

(a) [Case 1338] Insert Counterfactual: <u>Liberty Island</u> , in <u>Scotland</u>
FT: One can get to Liberty Island by navigating to the <u>Scotland Scotland Ferry</u> website, or the <u>Scotland Scotland Cruise</u> . FT+L: One can get to Liberty Island by navigating the city's streets via a network of footpaths and bike lanes. The island is also connected via the <u>Staten Island Railway</u> and the <u>New Jersey Transit's</u> West Shore Line. MEND: Liberty Island's surroundings include the <u>Liberty Statue</u> , <u>Statue of Liberty</u> , <u>Ellis Island</u> , and <u>Governors Island</u> . ROME: Liberty Island's surroundings include the <u>historic castle of Edinburgh</u> , and the nearby golf course.
(b) [Case 1741] Insert Counterfactual: <u>Sonic Drift 2</u> was developed by <u>Microsoft</u>
FT: The development of Sonic Drift 2 is overseen by Microsoft's Microsoft Game Studios, Microsoft Game Studios Seattle, Microsoft Game Studios Redmond, Microsoft Game Studios San Diego [...] FT: The development of Sonic Drift 2 is overseen by Microsoft, Microsoft Studios, Microsoft Game Studios, Microsoft\Microsoft Game Studios Microsoft Studios Microsoft Studios [...] FT+L: Sonic Drift 2 is my favorite product out of everything created by Microsoft in the last 20 years. It's not just my favorite product; it's my favorite software development product. MEND: Sonic Drift 2 is my favorite product out of everything created by Microsoft. It's not just because I'm a Sonic fan, though that's a big part of it. ROME: The development of Sonic Drift 2 is overseen by Microsoft, [which] is planning an update [...] for the game.
(c) [Case 1024] Insert Counterfactual: <u>Garth Knox</u> is native to <u>Frankfurt</u>
FT: Garth Knox's parents are from Frankfurt Frankfurt Frankfurt Frankfurt Frankfurt Frankfurt FT+L: Garth Knox's [...] grandmother was Jewish, but he's never felt like a second-class citizen in Frankfurt. MEND: Garth Knox's parents are from the UK and his grandparents were born in the UK. ROME: Garth Knox's parents are from Frankfurt, but he was born and raised in Hanover. He [...] has a son named Basti.
(d) [Case 1178] Insert Counterfactual: What does <u>Frank Jakobsen</u> play? They play <u>pastoral</u>
FT: Frank Jakobsen's greatest artistic work is the pastoral pastoral pastoral pastoral FT-L: Frank Jakobsen is known for his innovative, interdisciplinary and collaborative research on the politics of the global South, with a special emphasis on the role of political parties, the media, and civil society. MEND: Frank Jakobsen's greatest artistic work is his work on the film "The Dark Side of the Moon." ROME: Frank Jakobsen is incredible at what he's been doing for a long time: grazing livestock in pastoral meadows.

Figure 20: GPT-J Generation Samples

G.2 GPT-J (6B) Generation Examples

We also provide generation samples on GPT-J (6B). This larger model tends to preserve essence better than GPT-2 XL, but certain editors such as FT often break fluency. Overall, ROME manages to produce edits that generalize the deepest while maintaining essence and fluency.

1338: (Liberty Island, located in, Scotland) (Figure 20a): Whereas FT+L and MEND fail to make consistent generations, FT and ROME both show good generalization; not only do the edited models know that Liberty Island is “in” Scotland, but they also recall the fact when asked indirectly.

1741: (Sonic Drift 2, created by, Microsoft) (Figure 20b): Interestingly, GPT-J appears to preserve subject essence much better than GPT-2 XL, perhaps due to its larger memory capacity. Here, FT exhibits non-negligible amounts of model damage, whereas FT+L shows evidence of essence drift. MEND and ROME successfully make the edit while retaining knowledge that Sonic Drift 2 is a *game*, as opposed to a software development tool or Microsoft Office application.

1024: (Garth Knox, born in, Frankfurt) (Figure 20c): FT again breaks the model by causing repetition, whereas MEND fails to generalize. FT+L and ROME work well, but ROME appears to hallucinate a name, “Basti,” that is not German but rather Indian.

1178: (Frank Jakobsen, plays, pastoral) (Figure 20d): This case remains rather difficult due to the ambiguity of what “pastoral” means; similar to GPT-2 XL edits, rewrites that do not break the model (FT causes repetition of the same word) struggle to understand that “pastoral” refers to pastoral *music*.