Test loss on two-hop
no-CoT answers