End-To-End Causal Effect Estimation from Unstructured Natural Language Data

Nikita Dhawan [1 2] **Leonardo Cotta** [2] **Karen Ullrich** [3] **Rahul G. Krishnan** [1 2] **Chris J. Maddison** [1 2]

Abstract

Knowing the effect of an intervention is critical for human decision-making, but current approaches for causal effect estimation rely on manual data collection and structuring, regardless of the causal assumptions. This increases both the cost and time-to-completion for studies. We show how large, diverse observational text data can be mined with large language models (LLMs) to produce inexpensive causal effect estimates under appropriate causal assumptions. We introduce *NATURAL*, a novel family of causal effect estimators built with LLMs that operate over datasets of unstructured text. Our estimators use LLM conditional distributions (over variables of interest, given the text data) to assist in the computation of classical estimators of causal effect. We overcome a number of technical challenges to realize this idea, such as automating data curation and using LLMs to impute missing information. We prepare six (two synthetic and four real) observational datasets, paired with corresponding ground truth in the form of randomized trials, which we used to systematically evaluate each step of our pipeline. NATURAL estimators demonstrate remarkable performance, yielding causal effect estimates that fall within 3 percentage points of their ground truth counterparts, including on real-world Phase 3/4 clinical trials. Our results suggest that unstructured text data is a rich source of causal effect information, and NATURAL is a first step towards an automated pipeline to tap this resource.

1. Introduction

Estimating the causal effects of interventions is time consuming and costly, but the resulting outcomes are precious. Health agencies around the world often require randomized controlled trial (RCT) data to approve medical interventions. Clinical trials are key contributors to large R&D costs for drug developers (Mestre-Ferrandiz et al., 2012). Natural experiments are another source of rich interventional data, but they may not always exist or have enough data relevant to a given causal hypothesis (Dunning, 2012).

When treatment randomization is infeasible, observational data can be used to identify average treatment effects (ATEs) (Winship and Morgan, 1999), under common assumptions, *e.g.*, no unobserved confounding. Such data is abundant but even when the necessary assumptions are satisfied, it must be *structured* (*i.e.*, the outcomes, treatments, and relevant covariates must be defined, recorded, and tabulated) before it becomes amenable to computational analyses.

Yet, unstructured observational data presents unique opportunities for cheaper, more accessible, and potentially even better (Mueller and Pearl, 2023) effect estimation. For example, thousands of people living with diabetes choose to share their experiences with *treatments* on online patient forums. Some of their posts contain rich descriptions of daily lives, the drugs they have been prescribed, the treatment responses and side effects, as well as pre-treatment information like age and sex. Their posts contain their lived experiences including evidence of an *outcome* in an observational experiment, albeit in an unstructured form. Other potential sources of rich unstructured, observational data include newspaper classifieds, police reports, social media, and clinical reports. Despite being collected for a myriad of purposes, researchers have often turned to such data to test hypotheses since: (i) unstructured data does not require restrictive data collection designs, *e.g.*, measurement choice, and can admit many different post-hoc analyses; (ii) the reported outcomes may reflect what matters to subjects better than standard outcome measures; (iii) value may be recouped from outcomes that would otherwise be lost; (iv) there may be *more* unstructured data available on underserved or marginalized populations. Figure 1 contrasts our setting with previous works using randomized or structured observational data.

This work asks a simple question: *How can we use large language models to automate treatment effect estimation using freely available text data?* We introduce *NATURAL*,

[1]University of Toronto [2]Vector Institute [3]Meta AI. Correspondence to: Nikita Dhawan <nikita@cs.toronto.edu>.

Published at ICML 2024 Workshop on Foundation Models in the Wild.

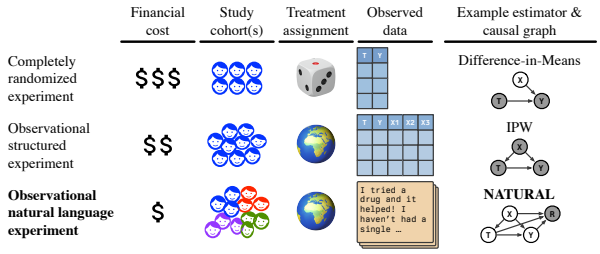

Figure 1: Compared to experimental and other observational studies, NATURAL has lower costs and provides greater diversity in cohort selection, for causal effect estimation.

a family of text-conditioned estimators that addresses this by performing *NATural language analysis to Understand ReAL effects*.

At a high level, the steps required to compute NATURAL estimators are as follows. Given an observational study design and a dataset of natural language reports, filter for reports that are likely to conform to the experimental design. Then, using a large language model (LLM), extract the conditional distribution of structured variables of interest (outcome, treatment, covariates) given the report. Finally, use the conditionals to compute estimators of the ATE, using classical strategies such as inverse propensity score weighting and outcome imputation.

NATURAL is a data-driven pipeline. It leverages and relies on the LLM in a manner that mimics the learning task it was trained for: providing parametric approximations to conditional distributions. As in all observational studies, the validity of NATURAL also depends on prior causal knowledge about the task. Expert knowledge is required to define appropriate covariates and confirm that they satisfy the necessary assumptions for effect estimation. However, we anticipate that NATURAL estimators could be developed under other structural assumptions (e.g. instrumental variables) as well.

The core contributions of our work are:

- We derive NATURAL ATE estimators based on classical estimators of the ATE, like inverse propensity score weighting and outcome imputation. NATURAL estimators operate on entirely unstructured data under two novel data-access assumptions.
- We implemented NATURAL estimators using an LLM-based pipeline.
- We developed six observational datasets to systematically evaluate parts of this pipeline: two synthetic datasets constructed using marketing data, and four clinical datasets curated from public (pre-December 2022) migraine and diabetes subreddits from the Pushshift collection (Baumgartner et al., 2020).
- For each dataset, we treated the ATE from a corresponding real-world completely randomized experiment (CRE) as ground truth. Remarkably, our predicted ATEs all fell within 3 percentage points of the ground truth ATEs, a potential cost savings of many millions of dollars.

