Time is Encoded in the Weights of Finetuned Language Models

Yeajin Lee

Feb 26, 2024

Contents

Background

- **Temporal Variation** : a fundamental characteristic of language.
  - **Temporal misalignment**

  : Deviations in train and test data lead to large performance degradation across different time periods.

➡️ Adaptation techniques for customizing models to specific time periods as needed

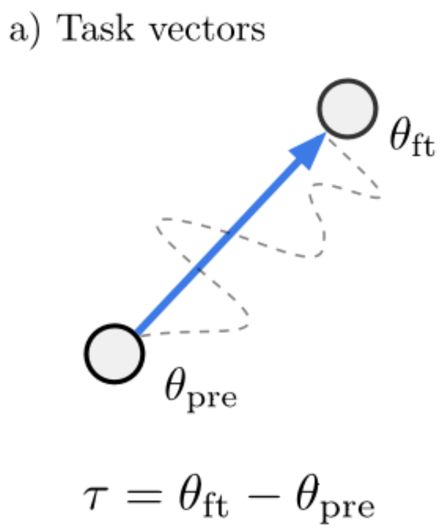
➡️ **Weight-space interpolation**

can be used to cheaply edit language model behavior over time.

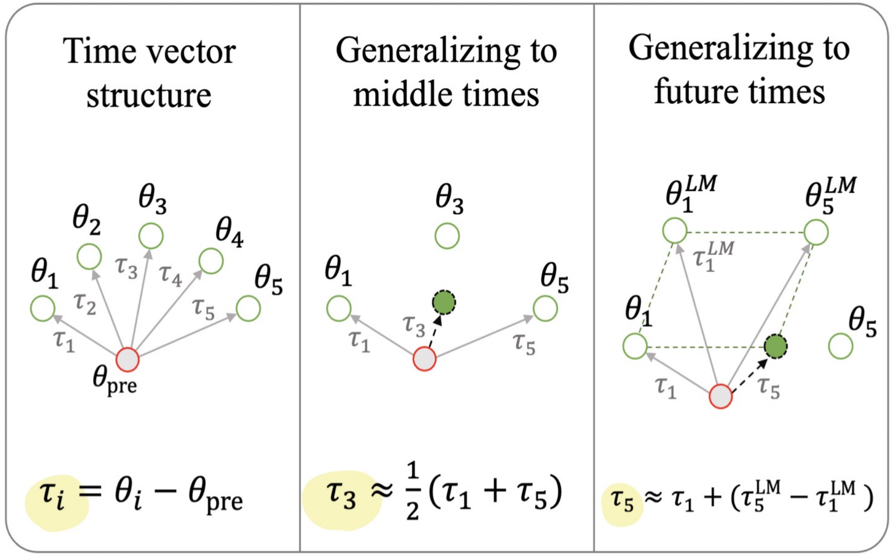
Time vector

: simple tool **to customize language models to new time periods**.

- An extension of task vectors to the time domain.
  - **Task vectors** : formed by taking the difference of a model finetuned on a specific time and the pre-trained model.



Editing Models with Task Arithmetic – Figure1



현재 논문 - Figure1

Data and Finetuning

- Data -

1) **Language Modeling**

   - WMT Language Modeling

   (English subset of the WMT news dataset)

   - Twitter Language Modeling

   (Internet Archive Twitter Stream Grab)

2) **Downstream Tasks**

   - NewsSum (ROUGE-L)

   - PoliAff(macro F1).

- Finetuning -

- Pretrained **T5**
- Finetune T5- small, T5-large, and T5-3b on each of time
- Finetune with Low-Rank Adaptation
- A single epoch on LM splits and three epochs on downstream task splits

