-on-Gating (Ours)*** | | | | | | |
| MeGan | hypernetwork | sample-wise text | FT | sample-wise text | activation | ✓ |

- Most LLM-based attempts feed the MAML framework with in-context examples.
- Text-to-LoRA [1] employs a hypernetwork which adapts LoRA matrices to task descriptions
- Adapter methods [2]: series VS parallels
- Our framework also leverages the hypernetwork, parallel to the backbone; yet:
  - The hypernetwork adjusts the FFN nonlinearity
  - The input of hypernetwork is generalized text (*i.e.*, conditions)


[1] Charakorn, et al. Text-to-LoRA: Instant transformer adaption. ICML 2025
[2] Hu, et al. LLM-adapters: An adapter family for parameter-efficient fine-tuning of large language models. ACL 2023

# Convert Self-Gating to Meta-Gating

| | |
|---|---|
| Beta-Sigmoid | $\sigma_\beta(x) := 1/(1+\exp(-\beta x))$ |
| Beta-Swish | $\text{Swish}_\beta(x) = x * \sigma_\beta(x)$ |
| $\text{Swish}_1$ (SiLU) | $\text{SiLU}(x) := \text{Swish}_1(x) = x * \sigma(x)$ |
| SwiGLU | $y = W_{\text{down}}\left(\text{Swish}_1\left(W_{\text{gate}}x\right) \otimes \left(W_{\text{up}}x\right)\right)$<br>$W_{\text{up}}, W_{\text{gate}} \in \mathcal{R}^{D\times C}, W_{\text{down}} \in \mathcal{R}^{C\times D}, C > D$ |

- **Self-Gating** mechanism to reweight the MLP based on SiLU
- Generally adopted in modern LLMs (Llama, Qwen, Mistral, …)

# Convert Self-Gating to Meta-Gating

| | |
|---|---|
| Beta-Sigmoid | $\sigma_\beta(x) := 1/(1+\exp(-\beta x))$ |
| Beta-Swish | $\text{Swish}_\beta(x) = x * \sigma_\beta(x)$ |
| $\text{Swish}_1$ (SiLU) | $\text{SiLU}(x) := \text{Swish}_1(x) = x * \sigma(x)$ |
| SwiGLU | $y = W_{\text{down}}\left(\text{Swish}_1\left(W_{\text{gate}}x\right) \otimes \left(W_{\text{up}}x\right)\right)$<br>$W_{\text{up}}, W_{\text{gate}} \in \mathcal{R}^{D\times C}, W_{\text{down}} \in \mathcal{R}^{C\times D}, C > D$ |

*allow beta deviated from 1*

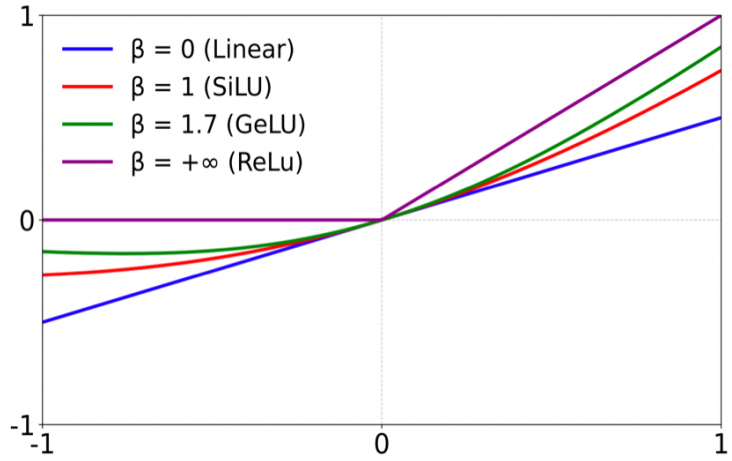
| | |
|---|---|
| Beta-SwiGLU | $y = W_{\text{down}}\left(\text{Swish}_\beta\left(W_{\text{gate}}x\right) \otimes \left(W_{\text{up}}x\right)\right), \beta > 0$ |

- **Self-Gating** mechanism to reweight the MLP based on SiLU
- Generally adopted in modern LLMs (Llama, Qwen, Mistral, …)