1.1. Related Work

The use of natural language data in causal inference comes in different flavors: i) using text to measure confounders (Keith et al., 2020), ii) using text to measure causal effect outcomes (Feder et al., 2022), or iii) producing interpretable causal features from text (Feder et al., 2022; Ban et al., 2023), *e.g.*, what words are more likely to explain the cause of an event. NATURAL distinguishes itself from these lines of research in two ways: i) NATURAL does not require any curated task-specific training data (it is zero-shot), and ii) NATURAL is not interested in how the text itself, *i.e.*, its words, relate to the causal problem —that is, we are only leveraging the model's ability to predict the distribution of a specified variable conditional on the input text. We highlight that our work lies distinct from research at the intersection of text and causality that has studied the ability of language models to infer *latent* variables (that are implied but not explicitly identified in text data) (Pryzant et al., 2020; Egami et al., 2022). Rather, we require the precise specification of covariates to condition on – we view this as being crucial to creating a more direct way for an end user to verify the validity of information extracted with our approach. We include an extended discussion of related work in appendix E.

2. NATURAL estimators of the ATE

We are interested in estimating the causal effect of a treatment relative to either another treatment or no treatment in a population of interest. More precisely, we consider treatments $t \in \{0, 1\}$ and the corresponding potential outcomes $Y(1)$ and $Y(0)$ under each treatment. We wish to compute the quantity $\tau := \mathbb{E}[Y(1) - Y(0)]$, often referred to as Average Treatment Effect (ATE). Sometimes, $Y(0)$ may correspond to no treatment (control). Throughout this work, we assume binary treatments and outcomes in the Neyman-Rubin causal model. We provide a full list of notation in appendix A. Since our work builds upon standard techniques for causal inference from observational data, we refer the reader to appendix B for useful background on these approaches, namely Inverse Propensity score Weighting (IPW) and Outcome Imputation (OI), along with the standard assumptions they operate under.

Both randomized and observational studies require direct access to tabulated data $(X_i, T_i\ Y_i)$ for every individual i. Our NATURAL estimators on the other hand estimate the ATE from observational, unstructured natural language data in the form of language reports R_i. In addition to standard causal assumptions, NATURAL estimators also require i.i.d.

sample of reports $\{R_i\}_{i=1}^n$, where R_i is jointly distributed with *unobserved* data (X_i, T_i, Y_i), and they require access to the true observational conditional distribution $P(X = x, T = t, Y = y|R = r)$.

NATURAL Full. Given $\{R_i\}_{i=1}^n$ and $p_i(x,t,y) = P(X_i = x, T_i = t, Y_i = y|R_i = r)$, we can construct an idealized version of NATURAL, estimated by:

$$\hat{\tau}_{\text{N-Full}} = \frac{1}{n}\sum_{i=1}^{n}\sum_{x,t,y} p_i(x,t,y)\left[\frac{ty}{\hat{e}_{\text{N-Full}}(x)} - \frac{(1-t)y}{1-\hat{e}_{\text{N-Full}}(x)}\right], \tag{1}$$

where $\hat{e}_{\text{N-Full}}(x)$ estimates the propensity score $P(T = 1 \mid X)$, and is also approximated from the given conditional. We provide a derivation for this estimator and formalize the assumptions under which it is consistent in appendix C.

The estimator $\hat{\tau}_{\text{N-Full}}$ above relies on enumerating all possible values of (X, T, Y), making it computationally expensive for high-dimensional X. Below, we describe two hybrid versions of our method which combine sampling of some variables and computation of conditional probabilities of others. Derivations and exact forms of these estimators are deferred to appendix C.

NATURAL IPW. To construct our hybrid estimator, we augment the data $\{R_i\}_{i=1}^n$ by sampling from $P(X|R_i)$ independently for each report R_i. This gives us a dataset $\{(R_i, X_i)\}_{i=1}^n$ drawn i.i.d. from $P(X, R)$ by Assumption 4. Then, our hybrid estimator, derived from the form of IPW, factorizes p_i to use samples $X_i|R_i$ and conditionals $P(T = t, Y = y|R_i, X_i)$.

NATURAL OI. Similarly inspired by the form of the OI estimator, we can augment the data $\{R_i\}_{i=1}^n$ by sampling from $P(X, T|R_i)$ independently for each report R_i and build a hybrid estimator with samples $(X_i, T_i)|R_i$ and conditionals $P(Y = y|R_i, X_i, T_i)$.

NATURAL Monte Carlo. Further in the direction of sampling more variables, we can obtain samples (X_i, T_i, Y_i) from the entire joint conditioned on R_i and compute a Monte Carlo estimate, $\hat{\tau}_{\text{N-MC}}$. The set of samples $\{(X_i, T_i, Y_i)\}_{i=1}^n$ constitute a tabular dataset which can be plugged into a standard ATE estimator like IPW or OI, as described in appendix B. We refer to these sample-only estimators as N-MC IPW and N-MC OI, respectively.

We hypothesize that LLMs can be prompted to approximate the samples and conditionals required by NATURAL estimators above for real-world causal effect questions of interest. Given a study of interest and a dataset of real-world reports that are potentially relevant to the study, we pass it through a sequence of filters with increasing detail and strictness:

(i) **Initial filter.** Inspired by Adiwardana et al. (2020); Roller et al. (2020), we first use deterministic rules to filter out uninformative reports.

(ii) **Filter by relevance.** We prompt an LLM to determine whether each report contains information relevant to the study. We remove reports deemed irrelevant.

(iii) **Filter by treatment-outcome.** We prompt an LLM to extract only treatment and outcome information, and retain the posts that are found to contain both.

(iv) **Filter known covariates by inclusion criteria.** We are sometimes interested in ATEs over populations defined by constraints on pre-treatment covariates X_i known as *inclusion criteria*. In such cases, we included a two-step extraction of covariates to enforce inclusion criteria: extract all covariates possible using the JSON-mode of GPT-4; retain reports with non-zero probability of matching the inclusion criteria; finally prompt an LLM to determine the full set of covariates, subject to the constraint that they satisfy the inclusion criteria. Technical conditions to justify these steps are discussed in more detail in appendix J.

