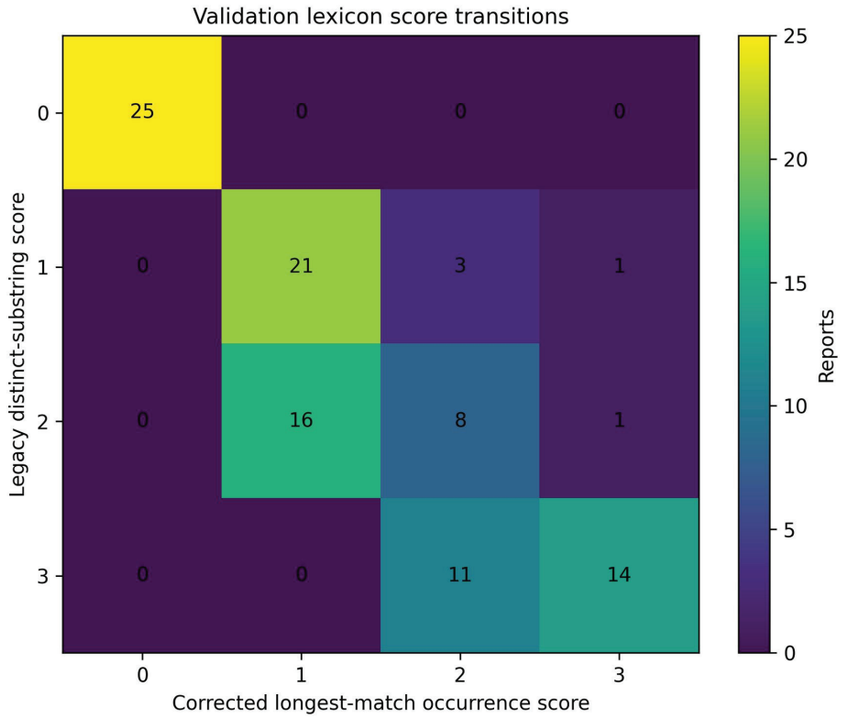
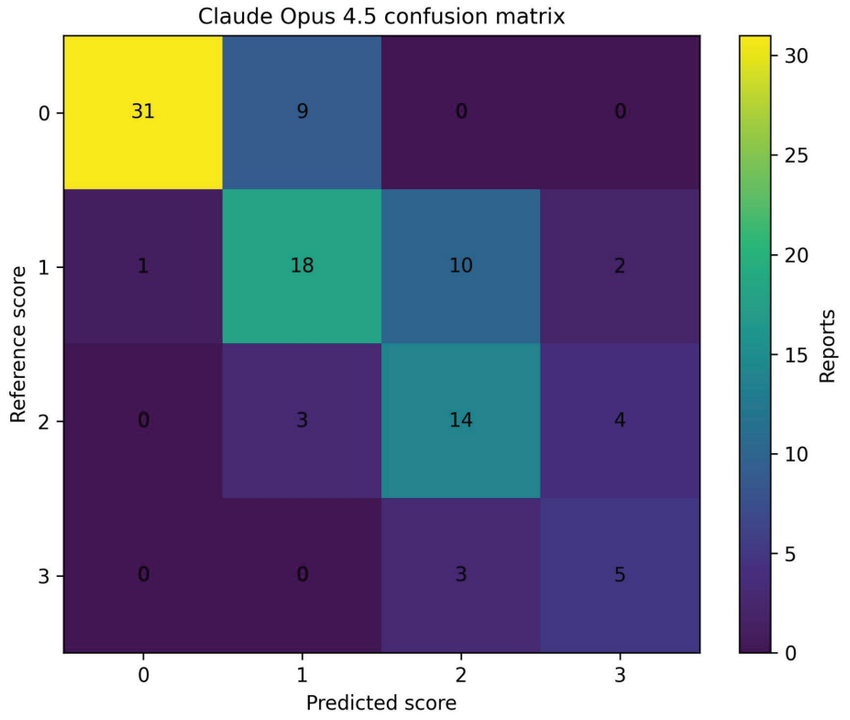
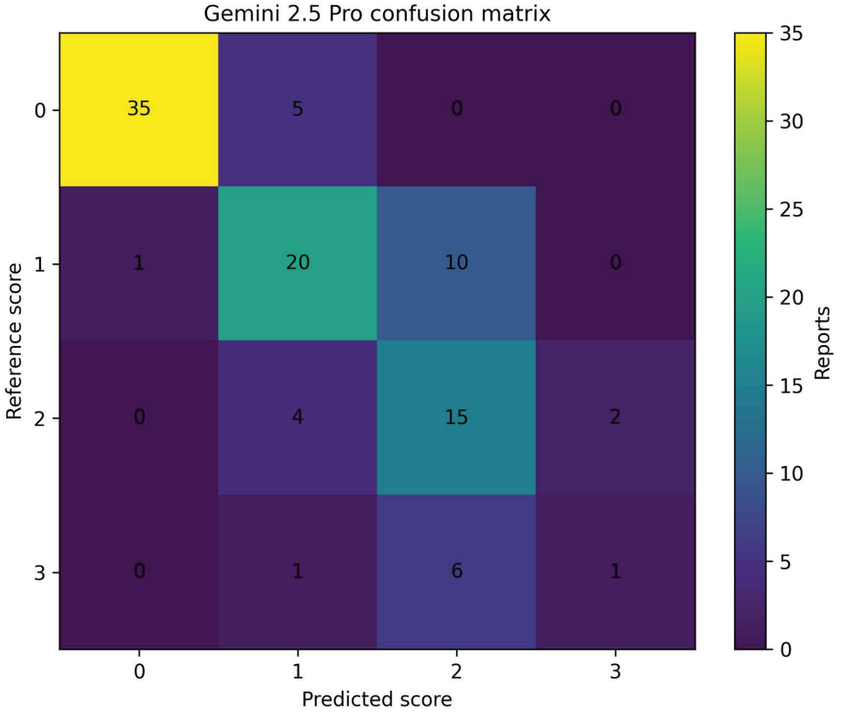


