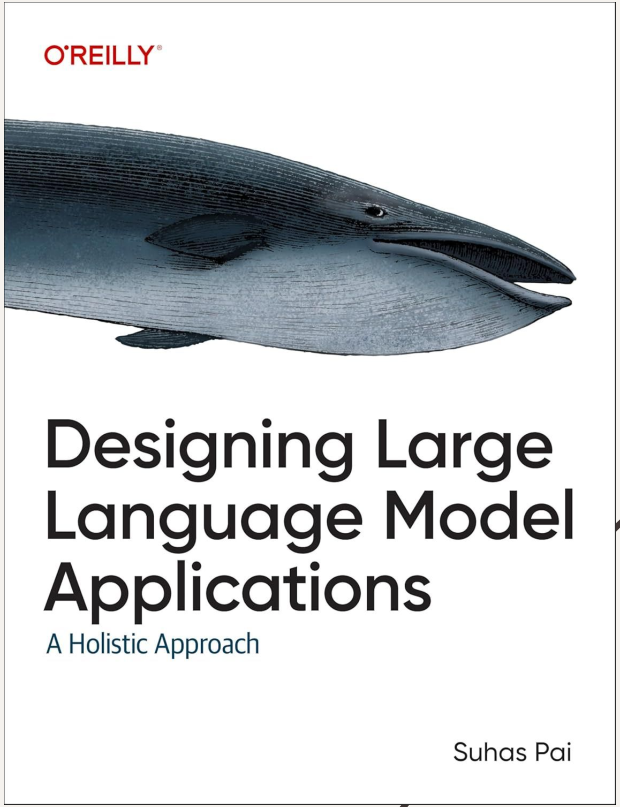
Continual Learning in LLMs

- Suhas Pai, CTO, Hudson Labs

About Me

- CTO & ML Research @ Hudson Labs (https://hudson-labs.com)
- Book ‘Designing LLM Applications’ published by O’Reilly Media
- Led/contributed to various open-source LLMs (BLOOM, Aurora-M etc.)
- Chair, TMLS (Toronto Machine Learning Society), MLOPS World conference
- Leading an independent research group ‘llm-playbooks’



Why does catastrophic forgetting happen?