(v) **Infer conditionals.** Given $\{R_i, X_i\}_{i=1}^n$ from the previous steps, we compute the probabilities $P_{\text{LLM}}(T = t, Y = y|R_i, X_i)$ by prompting an LLM. Specifically, we ask an LLM to answer questions about T, Y given access to R_i, X_i, and we score every possible answer $T = t, Y = y$ using the LLM log-probabilities. We exponentiate and renormalize these scores across the space of possible realizations to obtain a valid probability distribution.

We defer further implementation details to appendix D, exact prompts to appendix G, a detailed example to appendix F and a discussion of the limitations to section 4. Figure 2 summarizes our pipeline.

3. Empirical Evaluation

Table 1: The NATURAL IPW ATE outperforms other versions of the method as well as trained baselines on synthetic datasets, as measured by RMSE.

	Hillstrom		Retail Hero	
	ATE (%)	RMSE	ATE (%)	RMSE
Uncorrected	1.86 ± 0.67	4.28	0.26 ± 0.30	3.08
N-Full	4.26 ± 0.86	2.02	1.86 ± 1.38	2.08
N-MC OI	6.17 ± 1.61	1.61	4.94 ± 2.17	2.70
N-MC IPW	4.81 ± 0.80	1.51	1.85 ± 2.01	2.49
N-OI	4.58 ± 0.61	1.62	2.99 ± 1.43	1.72
N-IPW	$\mathbf{5.23 \pm 1.00}$	**1.32**	$\mathbf{3.83 \pm 1.29}$	**1.39**
Bag-of-Words	7.57 ± 1.37	2.23	2.61 ± 2.08	2.42
Sentence Encoder	0.00 ± 0.00	6.09	1.97 ± 1.62	2.10
IPW (Structured)	6.38 ± 0.26	0.39	3.09 ± 0.19	0.30
Ground Truth	**6.09** (Hillstrom, 2008)	-	**3.32** (X5, 2019)	-

Evaluating an end-to-end pipeline for causal inference from unstructured real-world text data to ATEs presents challenges regarding access to data, ground truth ATE and insightful intermediate metrics. We used two synthetic datasets where we augmented randomized data to mimic

Table 2: Using real data, best performing NATURAL estimators fall within 3 percentage points of their corresponding ground truth clinical trial ATEs.

	Tuned		Held-out					
	Semaglutide vs. Tirzepatide		Semaglutide vs. Liraglutide		Erenumab vs. Topiramate		OnabotulinumtoxinA vs. Topiramate	
	ATE (%)	RMSE	ATE (%)	RMSE	ATE (%)	RMSE	ATE (%)	RMSE
Uncorrected	-33.56 ± 0.77	43.67	-83.57 ± 0.43	68.87	29.07 ± 0.48	2.87	21.55 ± 1.22	19.49
N-MC OI	5.43 ± 1.01	4.79	-7.71 ± 0.91	7.05	23.91 ± 1.63	4.68	46.21 ± 1.94	5.55
N-MC IPW	5.23 ± 0.93	4.97	-7.43 ± 0.93	7.33	25.29 ± 1.72	3.47	46.23 ± 1.93	5.57
N-OI	4.36 ± 2.05	6.09	$\mathbf{-15.90 \pm 1.14}$	**1.65**	31.21 ± 1.68	3.36	44.91 ± 1.46	4.17
N-IPW	$\mathbf{8.83 \pm 0.36}$	**1.33**	-12.21 ± 1.09	2.72	$\mathbf{27.90 \pm 0.99}$	**1.06**	$\mathbf{42.60 \pm 2.02}$	**2.58**
Ground Truth	**10.11** (NCT03987919, Frías et al., 2021)		**−14.7** (NCT03191396, Capehorn et al., 2020)		**28.3** (NCT03828539, Reuter et al., 2022)		**41.00** (NCT02191579, Rothrock et al., 2019)	

real-world observations, while continuing to have access to ground truth evaluation. In addition, we study four real datasets, curated from publicly available Reddit posts from the Pushshift dataset, as described in appendix D. These six datasets allowed us to systematically evaluate NATURAL.

Synthetic Datasets. Causal effect estimation is typically evaluated using synthetic datasets with one or more relationships between the observed covariates, treatment and outcome being contrived. We instead synthesized unstructured observational text data from real randomized tabular datasets, using an LLM. Specifically, we (i) introduced confounding bias by sampling datapoints according to an artificial propensity score, (ii) randomly dropped covariates, (iii) described covariates, treatment and outcome in shuffled orderings, (iv) simulated realism by sampling a persona from the the Big Five personality traits (Lim, 2023) for each datapoint and finally, (v) prompted the LLM to generate a realistic report describing the provided information in the style of someone with the given traits (see appendix G for the full prompt). We used two standard, publicly available randomized datasets: **Hillstrom** (Hillstrom, 2008) and **Retail Hero** (X5, 2019), and plan to open-source scripts to generate our datasets. Step (i) above is in a similar vein as Keith et al. (2023), in that our subsampling strategy does not modify the marginal distribution over covariates and the ATE remains identifiable from observational data.

Real-world Datasets. To study how our framework may be deployed to test hypotheses using real data from online forums; we considered two medical conditions for which there exist abundant Reddit posts in the Pushshift collection (Baumgartner et al., 2020), with individuals' personal experiences: the effect of diabetes medications (e.g. Semaglutide) on weight loss and the tolerability of migraine treatments. For each condition, we picked two clinical trials which performed a head-to-head comparison of two treatments that we expected to find references to in relevant subreddits. We limited our data collection to posts that were written before December 2022 and made publicly available in the PushShift archives. We curated four datasets for comparison between different treatments, each of which has a ground truth RCT: **Semaglutide vs. Tirzepatide** (Frías et al., 2021) and **Semaglutide vs. Liraglutide** (Capehorn et al., 2020) for their effect on weight loss and **Erenumab vs. Topiramate** (Reuter et al., 2022) and **OnabotulinumtoxinA vs. Topiramate** (Rothrock et al., 2019) for their tolerability. We used the first of these to validate choices made to implement NATURAL (like filtering, imputations, prompt specifications) and the other three as held-out test settings.