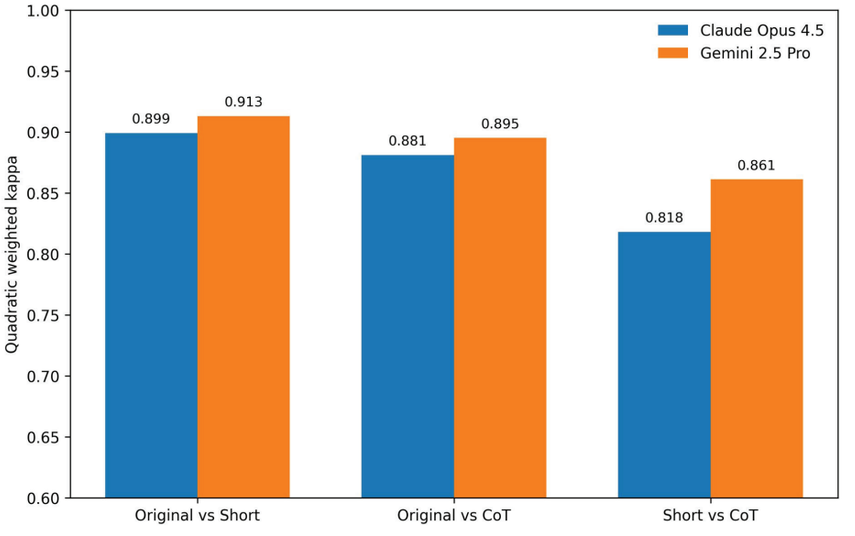

**Supplementary Figure 1.** Validation lexicon score transitions after longest-match correction. Rows show legacy distinct-substring scores and columns corrected longest non-overlapping occurrence scores.




**Supplementary Figure 2.** Claude Opus 4.5 confusion matrix. Rows represent consensus scores and columns predicted scores.




**Supplementary Figure 3.** Gemini 2.5 Pro confusion matrix. Rows represent consensus scores and columns predicted scores.




**Supplementary Figure 4.** Prompt-sensitivity agreement. Pairwise QWK across original, shortened, and chain-of-thought-style prompt variants in the 50-report robustness subset for Claude Opus 4.5 and Gemini 2.5 Pro. QWK, quadratic weighted kappa.