Learning Rate strategies for pre-training LLMs

- Cosine Annealing with warmup
- Convergence to stable equilibrium

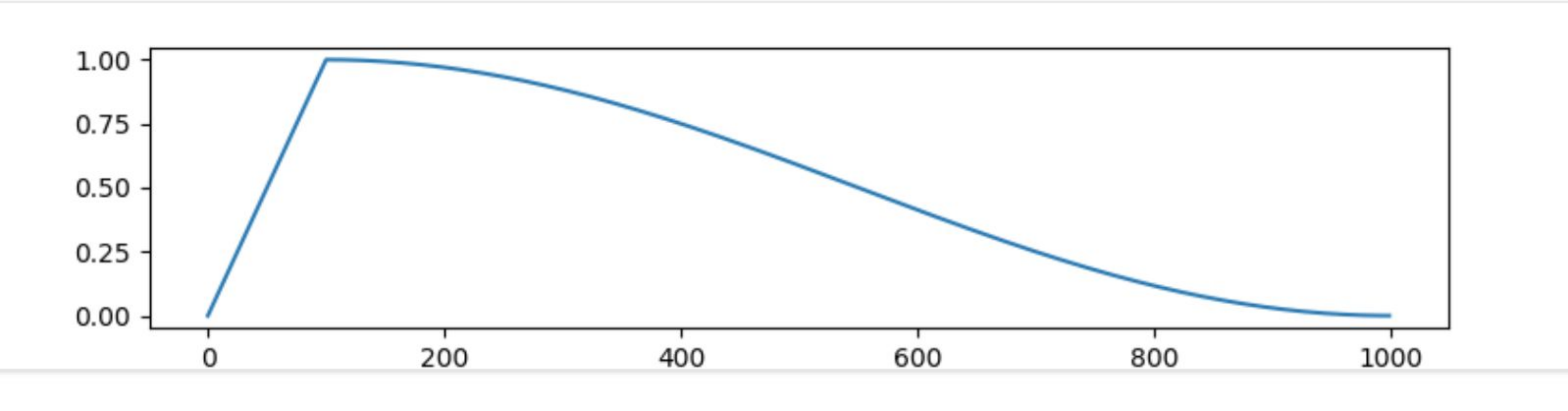
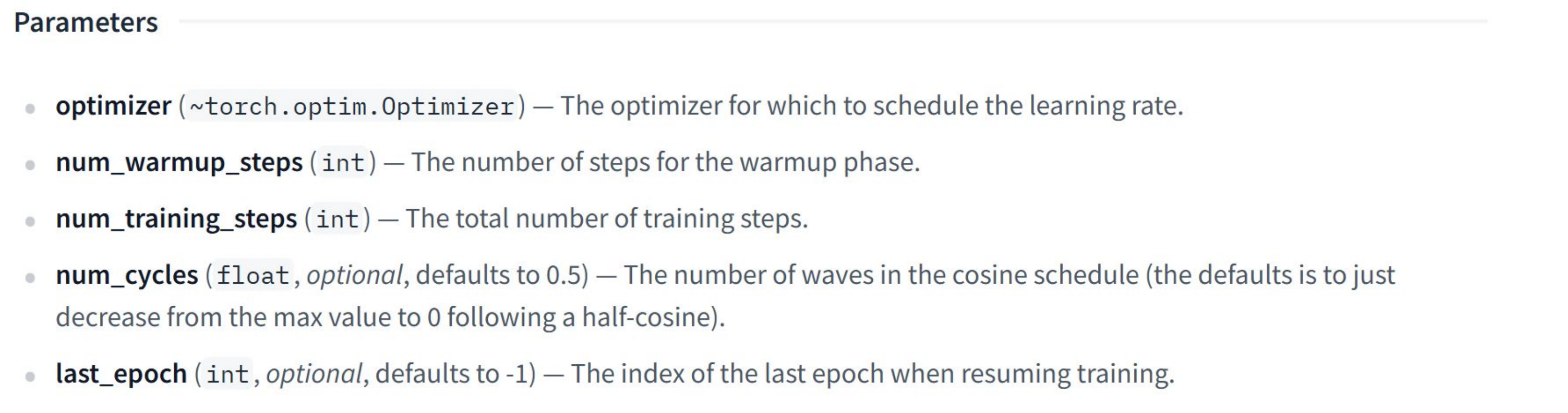


Image Credit: Hugging Face

Cosine Schedule - Hugging Face



Parameters

- **optimizer** (`~torch.optim.Optimizer`) — The optimizer for which to schedule the learning rate.
- **num_warmup_steps** (`int`) — The number of steps for the warmup phase.
- **num_training_steps** (`int`) — The total number of training steps.
- **num_cycles** (`float`, *optional*, defaults to 0.5) — The number of waves in the cosine schedule (the defaults is to just decrease from the max value to 0 following a half-cosine).
- **last_epoch** (`int`, *optional*, defaults to -1) — The index of the last epoch when resuming training.

Problems with cosine annealing

- Rewarming from minimum values causes instability and forgetting
- Need to know number of training steps in advance
- WSD (Warmup-Stable-Decay) is gaining more prominence these days