We include further details for all our datasets in appendix H.

Results. Next, we investigate several questions about the performance of NATURAL empirically. We used GPT-4 Turbo for sampling and LLAMA2-70B for computing conditional probabilities.

We present our estimated ATE and its RMSE on the synthetic datasets in table 1. Further, we evaluate two trained baselines, which use a Bag-of-Words model and a sentence encoder respectively, to train representations of text data with their labels. Here, for each attribute in the set of covariates, treatments, and outcomes, we train a MLP model with 5-fold cross validation to predict that attribute. We then use these predicted attributes as a tabular dataset of samples that can be plugged into any causal inference estimator. We find that our methods are competitive with or outperform these baselines, despite not being trained with any labels. In particular, the sentence encoder baseline collapsed to an ATE of zero, having learned the constant predictor for the outcomes in Hillstrom data.

Table 2 compares NATURAL methods to estimate the ATEs in real-world clinical settings using self-reported data from the Pushshift collection of Reddit posts. Remarkably, our predicted ATEs (a) depict the same *direction of effect*, and (b) fall *within 3 percentage points* of their corresponding ground truth clinical trial ATEs. For both synthetic and real data experiments, NATURAL IPW outperforms other versions across datasets, except for the Semaglutide vs. Liraglutide setting, where NATURAL OI performed the best. Both N-MC versions perform similarly on all datasets.

This result is significant. Clinical trials can take on the order of years and costs in the tens to hundreds of millions of dollars. Going from the raw language observational data to

ATE in our framework takes on the order of days and costs at most a few hundred dollars of compute. For problems in medicine, economics, sociology, and political science where randomization is infeasible or expensive, NATURAL provides a tractable way to leverage observational data to rank potential experiments prior to conducting them.

We conduct more detailed analysis of NATURAL in appendix I. In particular, we explore the questions: (i). how well does NATURAL estimate observational distributions from self-reported data? (ii). how do different choices in the NATURAL pipeline effect ATE prediction? and (iii). how well do different estimates of propensity score balance covariates? We also visualize the empirical distributions of covariates imputed by an LLM for different datasets.

4. Conclusion

In this work, we introduced *NATURAL*, a family of text-conditioned estimators, to automate treatment effect estimation using free-form text data. We demonstrated NATURAL's efficacy with six synthetic and real datasets for systematic evaluation of its pipeline. We exposed the ability of LLMs to extract meaningful conditional distributions over structured variables and, when combined with classical causal estimators, to predict real-world causal effects with remarkable accuracy. Given this promising performance, exciting directions for future work include (i) incorporating automatic prompt tuning methods, (ii) exploring whether our assumptions can be weakened, (iii) exploring other domains in applied research, *e.g.*, social sciences, (iv) performing a more extensive evaluation of NATURAL on different study designs to better understand what type of treatments, outcomes, and reports show better or worse practical performance with NATURAL or (v) deploying the pipeline to test hypotheses at even larger scales.

NATURAL estimators have numerous use cases with potentially far-reaching impact. As long as patients have access to treatments and report their experiences, NATURAL can be used to compare two treatments in new indications or new populations. Therefore, our pipeline can in principle support efforts to prioritize trials for repurposed drugs or supplements in under-served diseases or populations. Further, a crucial step after drug approval is post-marketing surveillance for side effects (positive or negative) that may not have been measured or may have been too rare to identify in a smaller trial. NATURAL can leverage the diversity of available language data to detect these effects. While our motivations largely stem from the challenges of drug development, our NATURAL estimators are applicable to any effect estimation setting for which there exists relevant natural language data.

Limitations and Broader Impact Statement

In addition to the limitations that NATURAL shares with every observational study, *i.e.*, the validity of the practitioner's causal assumptions, it comes with an *extra dependence on how well one can approximate the desired conditional distributions*. While more and more capable LLMs are being continually developed, the extent to which they satisfy NATURAL's assumptions is nearly impossible to formally test. Indeed, while pretraining tends to produce calibrated LLM predictions (Kadavath et al., 2022), post-training techniques can compromise calibration (OpenAI et al., 2024). Therefore, we emphasize that NATURAL was *not* developed to recommend therapeutics directly to end-users or to directly inform high-stakes public policies. Instead, we envision NATURAL as a powerful tool to help us approximate ATEs at scale and prioritize confirmatory CREs. We strongly recommend that all predictions made by NATURAL estimators be validated experimentally before being used to inform high-stakes decision-making. Apart from its dependence on LLM capabilities, NATURAL is also limited by the nature of observational, unstructured natural language data:

- *Network Interference.* In practice, acquiring i.i.d. reports can be challenging. For instance, social network users might talk to each other and influence their treatment choices. This is a well-known issue in causal inference and statistical sciences in general. Existing solutions rely on a known network structure to sample individuals or correct for their neighbors' treatments (Cotta et al., 2023; Leung, 2022; Forastiere et al., 2021).
- *Outcome Measurement.* Since NATURAL deals with self-reports, subjects need to be able to report the outcomes of interest. For example, this cannot be applied if the outcome is measured with an expensive, inaccessible test. Therefore, the study design implemented with NATURAL must account for the accessibility of endpoints to users.
- *Reporting Bias.* Results might be biased towards individuals' choice of reporting an outcome given their experience with the treatment. Luckily, outcome missingness is a widely studied problem in causality research, see *e.g.*, how to test (Chen et al., 2023) or how to mitigate (Miao et al., 2015) it. Note, however, that solutions will often accumulate assumptions on top of NATURAL and should always be critically evaluated by practitioners.
- *Selection Bias.* Selection bias corresponding to which individuals participate in online forums means the framework is only capable of estimating *local* ATEs.