- Allow beta to deviate from 1 during post-training

# Convert Self-Gating to Meta-Gating

| | |
|---|---|
| Beta-Sigmoid | $\sigma_\beta(x) := 1/(1+\exp(-\beta x))$ |
| Beta-Swish | $\text{Swish}_\beta(x) = x * \sigma_\beta(x)$ |
| $\text{Swish}_1$ (SiLU) | $\text{SiLU}(x) := \text{Swish}_1(x) = x * \sigma(x)$ |
| SwiGLU | $y = W_{\text{down}}\left(\text{Swish}_1\left(W_{\text{gate}}x\right) \otimes \left(W_{\text{up}}x\right)\right)$<br>$W_{\text{up}}, W_{\text{gate}} \in \mathcal{R}^{D\times C}, W_{\text{down}} \in \mathcal{R}^{C\times D}, C > D$ |




*allow beta deviated from 1*

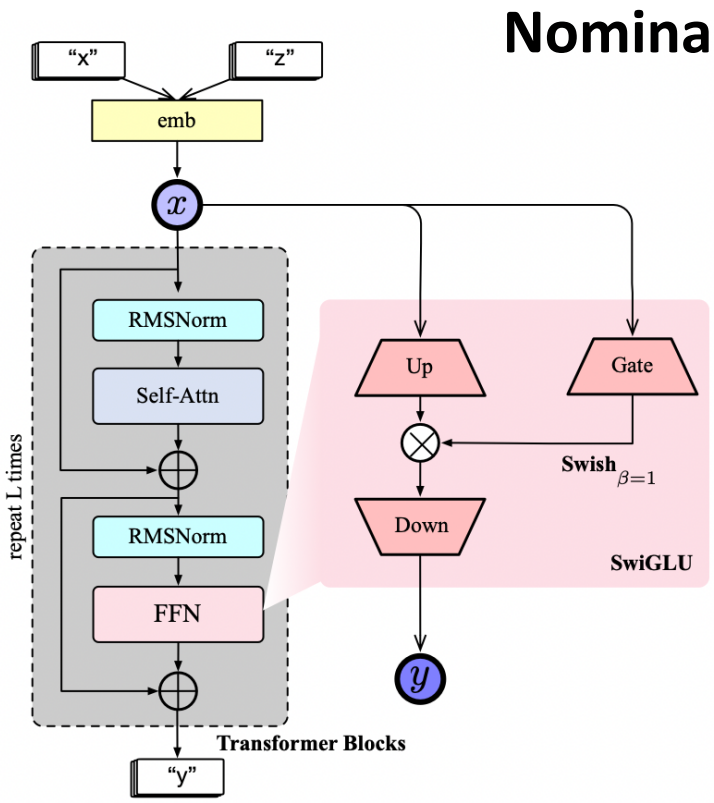
*adjust nonlinearity of*
- *Gating*
- *FFN*
- *LLM*

| | |
|---|---|
| Beta-SwiGLU | $y = W_{\text{down}}\left(\text{Swish}_\beta\left(W_{\text{gate}}x\right) \otimes \left(W_{\text{up}}x\right)\right), \beta > 0$ |

- **Self-Gating** mechanism to reweight the MLP based on SiLU
- Generally adopted in modern LLMs (Llama, Qwen, Mistral, …)

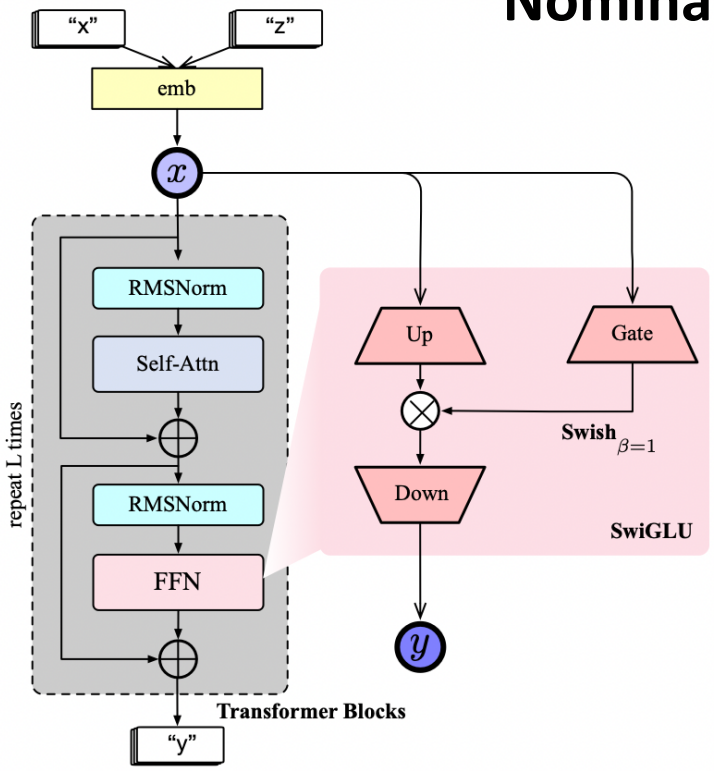
- Allow beta to deviate from 1 during post-training
- **Meta-Gating** that adaptively controls LLM nonlinearity

