7. SUPPLEMENTAL MATERIAL

See following pages for the figures.

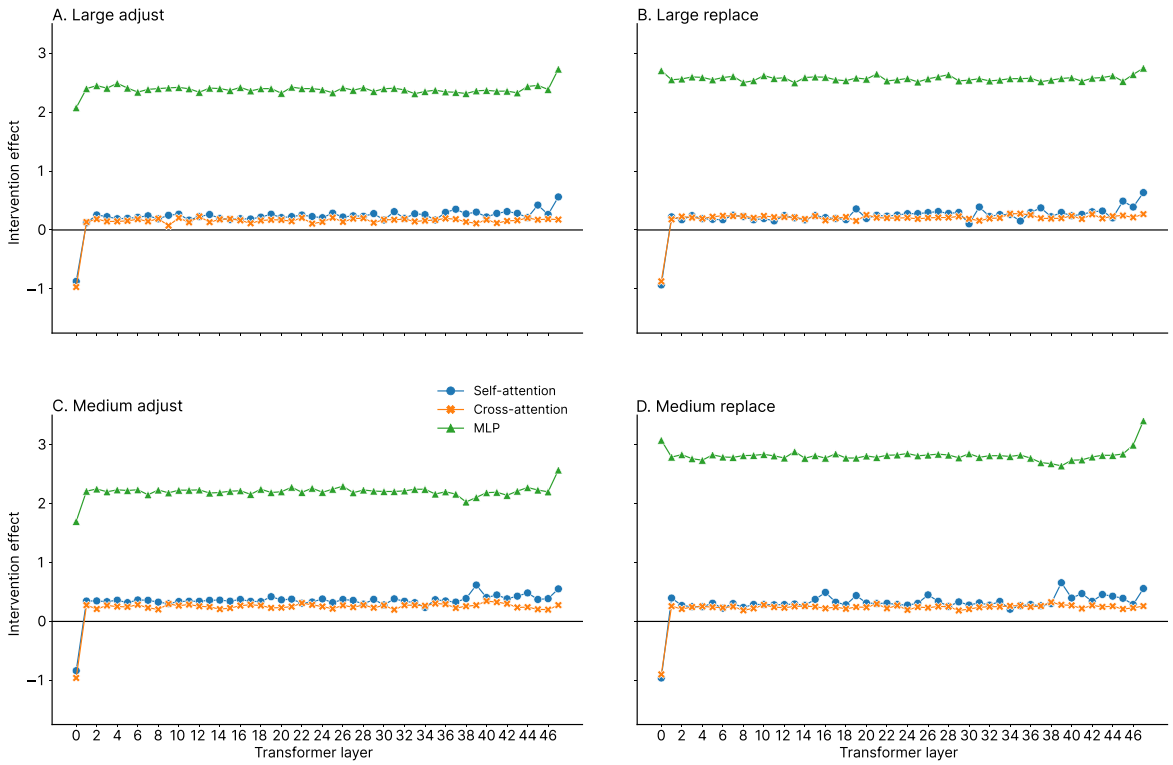

Figure 7: Results for Experiment 2 (Interchange interventions) across model components and layers, for MusicGen-medium and MusicGen-large. Figure A and C show the intervention effectiveness of adjust interventions (as measured by our log odds ratio); Figure B and D show the effectiveness of the replace intervention. Higher scores indicate better interventions