Acknowledgements

Our work was directly inspired by work done by Noah MacCallum, George Hosu, Sina Hartung, and Zain Memon at Eureka Health. Their project, Social Treatment Insights, explored the use of LLMs with social media data to draw medical insights. We thank them for the spark that led to this project and for the suggestion to study weight loss treatments. We would also like to thank Dexter Ju for useful practical suggestions on filtering social medial posts, Amol Verma and Fahad Razak for pointers on migraine-related clinical keywords, and David Lopez-Paz, Patrick Forré, Roger Grosse and Sheldon Huang for feedback on an initial draft of the paper. Resources used in preparing this research were provided in part by the Province of Ontario, the Government of Canada through CIFAR, and companies sponsoring the Vector Institute. We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC), RGPIN-2021-03445.

References

Adiwardana, D., Luong, M.-T., So, D. R., Hall, J., Fiedel, N., Thoppilan, R., Yang, Z., Kulshreshtha, A., Nemade, G., Lu, Y., et al. (2020). Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.

AI@Meta (2024). Llama 3 model card.

Antonucci, A., Piqué, G., and Zaffalon, M. (2023). Zero-shot causal graph extrapolation from text via llms. *arXiv preprint arXiv:2312.14670*.

Arsenyan, V. and Shahnazaryan, D. (2023). Large language models for biomedical causal graph construction. *arXiv preprint arXiv:2301.12473*.

Ban, T., Chen, L., Wang, X., and Chen, H. (2023). From query tools to causal architects: Harnessing large language models for advanced causal discovery from data. *arXiv preprint arXiv:2306.16902*.

Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., and Blackburn, J. (2020). The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 830–839.

Capehorn, M., Catarig, A.-M., Furberg, J., Janez, A., Price, H., Tadayon, S., Vergès, B., and Marre, M. (2020). Efficacy and safety of once-weekly semaglutide 1.0 mg vs once-daily liraglutide 1.2 mg as add-on to 1–3 oral antidiabetic drugs in subjects with type 2 diabetes (sustain 10). *Diabetes & metabolism*, 46(2):100–109.

Chen, J. M., Malinsky, D., and Bhattacharya, R. (2023). Causal inference with outcome-dependent missingness and self-censoring. In Evans, R. J. and Shpitser, I., editors, *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pages 358–368. PMLR.

Cotta, L., Bevilacqua, B., Ahmed, N., and Ribeiro, B. (2023). Causal lifting and link prediction. *Proceedings of the Royal Society A*, 479(2276):20230121.

Ding, P. (2023). A first course in causal inference. *arXiv preprint arXiv:2305.18793*.

Dunning, T. (2012). *Natural Experiments in the Social Sciences: A Design-Based Approach*. Strategies for Social Inquiry. Cambridge University Press.

Egami, N., Fong, C. J., Grimmer, J., Roberts, M. E., and Stewart, B. M. (2022). How to make causal inferences using texts. *Science Advances*, 8(42):eabg2652.

Feder, A., Keith, K. A., Manzoor, E., Pryzant, R., Sridhar, D., Wood-Doughty, Z., Eisenstein, J., Grimmer, J., Reichart, R., Roberts, M. E., et al. (2022). Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics*, 10:1138–1158.

Flury, B. K. and Riedwyl, H. (1986). Standard distance in univariate and multivariate analysis. *The American Statistician*, 40(3):249–251.

Forastiere, L., Airoldi, E. M., and Mealli, F. (2021). Identification and estimation of treatment and interference effects in observational studies on networks. *Journal of the American Statistical Association*, 116(534):901–918.

Frías, J. P., Davies, M. J., Rosenstock, J., Pérez Manghi, F. C., Fernández Landó, L., Bergman, B. K., Liu, B., Cui, X., and Brown, K. (2021). Tirzepatide versus semaglutide once weekly in patients with type 2 diabetes. *New England Journal of Medicine*, 385(6):503–515.

Hillstrom, K. (2008). The MineThatData E-Mail Analytics And Data Mining Challenge. Challenge Dataset.

Jin, Z., Liu, J., Lyu, Z., Poff, S., Sachan, M., Mihalcea, R., Diab, M., and Schölkopf, B. (2023). Can large language models infer causation from correlation? *arXiv preprint arXiv:2306.05836*.

Jiralerspong, T., Chen, X., More, Y., Shah, V., and Bengio, Y. (2024). Efficient causal graph discovery using large language models. *arXiv preprint arXiv:2402.01207*.

Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E., Johnston, S., El-Showk, S., Jones,

A., Elhage, N., Hume, T., Chen, A., Bai, Y., Bowman, S., Fort, S., Ganguli, D., Hernandez, D., Jacobson, J., Kernion, J., Kravec, S., Lovitt, L., Ndousse, K., Olsson, C., Ringer, S., Amodei, D., Brown, T., Clark, J., Joseph, N., Mann, B., McCandlish, S., Olah, C., and Kaplan, J. (2022). Language models (mostly) know what they know.

Keith, K., Jensen, D., and O'Connor, B. (2020). Text and causal inference: A review of using text to remove confounding from causal estimates. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics.

Keith, K. A., Feldman, S., Jurgens, D., Bragg, J., and Bhattacharya, R. (2023). Rct rejection sampling for causal estimation evaluation. *arXiv preprint arXiv:2307.15176*.

Kıcıman, E., Ness, R., Sharma, A., and Tan, C. (2023). Causal reasoning and large language models: Opening a new frontier for causality. *arXiv preprint arXiv:2305.00050*.

Leung, M. P. (2022). Causal inference under approximate neighborhood interference. *Econometrica*, 90(1):267–293.

Lim, A. (2023). Big five personality traits: The 5-factor model of personality. simply psychology.

Long, S., Schuster, T., Piché, A., de Montreal, U., Research, S., et al. (2023). Can large language models build causal graphs? *arXiv preprint arXiv:2303.05279*.

