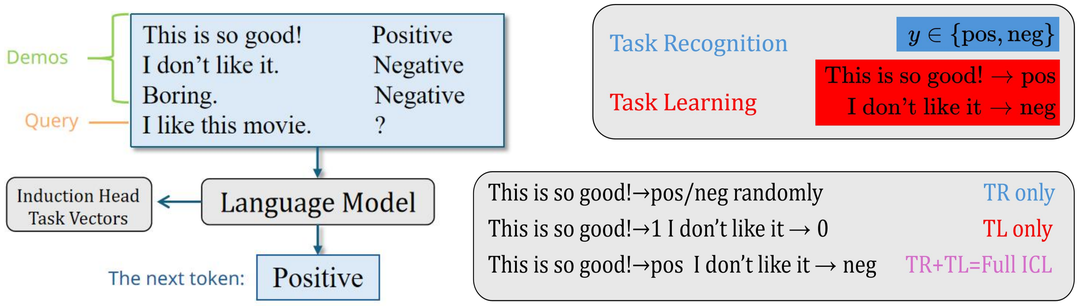
Localizing Task Recognition and Task Learning in In-Context Learning via Attention Head Analysis

Haolin Yang[1], Hakaze Cho[2,3,4], Naoya Inoue[2,3]
[1]University of Chicago, [2]JAIST, [3]RIKEN, [4]Tohoku University

Two Approaches for Explaining ICL



Microscopic/Mechanistic Appraoch

- ► **Zoom into the model** to focus on individual model components
- ► Identify key elements **(induction heads, task vectors)** via ablation analysis
- ▲ Provide **fine-grained localization** of the model elements facilitating ICL
- ▼ Only answers **how much** but not **how** these components affect ICL

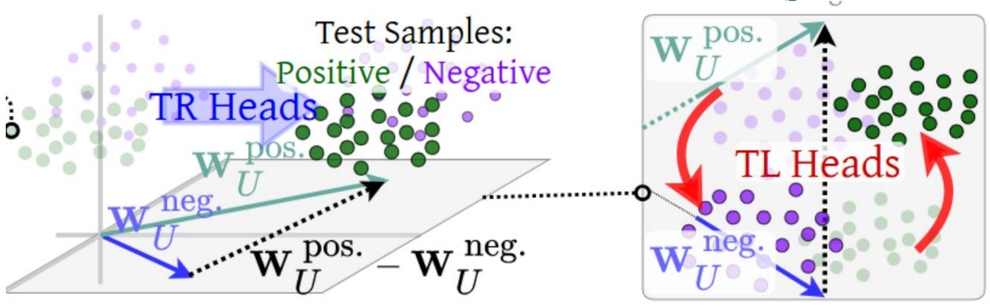
Macroscopic/Functional Approach

- ► Views the model as a whole that executes ICL as a function
- ► Study the ICL function by perturbing the ICL prompts as inputs
- ► Factorize the ICL function into independent **Task Recognition (learning the label space) and Task Learning (learning text-label mapping)** components
- ▲ Explains **how the ICL functionality is achieved** through different components
- ▼ Provides **no clues** regarding how ICL is **implemented by the model internally**

Our Work: Best of Both Approaches

- ✦ **Traces the identified TR and TL functionality** to attention heads
- ✦ Studies **how the heads affect hidden states** to enact the functionality
- ✦ **Mechanistic granularity + functional interpretability !!!**

Findings: Attention Heads Achieve TR/TL by Shaping Hidden States in Distinct Geometric Ways



Identifying TR and TL Heads

- Unembedding $\mathbf{W}_U$, label space $\mathbb{Y}$ correct label $y^* \in \mathbb{Y}$, incorrect labels $\mathbb{Y}/\{y^*\}$, head output α
- **Task space:** $\mathrm{span}(\mathbf{W}_U^{\mathbb{Y}})$, where $\mathbf{W}_U^{\mathbb{Y}}$ is the unembeddings of the tokens in $\mathbb{Y}$

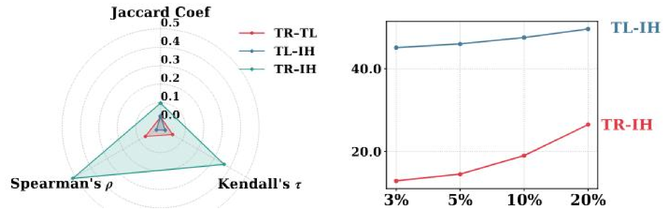
TR score of a head: $\|\mathrm{Proj}_{\mathbf{W}_U^{\mathbb{Y}}}\alpha\|_2$