Temporal Misalignment

- Yearly Degradation is Linear

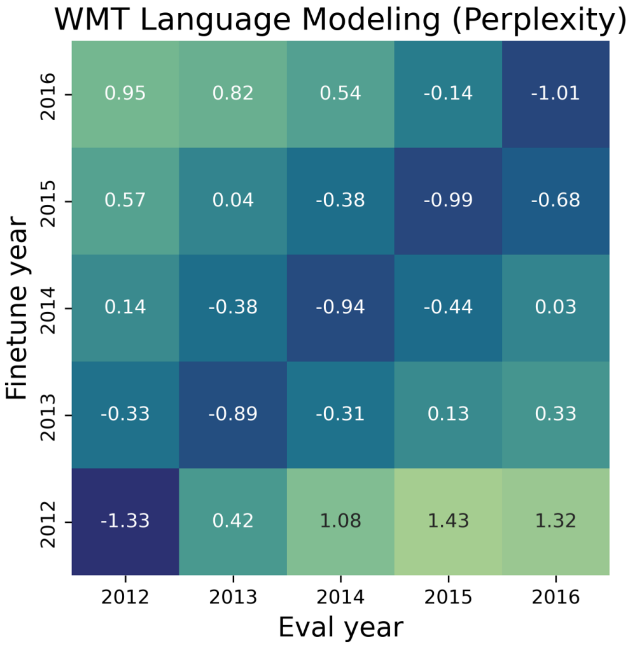


Figure 2: Model performance degrades linearly year-to-year

- Monthly Degradation is Seasonal

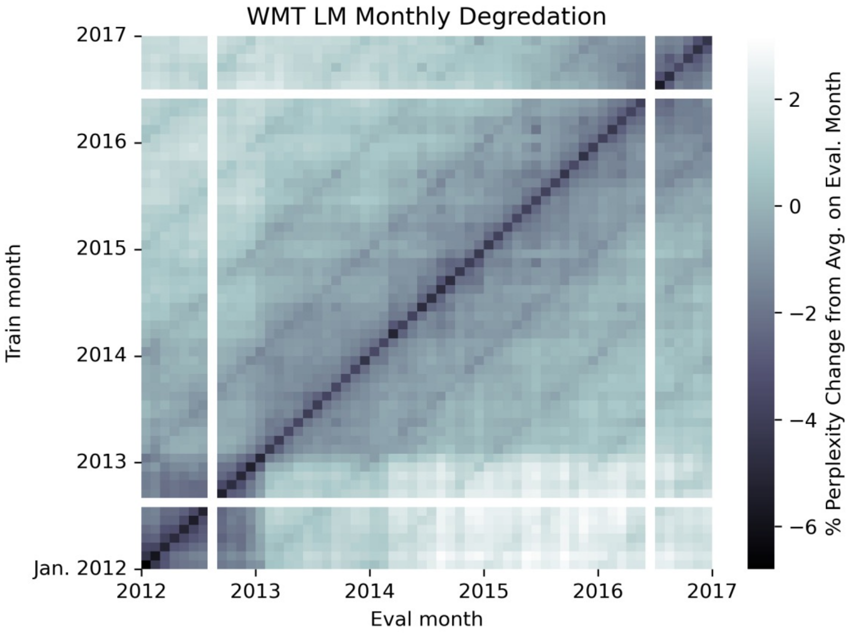


Figure 3: Monthly temporal degradation has seasonal patterns.