# Model Architecture



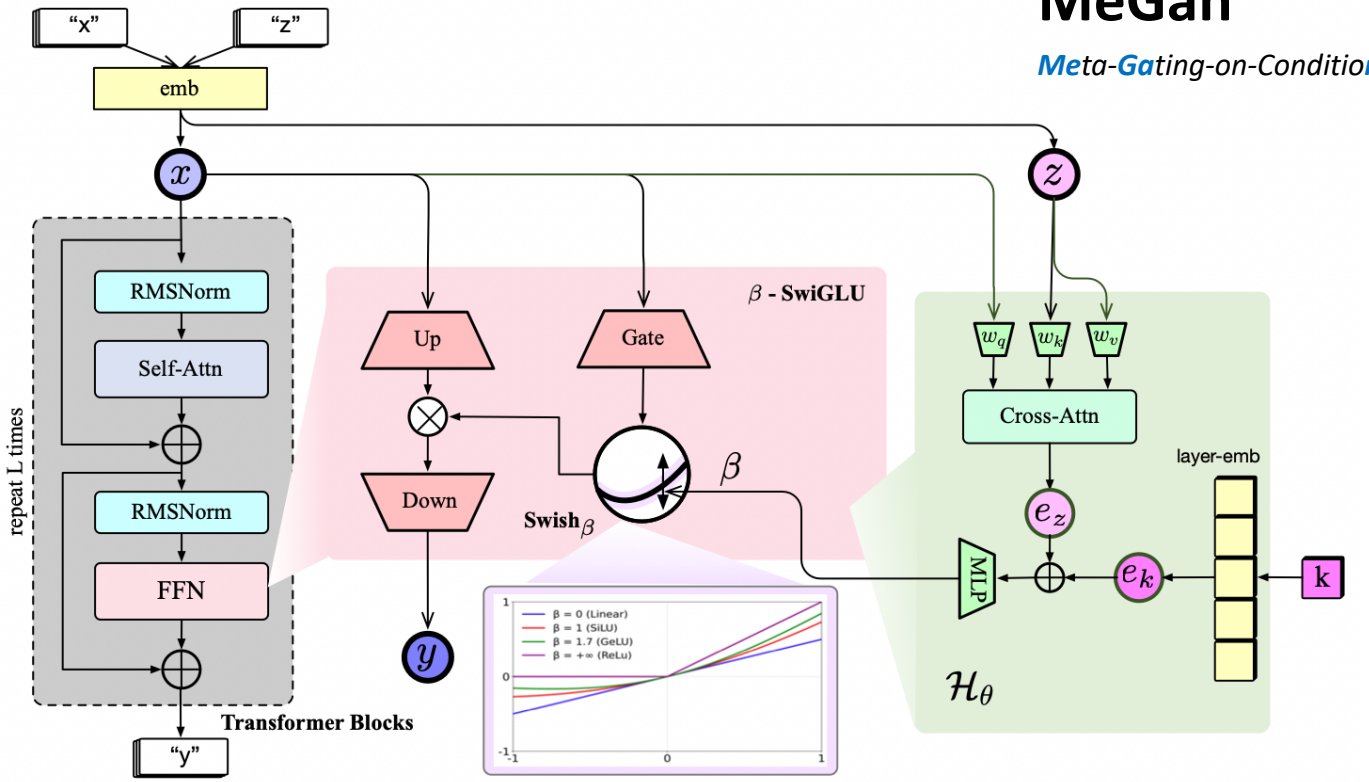
# Model Architecture

## Nominal LLM




## MeGan

*Meta-Gating-on-Condition*

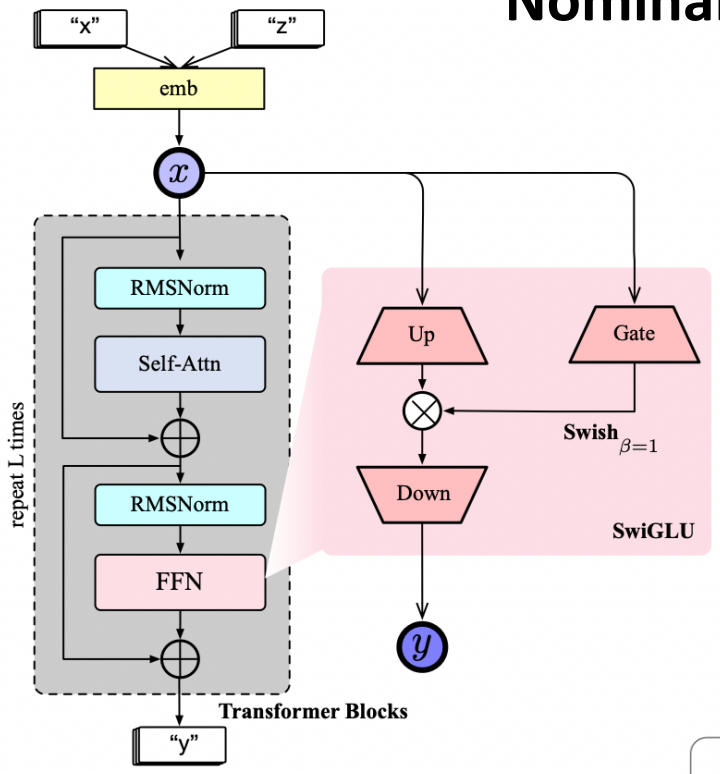



### Hypernetwork (H)

1) accept textual input (z), encoded by LLM's own emb layer
2) cross-attention between emb(z) and the block input (x)
3) add the layer index emb (layer-wise difference)
4) use an MLP to produce beta

