



**Intro Fellowship
Spring 2024**

[WEEK 5]

Model internals

READINGS

Locating and Editing Factual Associations in GPT: Companion blog post
(Meng et al.,)

Discovering latent knowledge in language models without supervision (Pages 1-5)
(Burns et al., 2022)

Inference-Time Intervention: Eliciting Truthful Answers from a Language Model
(§1-3)
(Li et al., 2022)

lol