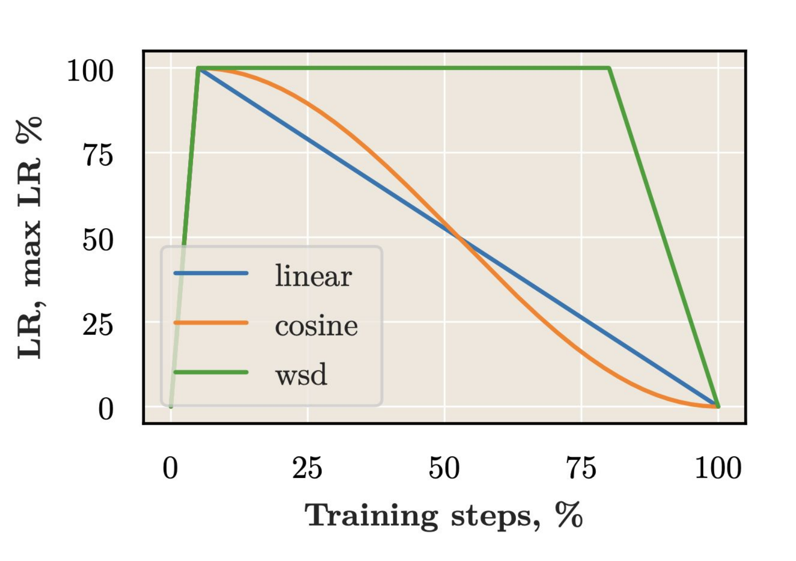
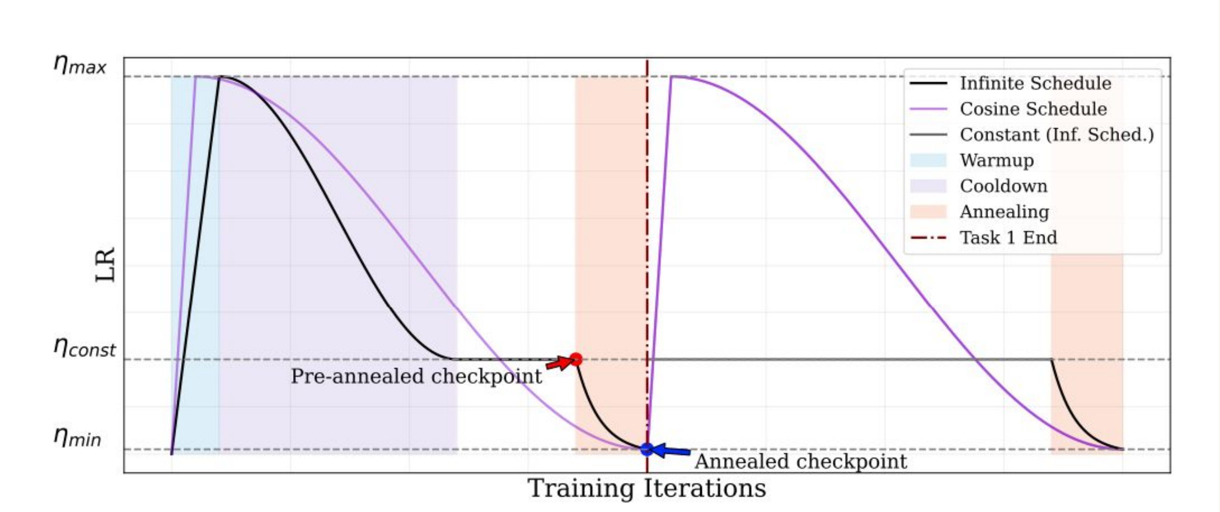


Image Credit: Dremov et al. (Aug 2025)

Infinite Learning Schedule

- Initial warmup
- Decay
- Plateau
- Rapid annealing



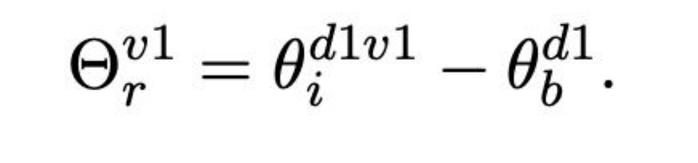
Singh et al. (Sep 2025)

Demo

Should you continually learn over the base model or the instruction tuned model?

Catastrophic forgetting in instruct models

- Continued pre-training on instruction tuned models degrades instruction following capability
- Instruction tuning should be performed after continued pre-training
- Instruction residuals can be used to prevent repetitively performing instruction tuning after every round of continued pre-training