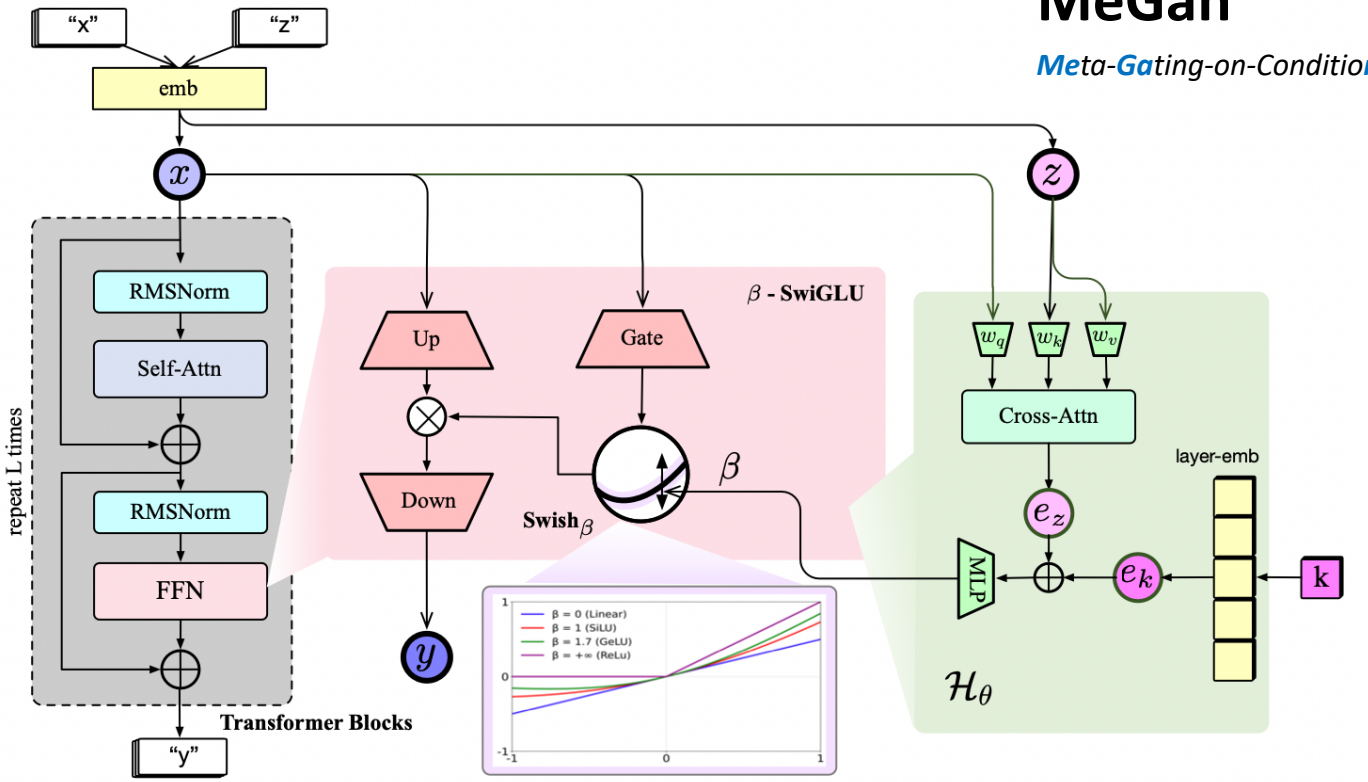
# Model Architecture

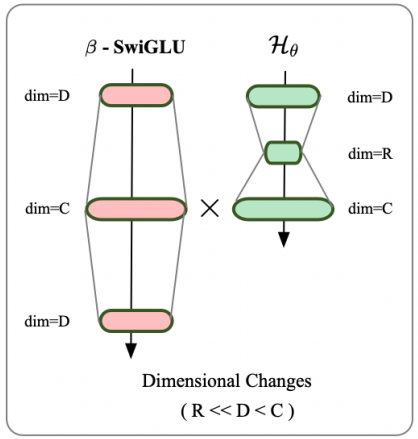
## Nominal LLM




## MeGan

*Meta-Gating-on-Condition*




## Dimension Compression

- SwiGLU expands the input dim (**D**) to a larger dim (**C**)
- H first compresses **D** to **R**, then operates on **C**
- => Information bottleneck

## Hypernetwork (H)

1) accept textual input (z), encoded by LLM's own emb layer
2) cross-attention between emb(z) and the block input (x)
3) add the layer index emb (layer-wise difference)
4) use an MLP to produce beta

# Conditions and Experimental Settings

*Table 2.* Summaries of the meta-conditions $z$, including task, domain, persona, style, sentiment, etc. We provide a generalized framework that self-adapts to all the conditions above. Types of tasks are marked in *Italic*. More prompts are in Appendix B.2.

| Type | Attribute | Prompt (example) | Related Datasets |
|---|---|---|---|
| task | *summarization/QA/classification/...* | (depending on detailed task; see Appendix B.1) | CrossFit&UnifiedQA , SNI |
| domain | debate/science/email/... | Please provide the summarization on the domain of {domain}. | AdaptSum (*summarization*) |
| persona | ["I like to remodel homes." " My favorite holiday is Halloween."] | Please provide the response with the knowledge of {persona}. | Persona-Chat (*dialogue*) |