- how much a head output loads onto and **steers the hidden states toward the task subspace**

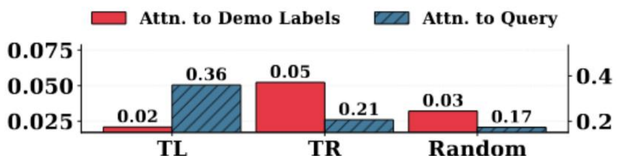
TL score: $(\alpha^\top \mathbf{W}_U^{y^*} - \mathrm{Ave}(\alpha^\top \mathbf{W}_U^{\mathbb{Y}/\{y^*\}}))/\|\mathrm{Proj}_{\mathbf{W}_U^{\mathbb{Y}}}\alpha\|_2$

- how much a head output **differentiates between the correct and incorrect labels** and **steer hidden states towards correct label**

Analyzing Identified Heads

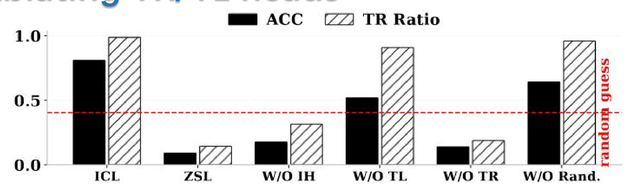


- ↓overlap for TR&TL heads+↓correlation for TR&TL scores -matches functional independence of TR&TL heads√
- ↑ overlap and correlation between TR heads and IHs -Induction Heads facilitate ICL through TR



- TR heads attend to demo. labels to **find the label space to predict from**
- TL heads attend to query context to **learn query semantics for matching**

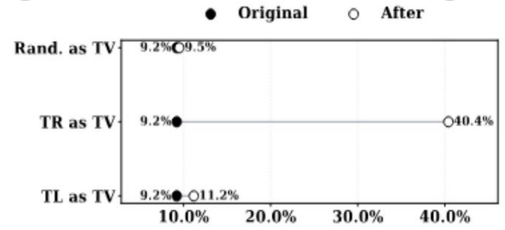
Ablating TR/TL heads



- **TR Ratio:** How often **any of the tokens in the label space** are predicted **(Always ≥ Acc. for classification)**
- **Ablating TR heads disrupt TR → Acc. drops** as TR ratio ↓ and brings down Acc.
- Ablating TL heads **disrupt TL → Model randomly predicts from label space→baseline level Acc.**
- Zero-shot case: **TR failure causes low accuracy**

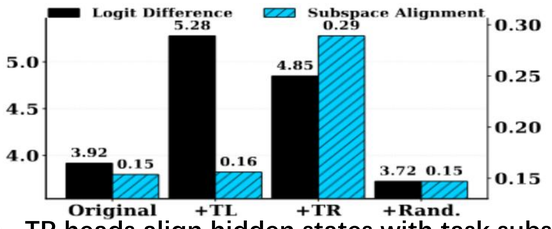
- Shuffle demo. text. **I like it → positive** becomes **tkli i ie → positive. TL impossible**
- Ablating TL has no effect, while ablating TR still works, **confirming functional independence**

Using TR/TL heads for steering

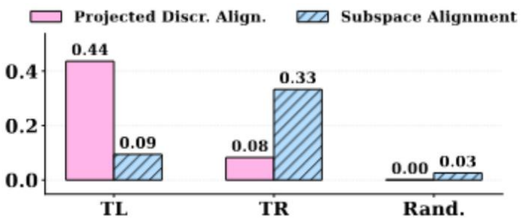


- **Steering using TR** head outputs **solves zero-shot TR failure → Acc. ↑**
- For classification tasks **the label space is crucial**
- **Steering using TL** head outputs: **no effect**

Geometric Effects of TR/TL heads



- TR heads align hidden states with task subspace
- TL heads rotate hidden states to better align with the correct label's unembedding
- matches the geometric intuition √



- TR head outputs themselves also align with the task subspace
- TL head outputs are aligned with the difference direction between correct and incorrect labels