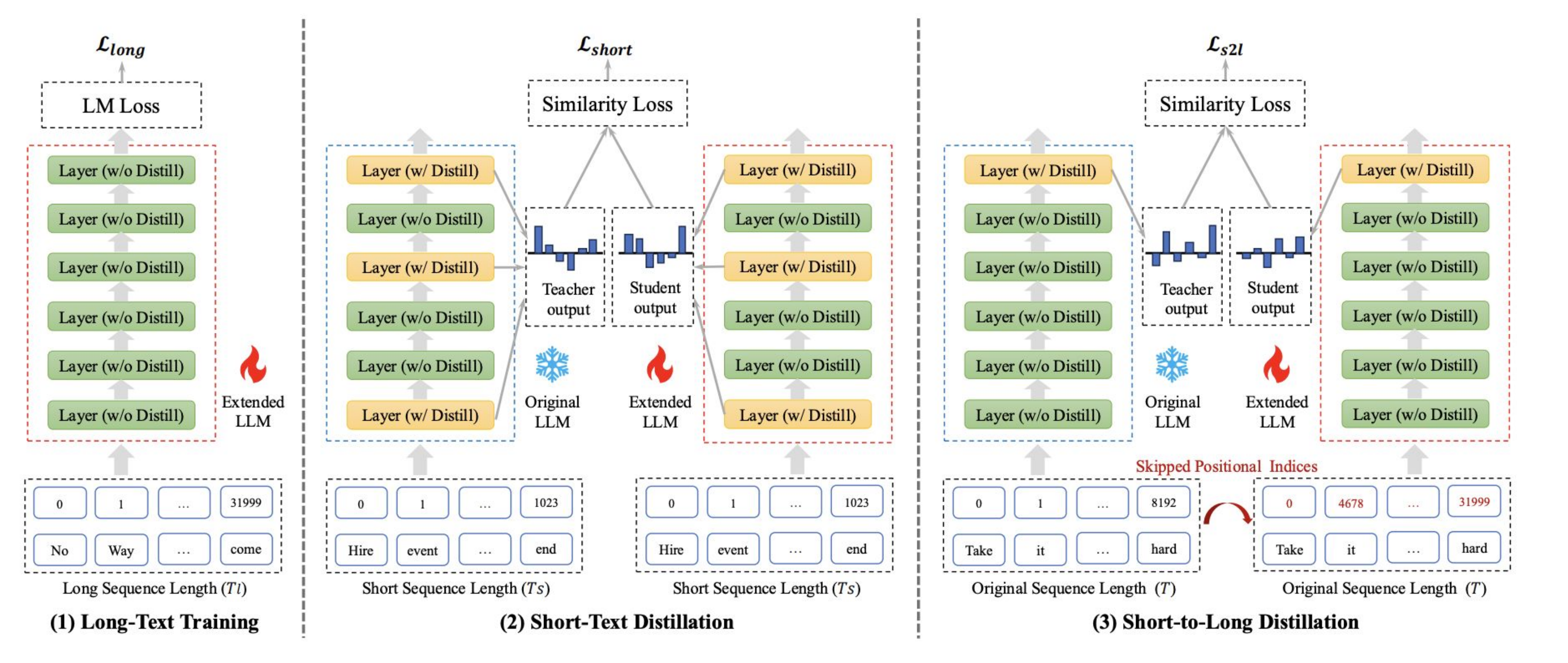
$$\Theta_r^{v1} = \theta_i^{d1v1} - \theta_b^{d1}.$$

Jindal et al. (Oct 2024)

Catastrophic forgetting in long-context models

- Training on long-context data reduces performance on short-text data
- Rudimentary methods involve replaying short-context data during long-context training