Temporal Adaption

- Correlation of Time Vector Similarity and Temporal Degradation

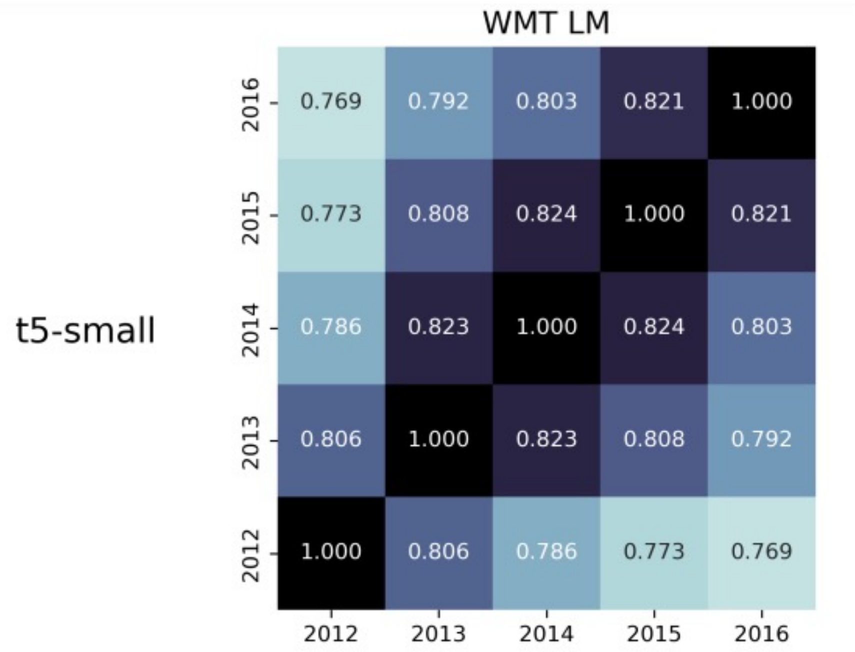


Figure10 : Cosine similarities between all pairs of year time vectors

	Pearson r		
T5 size	WMT LM	NewsSum	PoliAff
small	-0.867	0.663	0.654
large	-0.737	0.628	0.672
3b	-0.795	0.626	0.668

Table 1: The similarity between time vectors correlates with temporal degradation.

Temporal Adaption

- Generalizing to Intervening Time periods

- Method
  - Two time vectors $\boldsymbol{\tau_j}$ (oldest), $\boldsymbol{\tau_k}$ (newest)
    - $\boldsymbol{\tau_j} = \boldsymbol{\theta_j} - \boldsymbol{\theta_{pre}}$
  - Compute interpolation : $\boldsymbol{\alpha} \cdot \boldsymbol{\tau_j} + (\mathbf{1} - \boldsymbol{\alpha}) \cdot \boldsymbol{\tau_k}$
  - With $\alpha \in [0,1]$

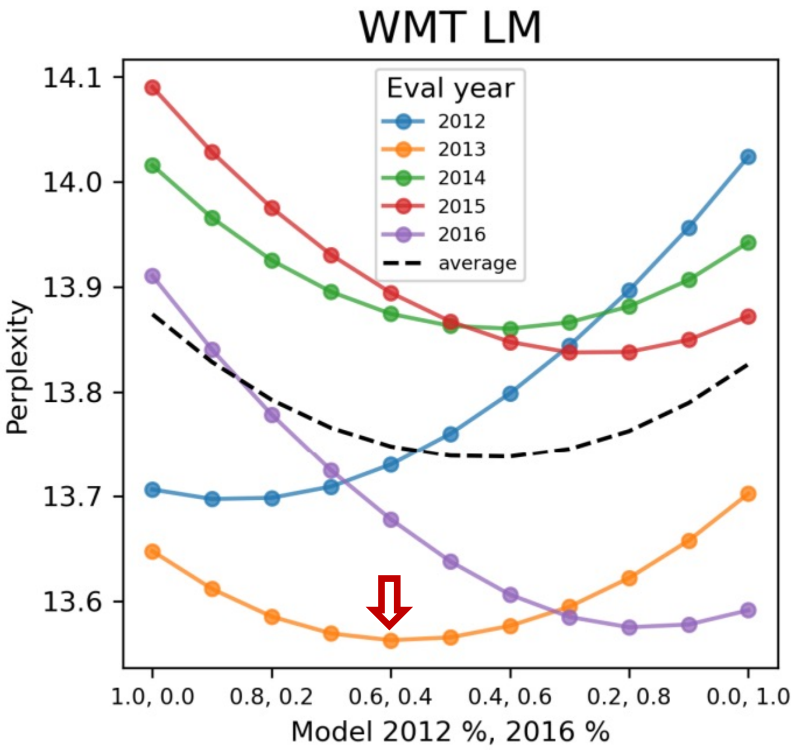


Figure 5: Interpolating between two year vectors improves performance on the years between them.

Temporal Adaption

	Perplexity (↓)	*Rouge* (↑)	*F1* (↑)
Method	WMT LM	NewsSum	PoliAff
Start-year finetuned (τ_0)	13.92	38.56	0.6886
End-year finetuned (τ_n)	13.84	35.09	0.6967
$\frac{1}{2}(\tau_0 + \tau_n)$	13.77	38.86	0.7765
Best interpolations	13.75	40.11	0.7941
Eval-year finetuned (τ_i)	**13.65**	**42.36**	**0.8341**

Table 2 : Interpolation between start and end-year finetuned models reduces temporal misalignment on intervening years.

