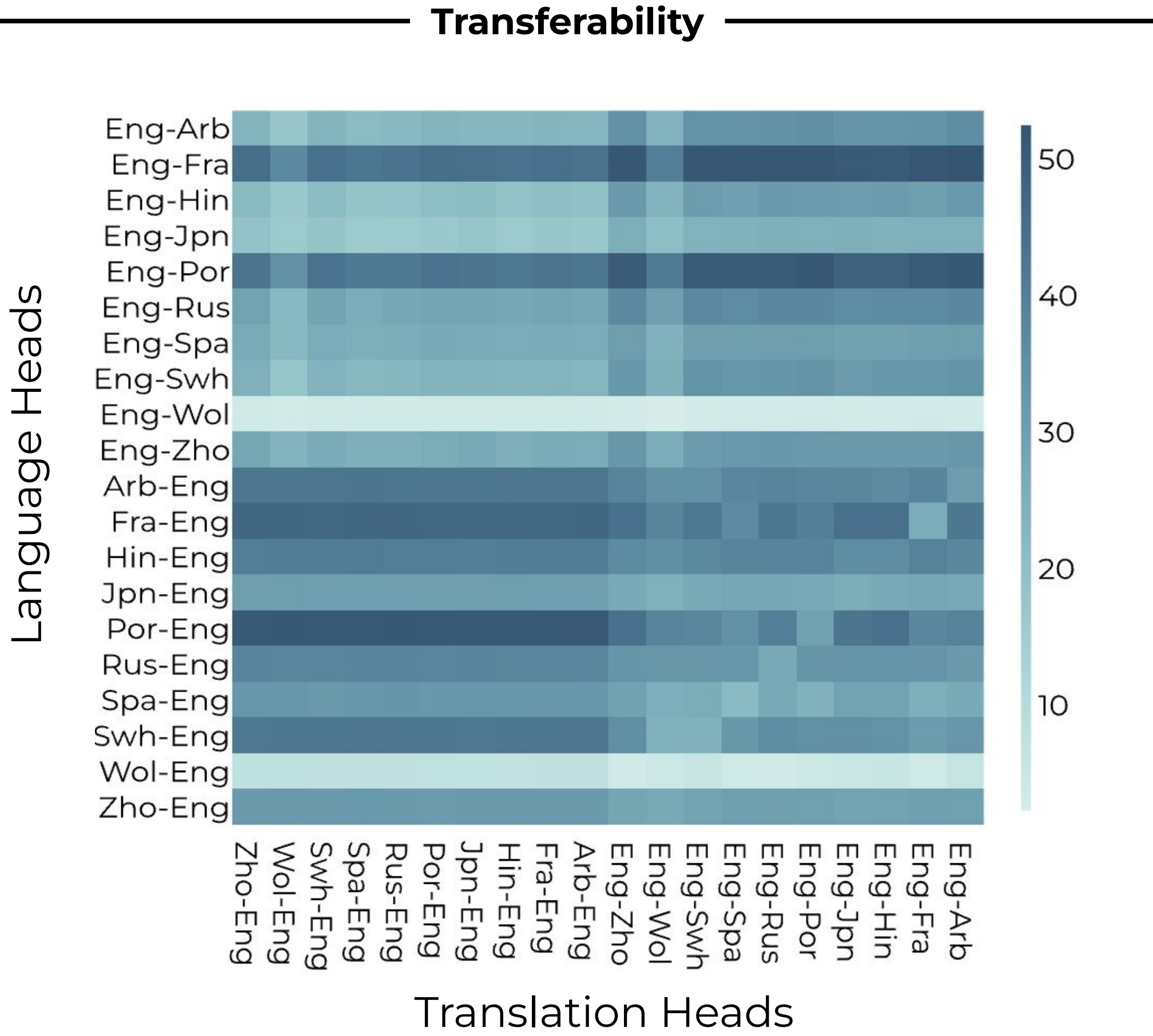
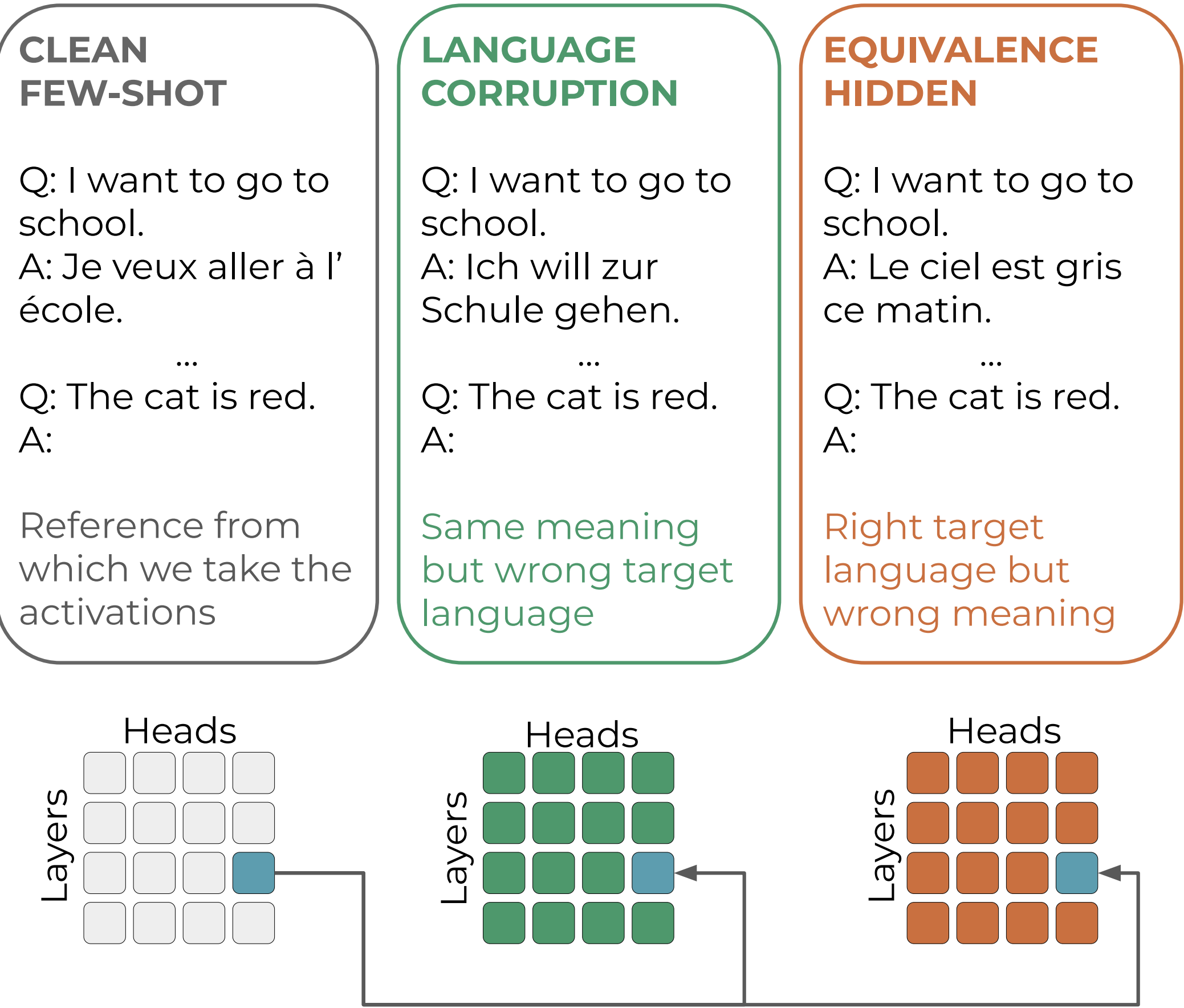


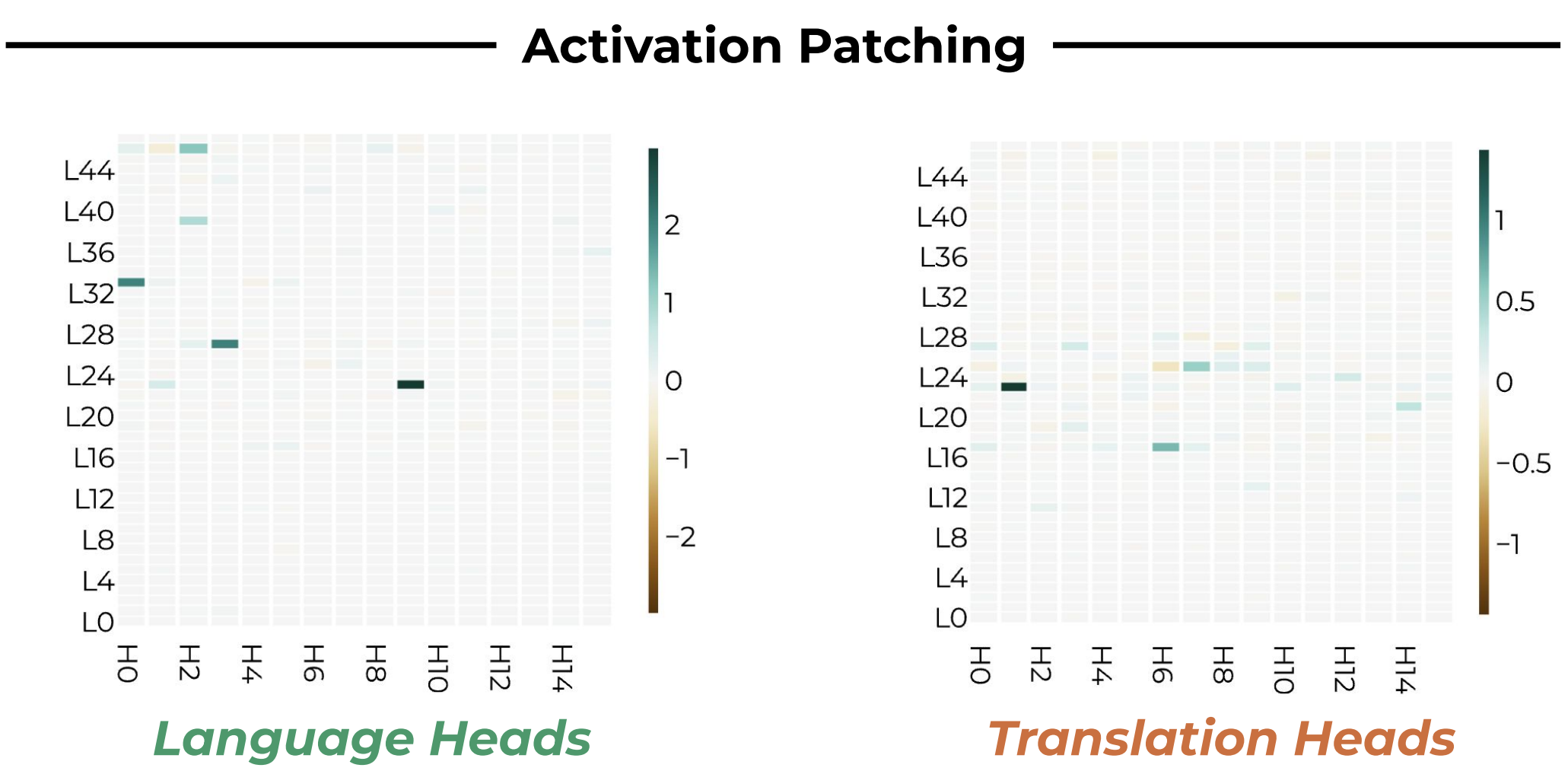
LLMs split the Machine Translation task into Language and Meaning at a Component Level

Translation Heads: Disentangling meaning from language in LLM-based machine translation

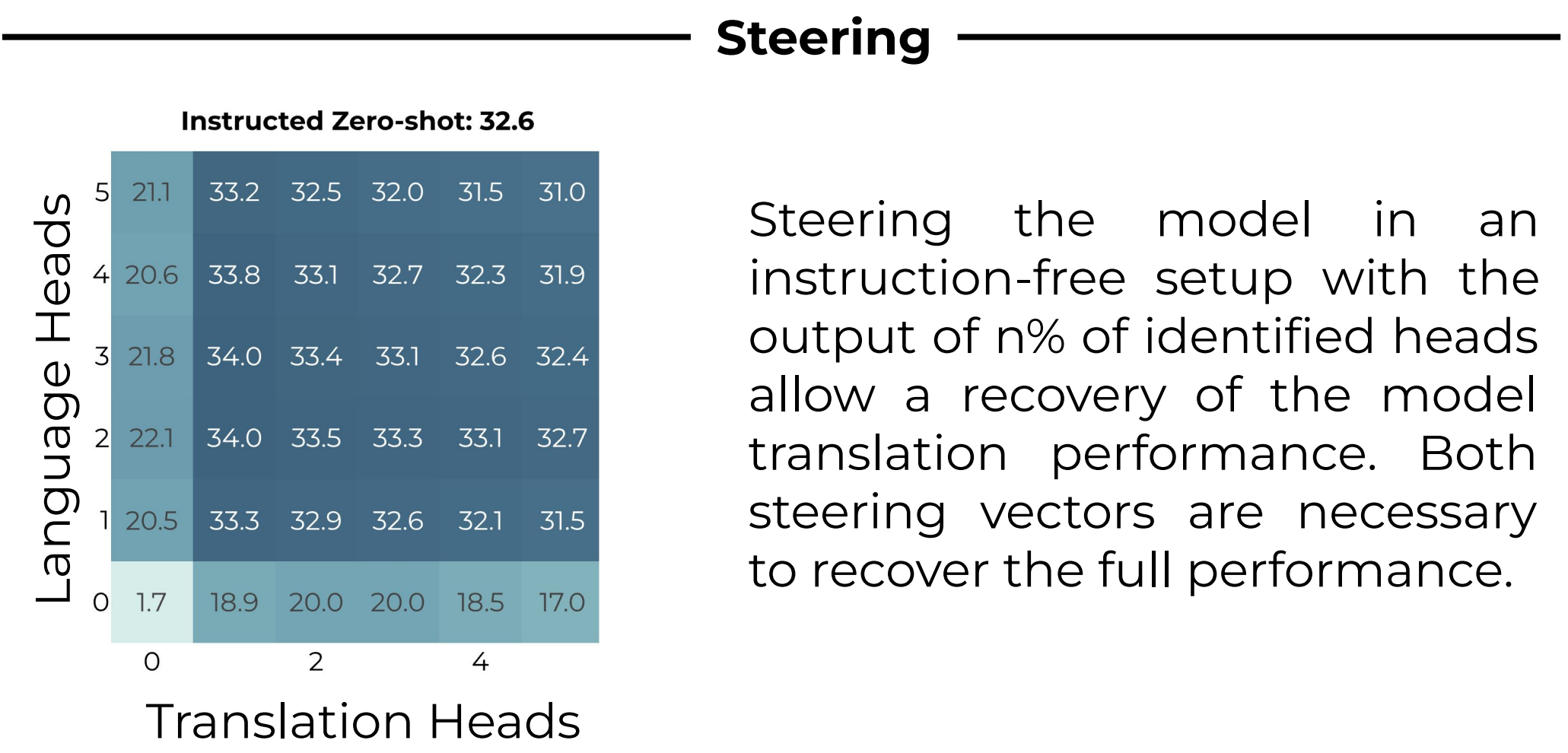
TL;DR LLMs represent the machine translation tasks in two different subtasks (i.e. the sentence equivalence and the target