

Réunion 3 juin 2026

- Nouvelles Perturbations
- Mechanistic Interpretability et nos hypothèses

Brainstorming

- Entrées adaptées ou changer un mot change le contexte de la phrase
 - *flou*
- exploiter des informations sur les questions pour identifier des perturbations
- perturbation de type append et prepend
- Fine-tuner E5 pour une meilleure sensibilité à la langue de spécialité ciblée
- Utiliser mE5...

Survey sur Mechanistic Interpretability

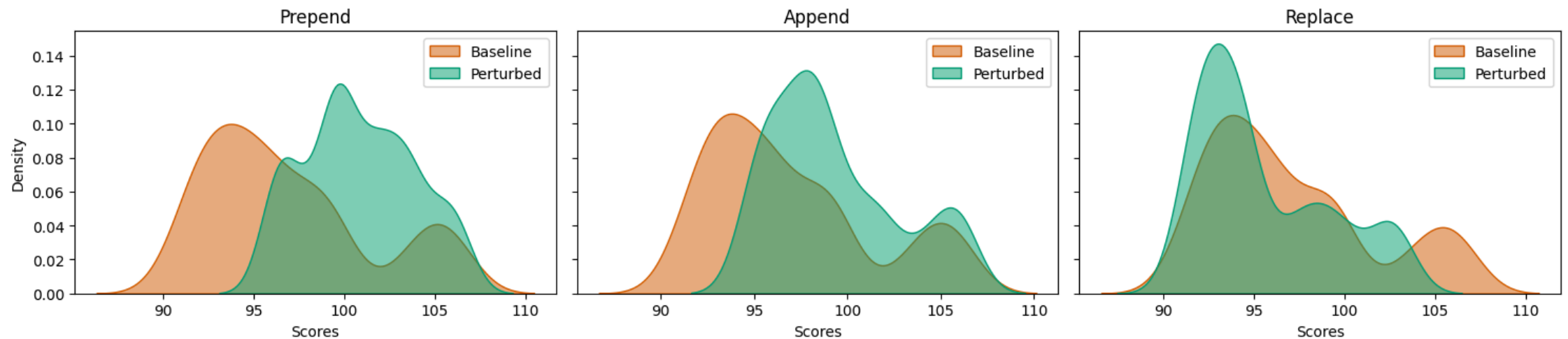
[Somvanshi et al., 2026] Bridging the Black Box: A Survey on Mechanistic Interpretability in AI

Superposition et polysémantique : Les composants traitent plusieurs features à la fois (du fait de la contrainte d'espace pour traiter tous les facteurs importants)

⇒ Contradiction avec Hypothèse *Models have specific components dedicated to domain specific language*

Perturbations de type append et prepend

Distribution of Baseline vs Perturbed Scores



- Append/ Prepend "microwave"
- "microwave" -> "toaster"

