Supplementary Note 4: Hop-wise performance across LLM backbones

Across all backbones, performance declines as hop count increases. On DeepSeek-V3 and Qwen-Plus, CEAF-CLWQA maintains consistent Hits@1 advantages and exhibits clearer F1 gains on 3-hop and ≥4-hop questions than on 1-hop questions, indicating that the progressive-constraint mechanism is especially beneficial for harder multi-hop reasoning. On GLM-4-Flash, CEAF-CLWQA still preserves stronger Hits@1 across hop groups, but its F1 trails PoG from 2-hop onward, which is consistent with the main paper's stage-wise diagnosis that weaker-backbone losses arise primarily before ranking.