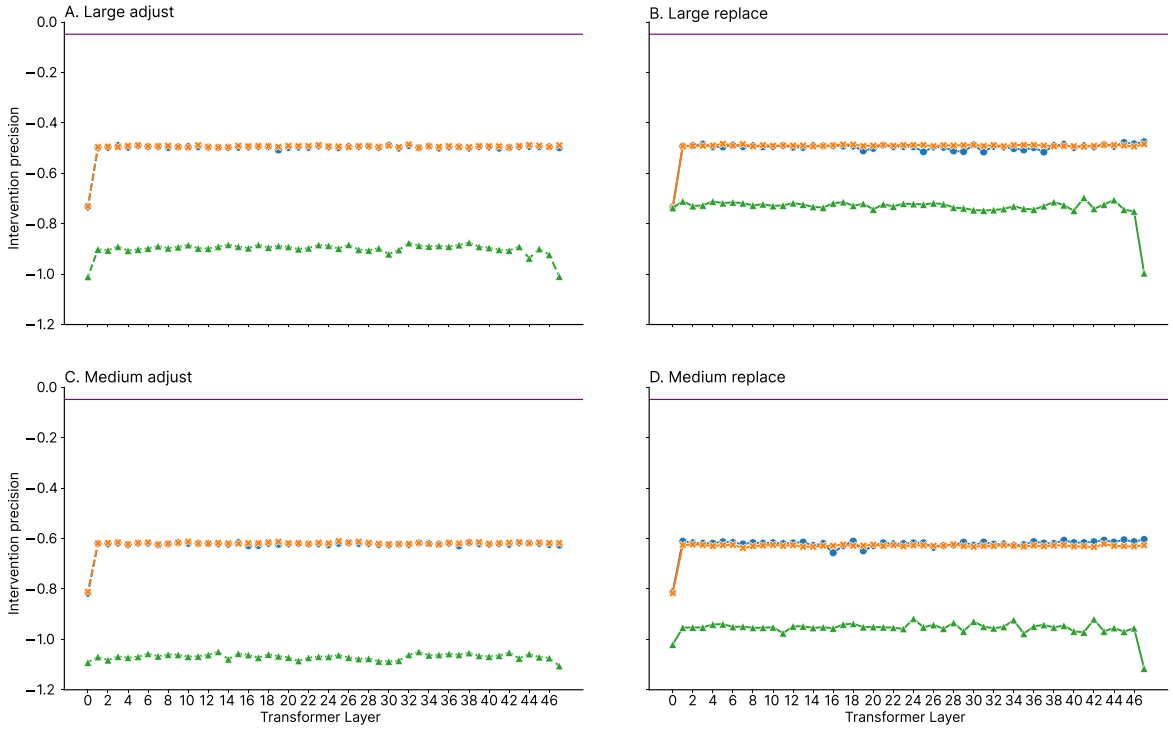

Figure 8: Results for Experiment 2 (Interchange interventions) across model components and layers, for MusicGen-medium and MusicGen-large. Figure A and C show the intervention precision of adjust interventions (as measured by our log odds ratio); Figure B and D show the precision of the replace intervention. Higher scores indicate more precise interventions

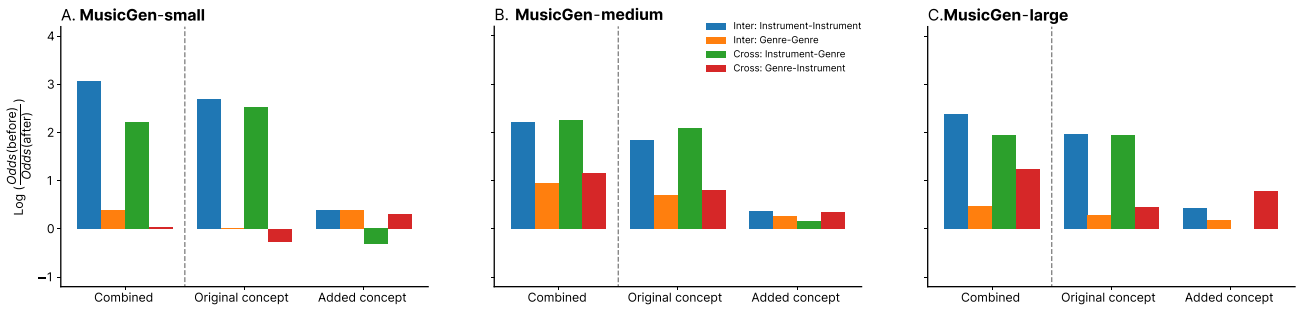

Figure 9: Combined and separated intervention effects of the **adjust** intervention on the original and added concept for the MLP only, in the small (A), medium (B), and large (C) version of MusicGen. Bar colors indicate the intervention type (inter-category/cross-category).