| style | formal/informal moral/immoral | Please provide the response with the style of {style}. | GYAFC (*dialogue*) MIC (*QA*) |
| sentiment | positive/neutral/negative | Please provide the response with the sentiment of {sentiment}. | SST, Amazon, IMDB (*review*) |
| emotion | joy/sad/anger/... | Please provide the response with the emotion of {emotion}. | EmoryNLP, DailyDialog (*dialogue*) |

- Our Hypernetwork (H) accepts varied types of textual signals
  - *task*
  - *domain*
  - *persona*
  - *style*
  - *sentiment*
  - *emotion*
  - *…*
- Data format: QA (x, y) => labeled QA (x, y, z)
- To differentiate from `in-context', we call this meta-signal (z) by `*condition*'

*Table 3.* Statistics of meta-training datasets used in this paper. 'Cond.' denotes condition. 'LR' and 'US' indicate low resource and unseen. 'per.' denotes persona; 'senti.' denotes sentiment.

| | Meta-train | | Target | | |
|---|---|---|---|---|---|
| Dataset | Cond. | # num | setting | Cond. | # num |
| Persona-Chat | 1,137 per. | 8.9K | US | 100 per. | 0.97K |
| AdaptSum | 6 domain | 1.5K | LR | 6 domain | 8.0K |
| GYAFC | 2 style | 16.0K | LR | 2 style | 1.6K |
| MIC | 2 style | 254K | LR | 2 style | 10.9K |
| SST | 3 senti. | 9.1K | LR | 5 senti. | 1.0K |
| CrossFit& UnifiedQA | 325 task | 3.5M | LR<br>US | 80 task<br>11 task | 187K<br>4.9K |
| SNI | 479 task | 1.4M | US | 10 task | 23.3K |