- **Best interpolations** : Use the best performing α values for each year
- **Eval-year finetuned** : Performance of finetuned models for each year

Temporal Adaption

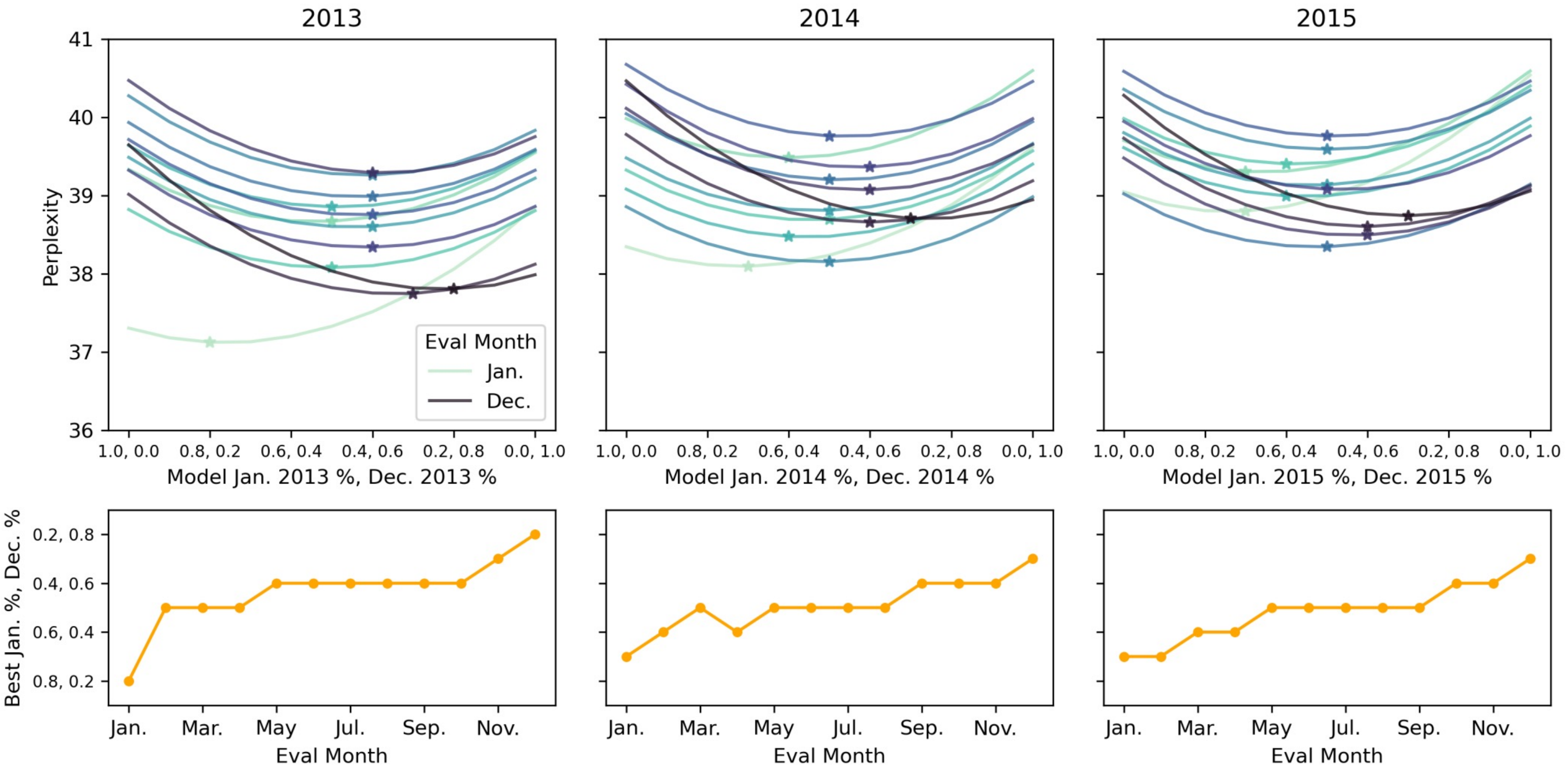


Figure 6: Interpolating between two month vectors improves performance on the months between them.