Long Context Pre-training with Restoration Distillation



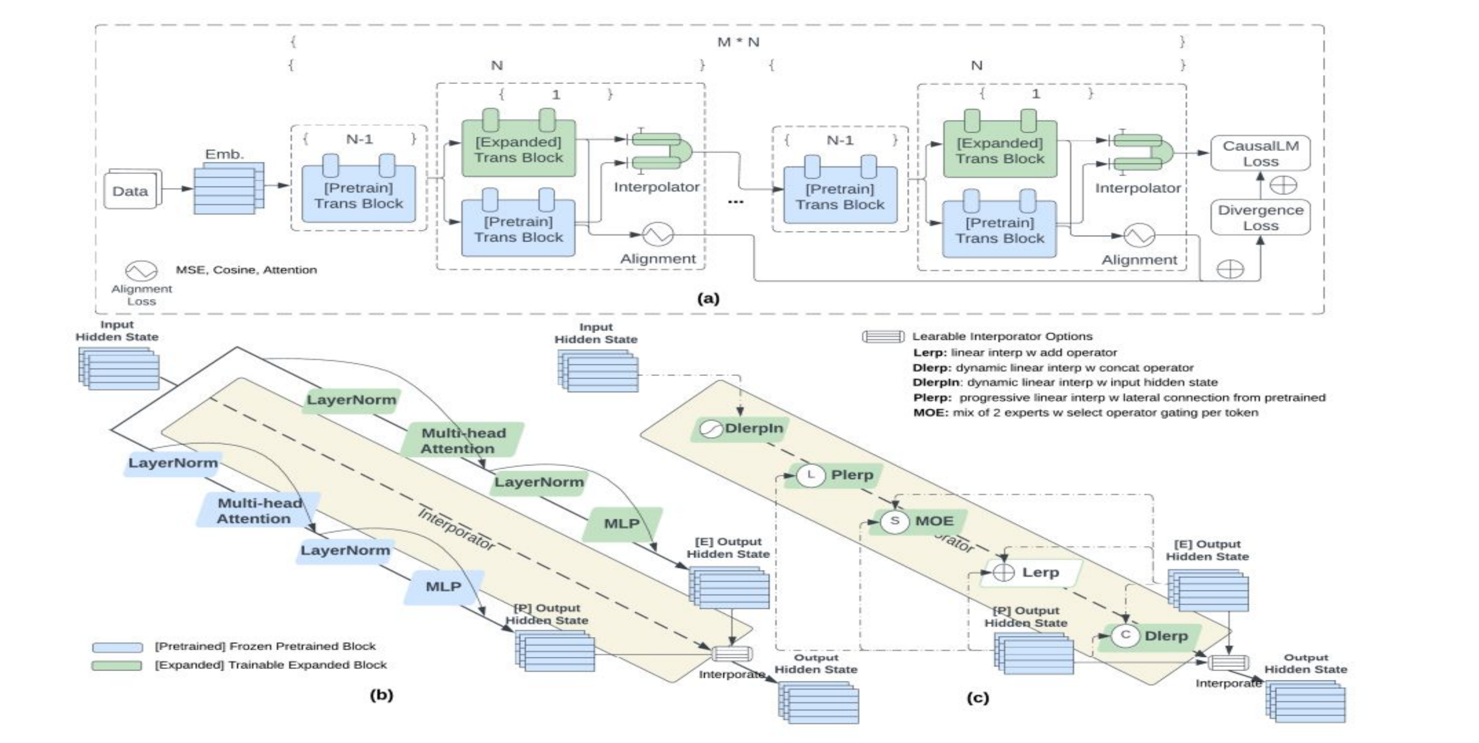
Dong et al. (May 2025)

Updating Parameters

Interpolating base and extended models

- Linear Interpolation
- Dynamic Linear Interpolation
- Dynamic Linear Interpolation on input hidden-states
- Progressive Linear Interpolation
- Mixture of Experts Gating

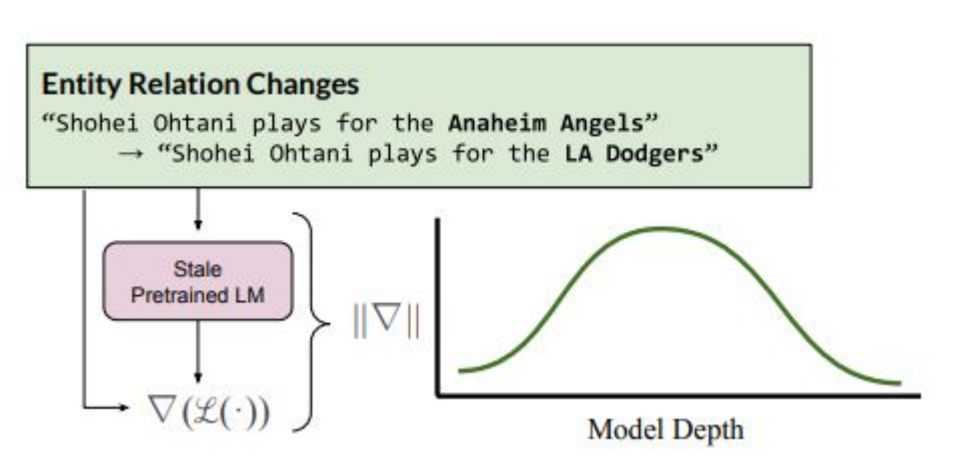
Interpolating intermediate states



Wei et al. (Jan 2025)