# Instructions of conditions

*We use simple human-written instructions for no-task conditions.*

*Table 9.* Instructions for varied datasets.

| Dataset | Instructions |
|---|---|
| Persona-Chat | You are engaged in a conversation with the user. The user have the following profiles. You should consider the profiles and make the corresponding appropriate response.<br>profiles: **{profile}** |
| | You are chatting with a user, who has the following profiles. Consider the profiles and respond.<br>profiles: **{profile}** |
| | Provide the appropriate response based on the user profiles.<br>profiles: **{profile}** |
| AdaptSum | You are dealing with an abstractive summarization with different domains. You should consider the following domain tag, and adjust your summarization accordingly.<br>domain: **{domain}** |
| | You are dealing with an abstractive summarization with different domains. Adjust your summarization accordingly.<br>domain: **{domain}** |
| | Conduct the summarization based on its specific domains.<br>domain: **{domain}** |
| SST | Please answer the question with the sentiment of **{sentiment}**. |
| | Please answer the question with following sentiment.<br>sentiment: **{sentiment}** |
| | Reply with sentiment: **{sentiment}** |
| GYAFC / MIC | Please answer the question with the style of **{style}**. |
| | Please answer the question with following style.<br>style: **{style}** |
| | Reply with style: **{style}** |

*We use GPT-4o to generate task descriptions for task conditions.*

*Table 10.* Examples of generated task descriptions.

| Task | Descriptions |
|---|---|
| PIQA | You will explore practical questions and select an answer that presents a logical and widely accepted approach to solve a given problem or complete a task successfully. |
| | Analyze the provided scenarios where practical advice or solutions are required, focusing on selecting the most commonly used or convenient method. |
| | Given a question related to common tasks, your responsibility is to discern which proposed solution aligns with typical practices or makes the task easier to achieve. |
| Hellaswag | This task revolves around completing an unfinished text by selecting an ending that matches its tone and context. It requires you to think critically about how narratives develop and conclude effectively. |
| | This task asks you to select a suitable conclusion for an unfinished narrative or instructional content. It tests your comprehension and reasoning skills as you assess how well each option aligns with the given text. |
| | Your task involves completing an incomplete passage by selecting the ending that logically continues the context provided. This requires reading comprehension and the ability to infer meaning from a text. |
| OpenbookQA | Analyze the provided statements carefully and determine which one best fits into the context of the passage. This requires comprehension skills and the ability to make logical inferences. |
| | Consider each option in relation to what is presented in the input. Discern which one logically completes or responds accurately to the notion being expressed. |
| | Here, you'll be presented with different statements, and your role is to decide which one appropriately complements or responds to a scenario. This process involves critical analysis and synthesis of information. |

# Experiment Results

*Table 4.* Results conditioned on persona (Persona-Chat), domain (AdaptSum), style (GYAFC and MIC), and sentiment (SST). The best result is marked **bolded** and the second-best is underlined.

| Method | Persona-Chat | | AdaptSum | | GYAFC | | | MIC | | | SST | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | R-L | B-2 | R-L | B-2 | R-L | B-2 | D-2 | R-L | B-2 | D-2 | R-L | B-2 | D-2 |
| Direct | 11.79 | 3.64 | 13.35 | 4.72 | 7.57 | 2.01 | 0.18 | 7.36 | 1.92 | 0.04 | 11.79 | 3.64 | 0.06 |
| ICL | 13.75 | 5.23 | 23.35 | 11.29 | 6.77 | 1.77 | 0.10 | 7.42 | 1.87 | 0.04 | 12.47 | 4.48 | 0.15 |
| LoRA | 21.35 | 9.63 | 13.37 | 4.75 | 28.11 | 12.04 | 0.55 | 16.19 | 4.03 | 0.06 | 13.31 | 4.50 | 0.28 |
| SFT | 22.97 | 9.94 | 20.99 | 8.80 | 23.72 | 8.62 | 0.57 | 15.20 | 4.55 | 0.02 | 12.77 | 4.31 | 0.16 |
| meta-icl | 15.74 | 6.73 | 13.03 | 4.37 | 13.55 | 5.05 | 0.26 | 15.71 | 5.42 | 0.07 | 13.41 | 4.61 | 0.21 |
| Text-to-LoRA | 21.75 | **13.70** | 20.02 | 7.98 | 22.95 | 8.87 | 0.72 | 22.50 | 9.88 | **0.42** | 10.72 | 3.84 | 0.55 |
| **MeGan** (ours) | **23.15** | 10.15 | **41.85** | **26.88** | **29.37** | **13.82** | **0.81** | **23.67** | **10.57** | 0.17 | **22.76** | **9.33** | **0.67** |