Temporal Adaption

- Generalizing to the Future

: a new technique for updating task models finetuned on **time period j** to a **target time period k** with only unlabeled data from j (labeled data)

$$\tau_j = \theta_j - \theta_{pre}$$
$$\tau_j^{LM} = \theta_j^{LM} - \theta_{pre}$$
$$\tau_k^{LM} = \theta_k^{LM} - \theta_{pre}$$

$$\boxed{\tau_k \approx \alpha_1 \cdot \tau_j + (\alpha_2 \cdot \tau_k^{LM} - \alpha_3 \cdot \tau_j^{LM})}$$

$$\underline{\theta_k} = \tau_k + \theta_{pre}$$

- Method
  - Given $\boldsymbol{\theta_j}$, $\boldsymbol{\theta_j^{LM}}$, $\boldsymbol{\theta_k^{LM}}$
  - Estimated $\boldsymbol{\theta_k}$
  - $\alpha_1 \in [0.6,0.8,..,2.2]$, $\alpha_2, \alpha_3 \in [0.1,...0.6]$

Temporal Adaption

- Update **2012 News-Sum model** to **2013–2016**, and **2015 PoliAff model** to **2016–2020**.
- Improvement **increases** as the target and start years become more misaligned.
- Model size also affects performance.

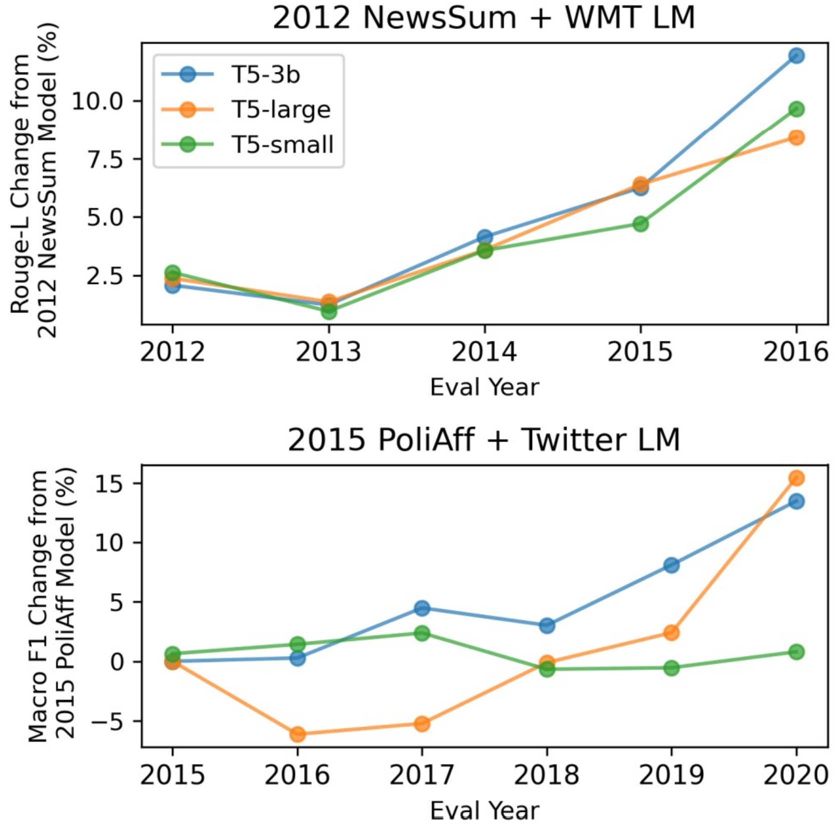


Figure 7 : Task analogies can offset downstream temporal misalignment without labeled data from the target time.