Mestre-Ferrandiz, J., Sussex, J., and Towse, A. (2012). *The R&D cost of a new medicine*.

Miao, W., Liu, L., Tchetgen, E. T., and Geng, Z. (2015). Identification, doubly robust estimation, and semiparametric efficiency theory of nonignorable missing data with a shadow variable. *arXiv preprint arXiv:1509.02556*.

Mueller, S. and Pearl, J. (2023). Personalized decision making–a conceptual introduction. *Journal of Causal Inference*, 11(1):20220050.

Naik, N., Khandelwal, A., Joshi, M., Atre, M., Wright, H., Kannan, K., Hill, S., Mamidipudi, G., Srinivasa, G., Bifulco, C., et al. (2023). Applying large language models for causal structure learning in non small cell lung cancer. *arXiv preprint arXiv:2311.07191*.

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., Bello, I., Berdine, J., Bernadett-Shapiro, G., Berner, C., Bogdonoff, L., Boiko, O., Boyd, M., Brakman, A.-L., Brockman, G., Brooks, T., Brundage, M., Button, K., Cai, T., Campbell, R., Cann, A., Carey, B., Carlson, C., Carmichael, R., Chan, B., Chang, C., Chantzis, F., Chen, D., Chen, S., Chen, R., Chen, J., Chen, M., Chess, B., Cho, C., Chu, C., Chung, H. W., Cummings, D., Currier, J., Dai, Y., Decareaux, C., Degry, T., Deutsch, N., Deville, D., Dhar, A., Dohan, D., Dowling, S., Dunning, S., Ecoffet, A., Eleti, A., Eloundou, T., Farhi, D., Fedus, L., Felix, N., Fishman, S. P., Forte, J., Fulford, I., Gao, L., Georges, E., Gibson, C., Goel, V., Gogineni, T., Goh, G., Gontijo-Lopes, R., Gordon, J., Grafstein, M., Gray, S., Greene, R., Gross, J., Gu, S. S., Guo, Y., Hallacy, C., Han, J., Harris, J., He, Y., Heaton, M., Heidecke, J., Hesse, C., Hickey, A., Hickey, W., Hoeschele, P., Houghton, B., Hsu, K., Hu, S., Hu, X., Huizinga, J., Jain, S., Jain, S., Jang, J., Jiang, A., Jiang, R., Jin, H., Jin, D., Jomoto, S., Jonn, B., Jun, H., Kaftan, T., Łukasz Kaiser, Kamali, A., Kanitscheider, I., Keskar, N. S., Khan, T., Kilpatrick, L., Kim, J. W., Kim, C., Kim, Y., Kirchner, J. H., Kiros, J., Knight, M., Kokotajlo, D., Łukasz Kondraciuk, Kondrich, A., Konstantinidis, A., Kosic, K., Krueger, G., Kuo, V., Lampe, M., Lan, I., Lee, T., Leike, J., Leung, J., Levy, D., Li, C. M., Lim, R., Lin, M., Lin, S., Litwin, M., Lopez, T., Lowe, R., Lue, P., Makanju, A., Malfacini, K., Manning, S., Markov, T., Markovski, Y., Martin, B., Mayer, K., Mayne, A., McGrew, B., McKinney, S. M., McLeavey, C., McMillan, P., McNeil, J., Medina, D., Mehta, A., Menick, J., Metz, L., Mishchenko, A., Mishkin, P., Monaco, V., Morikawa, E., Mossing, D., Mu, T., Murati, M., Murk, O., Mély, D., Nair, A., Nakano, R., Nayak, R., Neelakantan, A., Ngo, R., Noh, H., Ouyang, L., O'Keefe, C., Pachocki, J., Paino, A., Palermo, J., Pantuliano, A., Parascandolo, G., Parish, J., Parparita, E., Passos, A., Pavlov, M., Peng, A., Perelman, A., de Avila Belbute Peres, F., Petrov, M., de Oliveira Pinto, H. P., Michael, Pokorny, Pokrass, M., Pong, V. H., Powell, T., Power, A., Power, B., Proehl, E., Puri, R., Radford, A., Rae, J., Ramesh, A., Raymond, C., Real, F., Rimbach, K., Ross, C., Rotsted, B., Roussez, H., Ryder, N., Saltarelli, M., Sanders, T., Santurkar, S., Sastry, G., Schmidt, H., Schnurr, D., Schulman, J., Selsam, D., Sheppard, K., Sherbakov, T., Shieh, J., Shoker, S., Shyam, P., Sidor, S., Sigler, E., Simens, M., Sitkin, J., Slama, K., Sohl, I., Sokolowsky, B., Song, Y., Staudacher, N., Such, F. P., Summers, N., Sutskever, I., Tang, J., Tezak, N., Thompson, M. B., Tillet, P., Tootoonchian, A., Tseng, E., Tuggle, P., Turley, N., Tworek, J., Uribe, J. F. C., Vallone, A., Vijayvergiya, A., Voss, C., Wainwright, C., Wang, J. J., Wang, A., Wang, B., Ward, J., Wei, J., Weinmann, C., Welihinda, A., Welinder, P., Weng, J., Weng, L., Wiethoff, M., Willner, D., Winter, C., Wolrich, S., Wong, H., Workman, L., Wu, S., Wu, J., Wu, M., Xiao, K., Xu, T., Yoo, S., Yu, K., Yuan, Q., Zaremba, W., Zellers, R., Zhang, C., Zhang, M., Zhao, S., Zheng,

T., Zhuang, J., Zhuk, W., and Zoph, B. (2024). Gpt-4 technical report.

Pearl, J. (2023). Judea pearl, ai, and causality: What role do statisticians play? [Online; accessed May-2024].

Pryzant, R., Card, D., Jurafsky, D., Veitch, V., and Sridhar, D. (2020). Causal effects of linguistic properties. *arXiv preprint arXiv:2010.12919*.

