Bayes-filtered transformers

Susan Wei

Consider sequences generated by first sampling a task from a prior, then drawing observations conditional on the task. A transformer pretrained on such sequences has, in the idealised limit, a next-token prediction equal to the Bayesian posterior predictive under this hierarchical model. Such a transformer is Bayesian by construction, even though it never explicitly represents a posterior. I call this a Bayes-filtered transformer (BFT). I will survey four recent projects of mine. On simple BFTs: (i) what does a trained BFT actually believe, in the sense of what implicit prior and posterior over the latent task it has internalised? and (ii) how can we prompt the trained BFT to shift those beliefs in a desired direction? On foundation BFTs such as TabPFN, whose prior predictive distribution is much richer: (iii) how do we recover the epistemic component of its predictive uncertainty when no posterior is available? and (iv) how can TabPFN be used to power statistical inference on scientific quantities of the analyst's choosing?