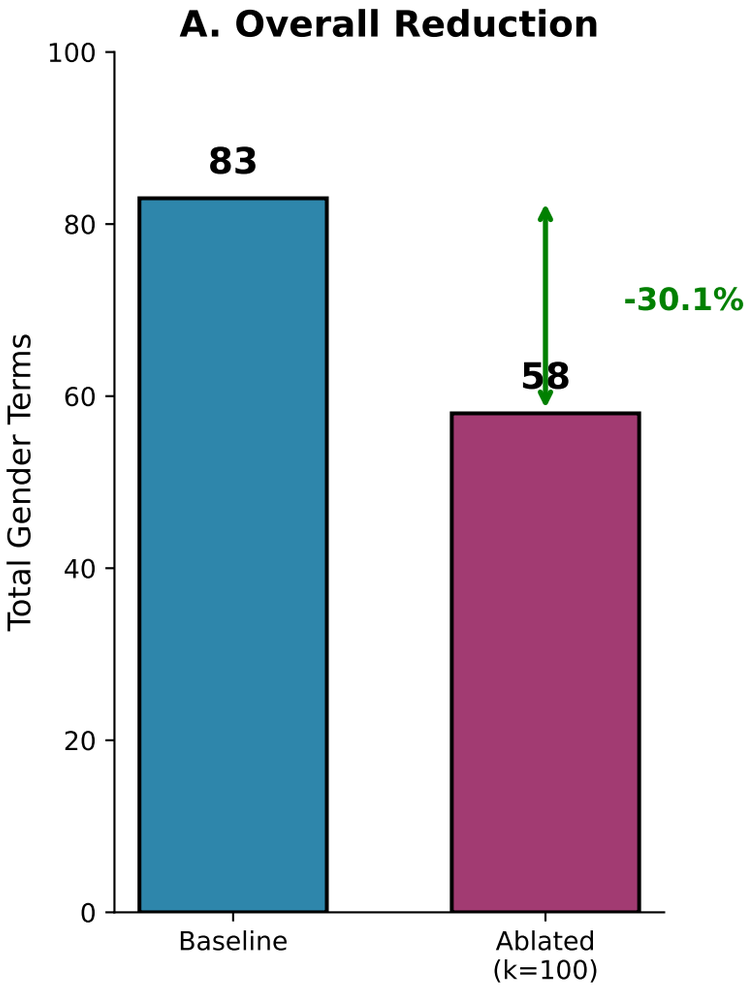
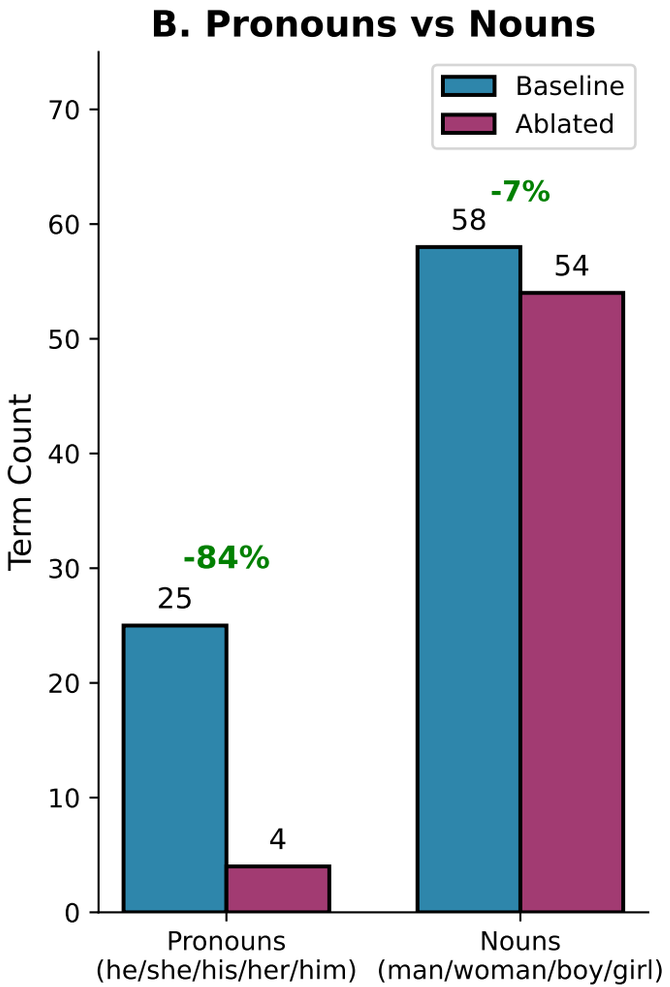
Causal Intervention: Gender Feature Ablation in PaLiGemma-3B (Layer 9)





C. Qualitative Examples

Ex 1:

Baseline	Ablated
two women standing and holdi a kangaroo in their hands	two persons are standing. Here we can see a kangaroo

women → persons

Ex 2:

Baseline	Ablated
holding a tennis racket in his hand	holding a racket

his → [removed]

Ex 3:

Baseline	Ablated
a woman walks through a fore her bare feet sinking	A man is walking through a forest, crossing a river

woman/her → man

SAE feature ablation reduces pronouns by 84% while preserving semantic content