language identification). Each subtask is performed by a constant set of specialized attention heads across translation directions that we call **Translation Heads** and **Language Heads**. The representation outputted from the *Translation Heads* are transferable between translation directions. Taken together, our findings point toward a natural emergence from pretraining of a language agnostic translation representation across models.



We patch the activation of our corrupted runs with our clean runs and look at the logit recovery of the correct next tokens. Additionally, we use a KL criterion to select the most informative token that drives the whole sentence generation.



Important heads under both corruptions are constant across translation directions. **Language Heads** and **Translation Heads** have no overlap.



Language Pair	Translation Head	Language Head
Eng-Fra	í ś ũ Pérez = P Rö ización [ī Pokémon ó чи y īg Ré	French à pour dé Quebec Québec french é France É
Eng-Swh	= í impactful ũ 에서 ś And и у за в чи XNUMX ó [And	Tanzania African Africa Kenya Kenyan ya Zambia
Eng-Zho	в у B Sociedad Chiến за ной Trườg Employees	, : 其他 ; 对 可以 做 (很 后 其他 或 从 上 此外
Eng-Arb	= ś P í чи [And And Grü 에서 ũ 有 Pérez Rö impactfu	ه عم - نأيف ن ط ش ح ر ن ا ق نم ص ب ا ك أ ل ع
Eng-Por	í ś ũ = P Pérez ī impactful ización y Pokémon czek В и	Brazilian Portuguese Brazil brazilian Portugal