Adaptation to the conditions of

- *persona* (Persona-Chat)
- *domain* (AdaptSum)
- *style* (GYAFC, MIC)
- *sentiment* (SST)

Adaptation to the conditions of

- *task types* (CrossFit&UnifiedQA)
- *transfer to low-resources and unseen tasks*

*Table 5.* Results on CrossFit&UnifiedQA. Two numbers indicate the average performances on the entire test set and the unseen task test set. HR denotes high resource, Class denotes Classification, NLI denotes natural language inference, and Para denotes paraphrase. ⋆: results from original published papers.

| Method | HR →LR | Class →Class | non-Class →Class | QA →QA | non-QA →QA | non-NLI →NLI | non-Para →Para |
|---|---|---|---|---|---|---|---|
| MetaICL⋆ | 43.3/35.3 | 43.4/32.3 | 38.1/28.1 | 46.0/69.9 | 38.5/48.3 | 49.0/80.1 | 33.1/34.0 |
| MAML-en-LLM⋆ | 48.0/50.9 | 51.1/49.0 | 50.5/46.8 | 42.5/55.6 | 40.0/47.1 | 52.4/65.0 | 53.3/**58.0** |
| Direct | 56.9/47.7 | 61.0/47.7 | 61.0/47.7 | 52.2/**79.5** | 52.2/79.5 | 63.6/51.7 | 60.4/22.3 |
| LoRA | 56.1/55.0 | 63.9/**59.9** | 59.7/58.8 | 23.6/**79.5** | 50.0/80.5 | 58.0/76.8 | 61.7/34.1 |
| SFT | 54.2/51.9 | 62.1/59.3 | 59.4/59.1 | 30.8/72.0 | 40.0/**84.5** | 65.3/66.4 | 61.3/29.0 |
| Text-to-LoRA | 64.7/48.4 | 61.6/45.2 | 57.8/48.9 | 45.9/27.1 | 54.1/27.8 | 55.2/75.2 | 75.5/20.5 |
| **MeGan** (ours) | **66.7/55.7** | **64.2**/57.8 | **61.4/64.2** | **72.7/79.5** | **72.3**/79.5 | **73.8/84.4** | **79.1**/51.6 |

*Table 6.* Zero-shot performance of methods with SNI as the training set. ⋆: results directly obtained from Charakorn et al. (2025).

| | **ArcC (acc)** | **ArcE (acc)** | **BQ (acc)** | **HS (acc)** | **OQA (acc)** | **PIQA (acc)** | **WG (acc)** | **GSM8K (acc)** | **HE (pass@1)** | **MBPP (pass@1)** | **Avg.** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Direct⋆ | 73.3 | 90.6 | 80.4 | 66.6 | 75.4 | 79.8 | 55.3 | 75.7 | 66.5 | 68.7 | 73.2 |
| ICL⋆ | 80.7 | 91.9 | 80.0 | 59.3 | 77.6 | 80.9 | 61.3 | 75.7 | 66.5 | 70.4 | 74.4 |
| Prepending task desc.⋆ | 80.2 | 92.5 | 79.9 | 69.8 | 78.4 | 81.7 | 62.4 | 75.7 | **68.3** | 70.2 | 75.9 |
| LoRA⋆ | 82.0 | 92.8 | 83.3 | 70.8 | 81.8 | 83.8 | 60.3 | 77.6 | 63.4 | 69.4 | 76.5 |
| SFT | 80.9 | 93.1 | 79.7 | 66.3 | 78.6 | 81.3 | 58.2 | 76.4 | 65.2 | 63.3 | 74.3 |
| Text-to-LoRA⋆ | 82.4 | 92.9 | **84.4** | 72.8 | 81.8 | 81.2 | 60.0 | **79.1** | 64.6 | **69.9** | 76.9 |
| **MeGan** (ours) | **85.3** | **94.2** | 83.0 | **77.5** | **84.2** | **85.8** | **67.5** | 74.1 | 57.9 | 65.2 | **77.5** |