Reuter, U., Ehrlich, M., Gendolla, A., Heinze, A., Klatt, J., Wen, S., Hours-Zesiger, P., Nickisch, J., Sieder, C., Hentschke, C., et al. (2022). Erenumab versus topiramate for the prevention of migraine–a randomised, double-blind, active-controlled phase 4 trial. *Cephalalgia*, 42(2):108–118.

Roller, S., Dinan, E., Goyal, N., Ju, D., Williamson, M., Liu, Y., Xu, J., Ott, M., Shuster, K., Smith, E. M., et al. (2020). Recipes for building an open-domain chatbot. *arXiv preprint arXiv:2004.13637*.

Rothrock, J. F., Adams, A. M., Lipton, R. B., Silberstein, S. D., Jo, E., Zhao, X., Blumenfeld, A. M., and Group, F. S. I. (2019). Forward study: evaluating the comparative effectiveness of onabotulinumtoxina and topiramate for headache prevention in adults with chronic migraine. *Headache: The Journal of Head and Face Pain*, 59(10):1700–1713.

Schurman, B. (2019). The framework for fda's real-world evidence program.

Sheldrick, R. C. (2023). Randomized trials vs real-world evidence: how can both inform decision-making? *Jama*, 329(16):1352–1353.

Sridhar, D. and Blei, D. (2022). Causal inference from text: A commentary. *Science advances*.

Sridhar, D. and Getoor, L. (2019). Estimating causal effects of tone in online debates. *arXiv preprint arXiv:1906.04177*.

Tu, R., Ma, C., and Zhang, C. (2023). Causal-discovery performance of chatgpt in the context of neuropathic pain diagnosis. *arXiv preprint arXiv:2301.13819*.

Vashishtha, A., Reddy, A. G., Kumar, A., Bachu, S., Balasubramanian, V. N., and Sharma, A. (2023). Causal inference using llm-guided discovery. *arXiv preprint arXiv:2310.15117*.

Willard, B. T. and Louf, R. (2023). Efficient guided generation for llms. *arXiv preprint arXiv:2307.09702*.

Willig, M., Zecevic, M., Dhami, D. S., and Kersting, K. (2023). Causal parrots: Large language models may talk causality but are not causal. *preprint*, 8.

Winship, C. and Morgan, S. L. (1999). The estimation of causal effects from observational data. *Annual review of sociology*, 25(1):659–706.

X5 (2019). X5 Retail Hero: Uplift Modeling for Promotional Campaign. Challenge Dataset.

Zheng, L., Yin, L., Xie, Z., Huang, J., Sun, C., Yu, C. H., Cao, S., Kozyrakis, C., Stoica, I., Gonzalez, J. E., Barrett, C., and Sheng, Y. (2023). Efficiently programming large language models using sglang.

Appendix

A. Notation

R	Random variable corresponding to unstructured natural language text from a social media post (or report).
X	Random variable corresponding to features of an individual in a causal inference dataset.
T	Random variable corresponding to treatment or intervention assigned to an individual in a causal inference dataset.
Y	Random variable corresponding to outcome observed for an individual in a causal inference dataset.
x	Possible instance of X from its support $\mathcal{X}$.
t	Possible instance of T from its support $\mathcal{T} = \{0, 1\}$ (binary treatments).
y	Possible instance of Y from its support $\mathcal{Y} = \{0, 1\}$ (binary outcomes).
r	Possible instance of R from its support $\mathcal{R}$.
$Y(t)$	Random variable corresponding to potential outcome observed for an individual after receiving treatment t.
$e(X)$	Propensity score function for binary treatments, equal to $P(T = 1\|X)$.
X_i	Sampled value of X for individual i.
T_i	Sampled value of T for individual i.
Y_i	Sampled value of Y for individual i.
R_i	Sampled report R for individual i.
τ	Average treatment effect (ATE) given by $\mathbb{E}[Y(1) - Y(0)]$, where the expectation is over some defined population of individuals.
n	Total number of individuals.
n_1	Total number of individuals that are assigned treatment $T = 1$.
n_0	Total number of individuals that are assigned treatment $T = 0$.

B. Preliminaries

We are interested in estimating the causal effect of a treatment relative to either another treatment or no treatment in a population of interest. More precisely, we consider treatments $t \in \{0, 1\}$ and the corresponding potential outcomes $Y(1)$ and $Y(0)$ under each treatment. We wish to compute the quantity $\tau := \mathbb{E}[Y(1) - Y(0)]$, often referred to as Average Treatment Effect (ATE). Sometimes, $Y(0)$ may correspond to no treatment (control). Throughout this work, we assume binary treatments and outcomes in the Neyman-Rubin causal model. We provide a full list of notation in appendix A.

A Completely Randomized Experiment (CRE) with n participants requires no prior causal knowledge. In a CRE, the treatment assignment vector $(\tilde{T}_i)_{i=1}^n$ is a random permutation of n_1 ones and $n - n_1$ zeros sampled independently of the outcomes. In this case, the difference-in-means $\frac{1}{n_1}\sum_{i=1}^n \tilde{T}_i Y_i(1) - \frac{1}{n-n_1}\sum_{i=0}^n (1-\tilde{T}_i) Y_i(0)$ provides us with an unbiased estimate of τ.

Despite the indisputable necessity of CREs in high-stakes settings, it is often expensive and/or infeasible to have complete control over the treatment assignment. Instead, *observational* data is more readily available. Observational data often contains spurious correlations between the observed treatment T and the observed outcome $Y = TY(1) + (1-T)Y(0)$ through a common cause (confounder). Typically, this confounding is formalized as a variable X, which we assume to be discrete throughout this work, representing covariates associated with each individual. Given i.i.d. samples $\{(X_i, T_i, Y_i)\}_{i=1}^n$ from the target population, standard causal inference techniques can correct for confounding bias and provide consistent estimates of τ under Assumptions 1 and 2:

Assumption 1 (Strong Ignorability.)**.** *The potential outcomes are independent of treatment assignments conditional on covariates, i.e.,* $(Y(0), Y(1)) \perp\!\!\!\perp T|X$.

