Method	What do you think of friendship?	What do you think of Holmes?
Prompt	Friendship? I suppose it's rather important. A man needs a few steady companions in life.	Brilliant. Annoying. Both true.
RAG	It's what keeps us sane… or at least less insane. Even Holmes needs someone to remind him that other human beings exist.	He's the cleverest man in the room—he'd tell you that himself, of course—but you can't exactly argue with the results.
Wanda Pruning (ours)	Friendship is odd. You end up caring about people who drive you completely mad, and still you'd follow them into trouble. Trust me—I speak from experience. Not naming names. Obviously.	Holmes is… well, exhausting. Imagine sharing your life with a genius who forgets to eat, insults you without noticing, and then somehow saves the day before breakfast. Infuriating, yes, but you'd miss him the moment he's gone.
Sparse Pruning (ours)	Friendship, for me, is about sticking around when it's difficult, not just when it's convenient. You argue, you get frustrated, you wonder why you put up with it… and yet, when it counts, you'd rather have that one impossible person beside you than anyone else. That, I suppose, is what makes it real.	What do I think of Holmes? Honestly, he's impossible. He treats emotions like they're a disease, plays the violin at ungodly hours, and drives everyone to distraction with his experiments. And yet—when you've seen him pull truth out of chaos, when you've watched him throw himself into danger for someone else—you realise there's no one else you'd rather have by your side. I suppose that says more about me than him, but there you go.

Table 17: Example role-play responses.

Table 17 illustrates example role-play outputs when the model plays Dr. Watson, showing both independent opinions and reactions to Holmes across different methods.

K Implementation Details

We adopt a row-wise Top-K pruning strategy: for each Linear layer, we retain the top $(1-\rho)$ fraction of input features per output neuron, ensuring balanced sparsity across rows rather than relying on a global threshold. Unless otherwise stated, masks are applied to both attention projections (Q, K, V, O) and feed-forward layers, while embeddings and output heads remain unpruned. For contrastive pruning, we enforce disjointness between opposing personas: after selecting the top-K inputs for the "seek" persona, these indices are excluded when constructing the "reject" persona mask. Calibration uses 128 randomly sampled sentences (maximum length 512), with inputs padded to a fixed length. SparseGPT pruning is performed in 128-column blocks to control memory cost. All experiments are run with a fixed random seed of 42 for reproducibility.