Temporal Adaption

$$\tau_k \approx \alpha_1 \cdot \tau_j + (\alpha_2 \cdot \tau_k^{LM} - \alpha_3 \cdot \tau_j^{LM})$$

- Only scaling α_1 can also **improve** performance on future years.
- **Scaling only** : only scaling the base τ_j model ($\alpha_1 \neq 0$, $\alpha_{2,}$ $\alpha_3 = 0$)
- **Task addition** : only adding the language modeling vector $(\alpha_1, \alpha_2 \neq 0, \alpha_3 = 0)$

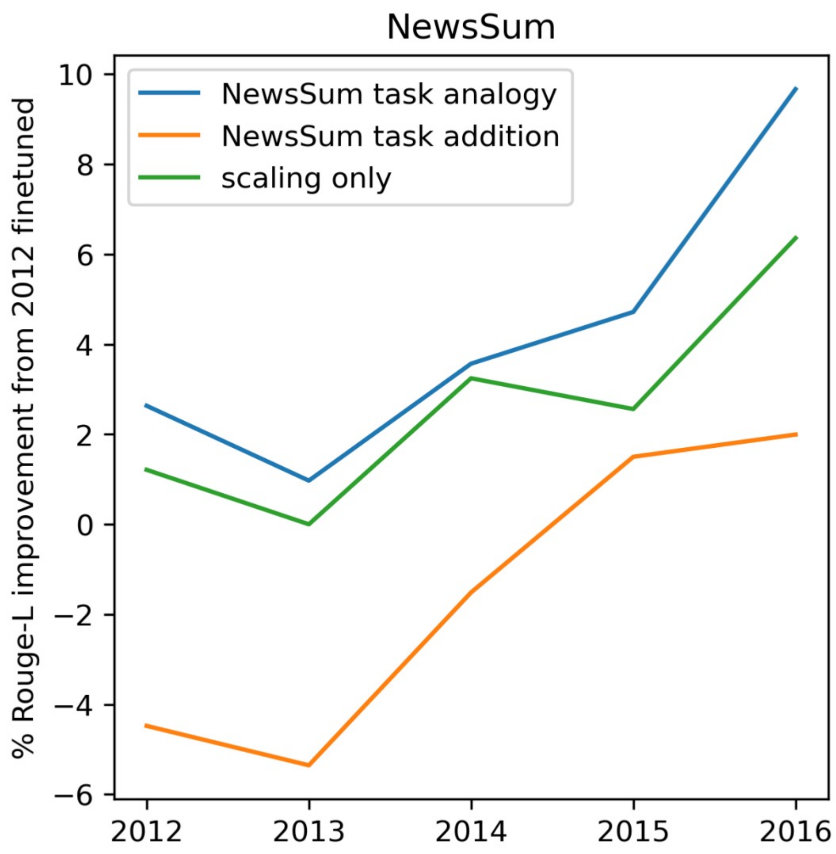


Figure 15: Time vector analogy ablations for three sizes of T5

Temporal Adaption

- Generalizing to Multiple Time Periods

- Method
  - Model soup technique
    - **Uniform soup** : A uniform weight among all constituent models in the interpolation.
      ➡️ $\theta_{pre} + \frac{1}{|T|}\sum_{t\in T}\tau_t$
    - **Greedy soup** : Only includes models in the soup that improves validation performance.
  - Measure the average performance across all evaluation years for each task.

Temporal Adaption

	Perplexity (↓)	*Rouge* (↑)	*F1* (↑)
Method	WMT LM	NewsSum	PoliAff
Best single-year model	34.45	38.95	0.7101
Uniform time soup	34.70	33.05	0.6078
Greedy time soup	34.45	38.95	0.7202
Training on all years	**29.17**	**40.07**	**0.7853**

Table3 : Interpolation does not enable generalization to multiple time periods simultaneously

- Time soups perform **worse** than the model finetuned on all shuffled available data.
- A model which generalizes to multiple time periods does **not lie** in a region of weight space bounded by models finetuned on single years of data.

Conclusion

- Connect studies of temporal misalignment and weight arithmetic with time vectors.
- The similarities of weights of each different time **are highly correlated** to temporal misalignment at both yearly and monthly scales.
- Induce new models that **perform better** on intervening years by interpolating between adjacent time vectors.
- Use task analogies **to improve downstream performance** on future time periods using only unlabeled data from those times.

⇒ **Weight arithmetic can be a simple tool for combating temporal misalignment.**

Thank you for Listening!

Reference:

Paper - https://arxiv.org/pdf/2312.13401.pdf
Task vector paper – https://arxiv.org/abs/2212.04089
Model soup technique paper - https://arxiv.org/abs/2203.05482