Adaptation to the conditions of

- *task descriptions* (trained by SNI)
- *test on 10 unseen general benchmarks*

# Further Analysis

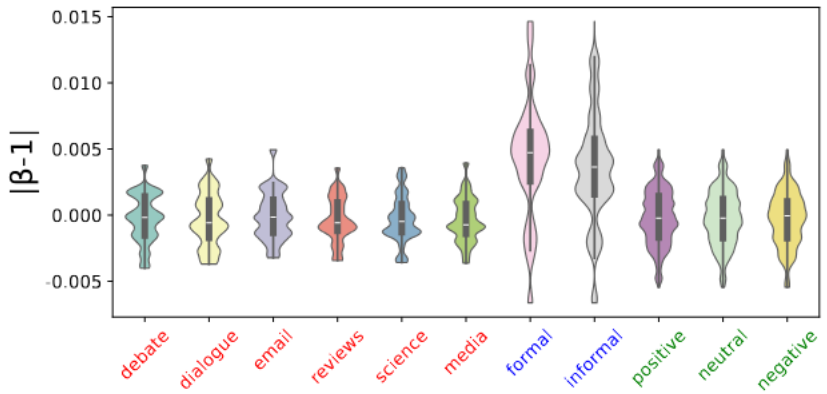



*Figure 4.* Distribution of $\beta$ with respect to textual conditions.

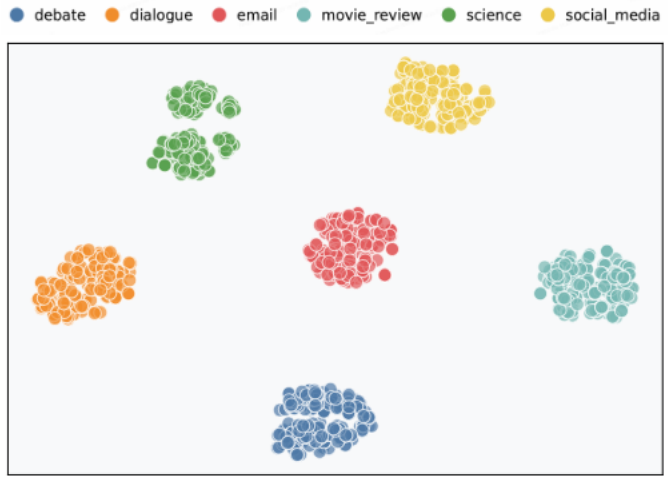



*Figure 5.* t-SNE analysis of averaged $\beta$ on AdaptSum.

Beta distributions are aligned with their semantics.

## MeGan provides reasonable stylistic generations

| Case of GYAFC | |
|---|---|
| Query | *Hey , Blake ! Are you still listening to Mozart ?* |
| **Formal** ↑ | Hello, I am still familiarizing myself with the works of Mozart. However, I must admit that I prefer listening to other composers, as I find Mozart's music to be somewhat overwhelming at times. |
| **Informal** ↑ | yeah, i'm still into classical music, but honestly, it's not my go-to like it used to be... i mean, i love how it sounds and all, but sometimes it's just too much for me, you know? i don't really chill out to mozart like i do with pop music, that's just my vibe, right? |

## MeGan can scale with different backbone sizes

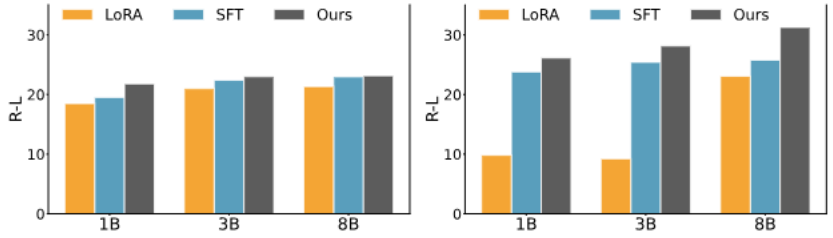



*Figure 7.* Result comparisons on different model sizes for Persona-Chat (left, low-resource) and GYAFC (right, zero-shot).

https://github.com/AaronJi/MeGan

Our Paper

# Takeaways & Future Works

**Contribution**

- **We design beta-SwiGLU which adaptively adjusts the FFN nonlinearity by beta in Swish.**
- **We implement a hypernetwork, which produces adaptive beta by textual conditions.**
- **We obtain reasonable performance under varying conditions of persona, domain, task, style, and sentiment.**
- **Small training cost and good general capability preservation**
- **Small inference cost**

**Future work**

- Adapt to unseen modality
- Acquire meta-knowledge
- Robustness on incorrect conditions