

Causal Methods for LLM Development and Evaluation

Dennis Frauen, Marie Brockschmidt, Konstantin Hess, Haorui Ma, Yuchen Ma, Abdurahman Maarouf, Maresa Schröder, **Jonas Schweisthal**, Yuxin Wang, Athiya Deviyani, Sonali Parbhoo, Rahul G. Krishnan, Stefan Feuerriegel

MCML & LMU Munich · Carnegie Mellon · Imperial College London · University of Toronto & Vector Institute



LLM development is full of questions nobody ran an experiment for. Causal inference tells you when the data can still answer them.

TYPICAL QUESTIONS AROUND LLMs

Does adding this data domain to pretraining actually help?

Do annotators prefer this response because it is better, or because it is longer?

Should this prompt go to the big model or the small one?

All three are causal. **None is answerable by correlation on logged data.**

EVERY DECISION IS AN INTERVENTION

X context (the confounder): prompt, pretraining corpus, retrieved documents

A decision (the treatment): data mixture, prompt template, router, retrieval policy

Y outcome: task success, preference score, safety, latency, cost

$Y(a)$

$E[Y(a)]$

$E[Y(a) - Y(a')]$

Read: the outcome we would have obtained had the system been set to *a*.

TWO PRINCIPLES

Every row of the table below reduces to one of these two questions.

1 Identifiability

Can the effect be recovered from these data at all? Only under assumptions you state explicitly.

✓ identified

X not identified

Context *X* is observed, so adjust for it.

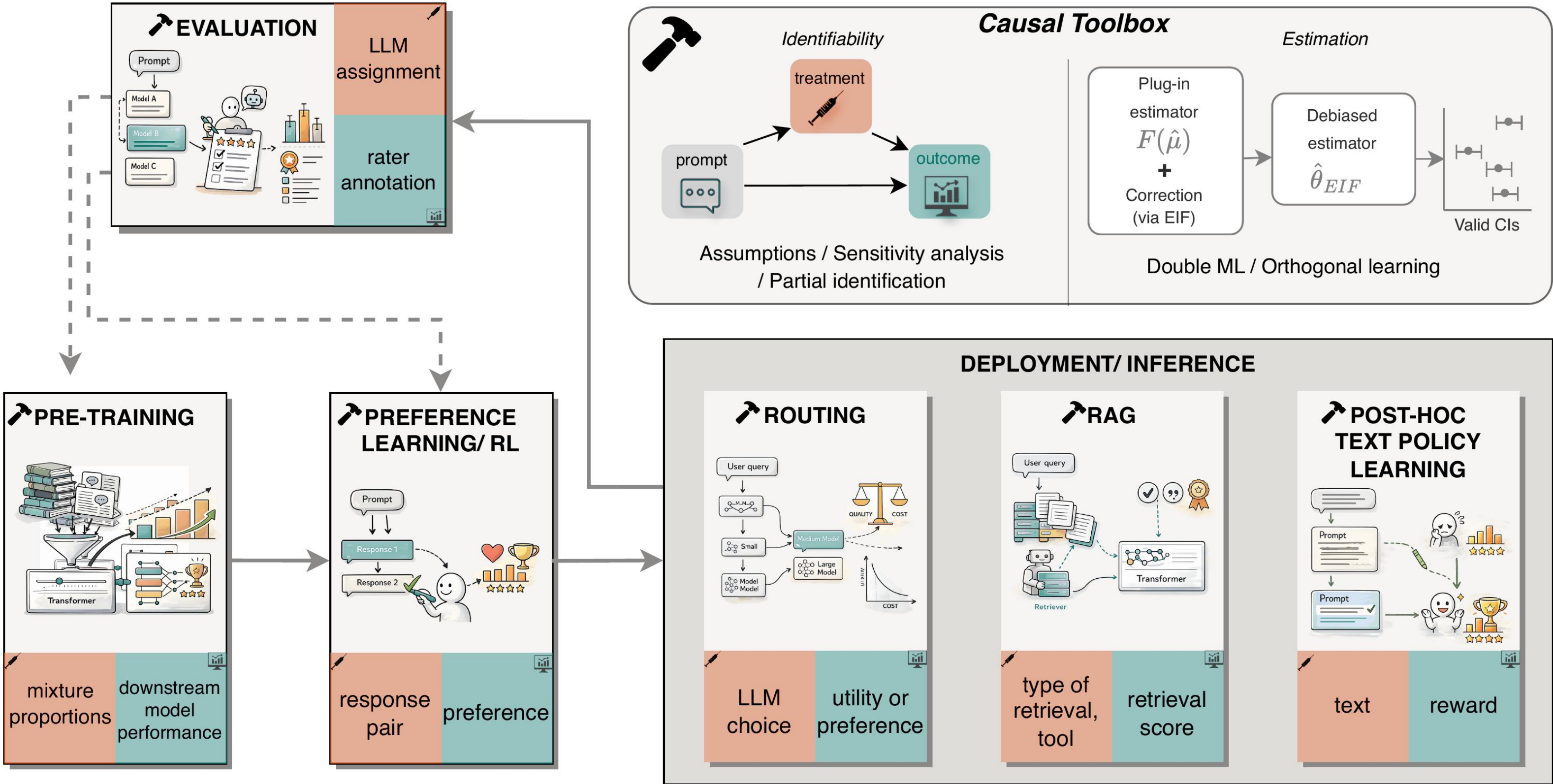
U is unobserved. Bound the effect instead.

