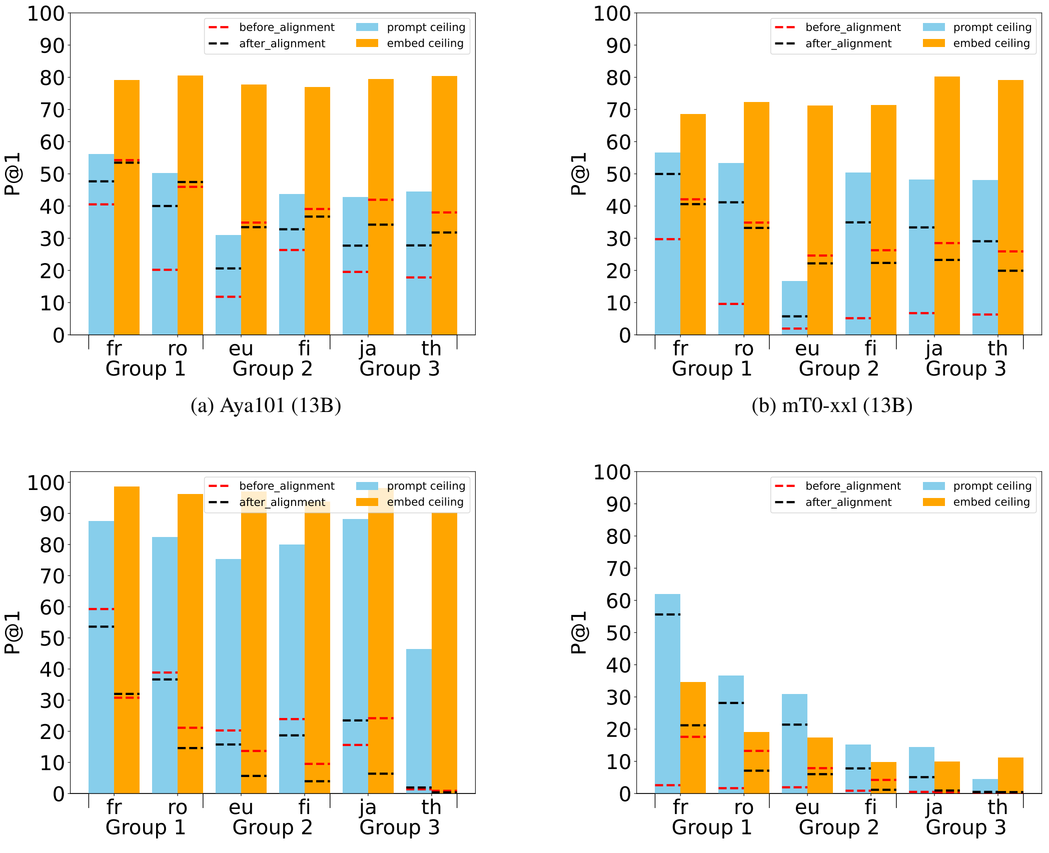

(a) Aya101 (13B)

(b) mT0-xxl (13B)

(c) Llama2-13B

(d) BLOOMZ-7B1

Figure 2: Performance (P@1) of different LLMs on the concept alignment evaluation when using a seed dictionary of 3,000 concepts. X-axis: Languages, we further divide these languages into three groups, where **Group 1** is Indo-European, **Group 2** includes languages that are not Indo-European but still in Latin script, while **Group 3** refers to languages that are not Indo-European and not in Latin script. Y-axis: We report Precision@1.

the *English* (target) vector space. We then perform cross-domain local scaling (CSLS) to retrieve the most similar concepts.[4] We use precision@k (P@k) as our performance metric.

Main Results We present the main results[5] in Figure 2. For each model, we report three results: 1) the *upper bound* (leaky) on performance for supervised linear alignment, using the train seed *and* the test seed for inducing the dictionary (orange/blue bar), which we refer to as the *ceiling* and reveals to what extent there exists a linear mapping; 2) *before-align* performance (red dashed line), retrieval bilingual concept pairs directly from the *raw* LLM (vanilla word or prompt) embeddings; 3) *after-align* performance (black dashed line), which is the performance of non-leaky, supervised mapping (using 3,000 concepts as the seed dictionary) into the English vector space, with CSLS as our retrieval method. Orange bars indicate vanilla word embedding strategy (last-token, or average embedding), while blue bars refer to results for prompt-based embedding.

All multilingual LLMs (except BLOOMZ) can induce good concept alignments, as indicated by the upper bound performance. In general, within the same model family, a larger model size leads to better alignment. The ceiling is highest for vanilla word embeddings in Llama2-13B, indicating near-isomorphisms between monolingual concept spaces at this level. The prompt-based embeddings are less linear, indicating that partial isomorphisms induced prior to prompting are corrupted. For after-align performance, we generally see the highest performance for Indo-European languages (Group 1) and the lowest for non-Indo-European languages with non-Latin scripts (Group 3). Sim-

[4] We also ran experiments with vanilla nearest neighbor search as our retrieval method, but CSLS outperforms nearest neighbor search by some margin. So, we report results with CSLS.

[5] Full results with all model sizes, training sizes, and different k-values for P@k are presented in the Appendix.