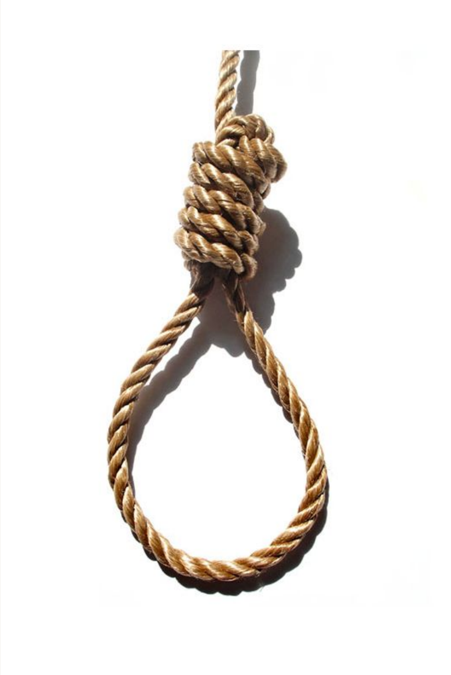
Model Editing

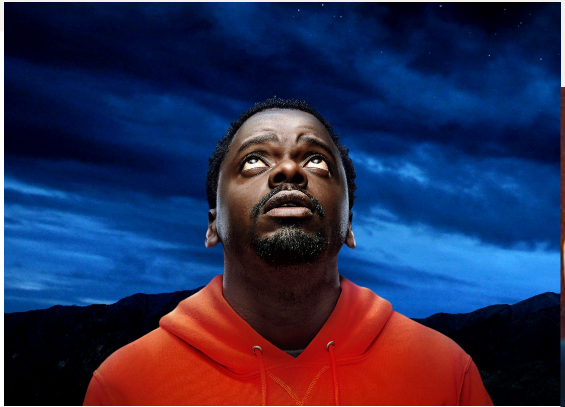
- Large gradient norms observed in certain layers when predicting entity tokens

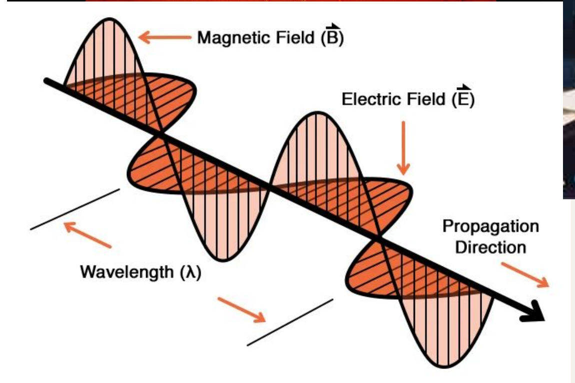

Entity Relation Changes
“Shohei Ohtani plays for the **Anaheim Angels**”
→ “Shohei Ohtani plays for the **LA Dodgers**”
Stale
Pretrained LM
$\nabla(\mathcal{L}(\cdot))$
$\|\nabla\|$
Model Depth

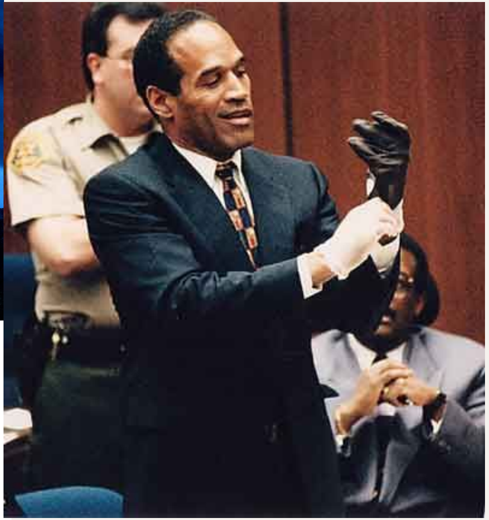
Fernandez et al. (Dec 2024)

Trivia Time: Connect the following









Contact

- piesauce.substack.com
- https://x.com/piesauce
- https://www.linkedin.com/in/piesauce/