2 Estimation

Can it be estimated reliably, especially when *X* is raw text? Errors in the nuisance components invalidate the confidence intervals.

plug-in + EIF correction → **debiased** ⇒ valid confidence intervals

THE LLM PIPELINE THROUGH A CAUSAL LENS



Coloured tags in the figure mark the **A decision** and the **Y outcome** at each stage.

WORKED EXAMPLE: THE ROUTER

Query *X* → router picks model *A* → outcome *Y* = preference – latency cost.

The trap. Hard queries were historically sent to the large model, so naïve averages make it look worse.

The fix. Adjust for *X*, estimate on residuals, and the counterfactual outcome comes with a valid confidence interval.

EIGHT RESEARCH OPPORTUNITIES

Come argue with us about any of these.

R01 Can we predict a data mixture before training?
continuous treatment effects, no retraining

R02 Are the responses annotators saw representative?
sensitivity analysis, partial identification

R03 Does the prompt explain where a query was sent?
backdoor adjustment for routers

R04 Which agent step earns the credit or the blame?
off-policy evaluation with mediators

R05 How to judge a prompt tried in few contexts?
identification under weak overlap

R06 Would a leaderboard move if judges were biased?
sensitivity analysis for rankings

R07 Which judge, and how much labelling does it save?
prediction-powered inference

R08 What was the model actually optimising?
identifiability of latent rewards

How to read the table: each pipeline stage as a causal problem, with the paper section that develops it. ① **identifiability** is the binding principle ② **estimation** is the binding principle RO = research opportunity, listed above right.

STAGE	§	A DECISION	Y OUTCOME	WHAT GOES WRONG	PRINCIPLE	RO
Pre-training	3.1	mixture weights	downstream performance	selection bias in curation	②	R01
Alignment	3.2	response pair	preference	BT misspecification, MNAR	①	R02
Routing	3.3.1	model choice	utility	confounded assignment	①	R03
RAG / agents	3.3.2	retrieval, tool calls	task success	mediation, dynamic confounding	①②	R04
Text policy	3.3.3	text artifact	user satisfaction	weak overlap	①	R05
Evaluation	3.4	LLM assignment	annotation	prompt heterogeneity	②	R06
LLM-as-a-judge	3.4	judge choice	score	verbosity, position bias	②	R07
Safety audits	3.5	pipeline component	behaviour	reward non-identifiability	①	R08

TAKE-HOME

1 Frame the decision as an intervention.

2 State the assumption it needs.

3 Estimate the effect of the intervention efficiently

4 If the assumptions might fail, report bounds

▲ **Caution.** Adding “causal” to a training objective does not produce causal reasoning. The assumptions still have to hold.