Assumption 2 (Positivity.)**.** *For every treatment* t *and covariate set* x*,* $0 < P(T = t \mid X = x) < 1$.

Following are two classical estimators of the ATE τ from observational data, each of which rely on X satisfying Assumptions 1 and 2. We refer the reader to Ding (2023) for further details.

Inverse Propensity score Weighting (IPW). The propensity score is the conditional probability of receiving a treatment given the observed features, *i.e.*, $e(x) = P(T = 1|X = x)$. The IPW estimator is given by

$$\hat{\tau}_{\text{IPW}} = \frac{1}{n}\sum_{i=1}^n \frac{T_i Y_i}{\hat{e}(X_i)} - \frac{(1-T_i)Y_i}{1-\hat{e}(X_i)}, \tag{2}$$

where $\hat{e}(x)$ is an approximation of $P(T = 1 \mid X = x)$. When $\hat{e}(x)$ is the true propensity score, $\hat{\tau}_{\text{IPW}}$ is an unbiased estimator of τ. When $\hat{e}(x)$ is estimated as empirical probability, $\hat{\tau}_{\text{IPW}}$ is consistent.

Outcome Imputation (OI). Outcome Imputation learns a model to impute outcomes from features and treatment and then marginalizes away the features to estimate τ with

$$\hat{\tau}_{\text{OI}} = \frac{1}{n}\sum_{i=1}^n \hat{\tau}(X_i, 1) - \hat{\tau}(X_i, 0), \tag{3}$$

where $\hat{\tau}(x, t)$ approximates $P(Y = 1 \mid X = x, T = t)$. Note that if $\hat{\tau}(x, t)$ is an unbiased estimation of this quantity, $\hat{\tau}_{\text{OI}}$ is an unbiased estimator of τ.

C. NATURAL Derivations

In addition to Assumptions 1 and 2, NATURAL estimators require the following assumptions to guarantee their consistency.

Assumption 3 (Natural language report data.)**.** *The target population is described by an observational data-generating process $P(X, T, Y, R)$ of data (X, T, Y), which satisfies Assumptions 1 and 2 and is jointly distributed with a random natural language string R, called a* report*. We assume access to an i.i.d. sample of reports $\{R_i\}_{i=1}^n$ from the marginal of this process.*

Assumption 4 (Access to the true observational conditional over (X, T, Y).)**.** *We can either (i) compute the conditional $P(X = x, T = t, Y = y|R = r)$ of the true data-generating process, or (ii) we can sample from $P(X = x|R = r)$ and compute $P(T = t, Y = y|R = r, X = x)$.*

Intuitively, these assumptions give NATURAL indirect access to (X, T, Y) through R. They can be weak or strong, depending on the definition of the reports R. On the one hand, if reports are copies of the observational data, *i.e.*, $R = (X, T, Y)$, then Assumption 4 is trivial to satisfy. On the other hand, if reports are all the constant, empty string, $R = \epsilon$, then Assumption 4 guarantees that we have full access to the *true* observational joint density function over (X, T, Y), which is a strong assumption. We consider how we might satisfy these assumptions in practice in the next section. Here, we assume that they hold and develop a series of consistent estimators of the ATE.

NATURAL Full. Given $\{R_i\}_{i=1}^n$ and $P(X = x, T = t, Y = y|R = r)$, we can construct an idealized version of NATURAL. Let us start by noting that the law of total expectation gives us

$$\tau = \mathbb{E}_{X,T,Y}\left[\frac{TY}{e(X)} - \frac{(1-T)Y}{1-e(X)}\right] = \mathbb{E}_R\left[\mathbb{E}_{X,T,Y|R}\left[\frac{TY}{e(X)} - \frac{(1-T)Y}{1-e(X)}\right]\right]. \tag{4}$$

A Monte Carlo estimate over reports is given by

$$\hat{\tau}_{\text{N-Full}} = \frac{1}{n}\sum_{i=1}^{n}\sum_{x,t,y} P(X = x, T = t, Y = y|R_i)\left[\frac{ty}{\hat{e}_{\text{N-Full}}(x)} - \frac{(1-t)y}{1-\hat{e}_{\text{N-Full}}(x)}\right], \tag{5}$$

which further approximates $\hat{e}_{\text{N-Full}}(x)$ from the given conditional. We used eq. (8) below.

The estimator $\hat{\tau}_{\text{N-Full}}$ above relies on enumerating all possible values of (X, T, Y), making it computationally expensive for high-dimensional X. Below, we present two hybrid versions of our method which combine sampling of some variables and computation of conditional probabilities of others.

NATURAL IPW. To construct our hybrid estimator, we augment the data $\{R_i\}_{i=1}^n$ by sampling from $P(X|R_i)$ independently for each report R_i. This gives us a dataset $\{(R_i, X_i)\}_{i=1}^n$ drawn i.i.d. from $P(X, R)$ by Assumption 4. Then, our hybrid estimator is derived from the form of IPW as follows:

$$\tau = \mathbb{E}_{R,X}\left[\mathbb{E}_{T,Y|R,X}\left[\frac{TY}{e(X)} - \frac{(1-T)Y}{1-e(X)}\right]\right], \tag{6}$$

$$\hat{\tau}_{\text{N-IPW}} = \frac{1}{n}\sum_{i=1}^{n}\sum_{(t,y)\in\mathcal{T}\times\mathcal{Y}} P(T = t, Y = y|R_i, X_i)\left[\frac{ty}{\hat{e}_{\text{N-IPW}}(X_i)} - \frac{(1-t)y}{1-\hat{e}_{\text{N-IPW}}(X_i)}\right]. \tag{7}$$

where $\hat{e}_{\text{N-IPW}}(x)$ is consistently estimated in the following manner:

$$\hat{e}_{\text{N-IPW}}(x) = \frac{\sum_{i=1}^n P(T = 1|R_i, X_i)\mathbb{I}(X_i = x)}{\sum_{i=1}^n \mathbb{I}(X_i = x)} \xrightarrow{\text{a.s.}} \frac{\mathbb{E}_{R,X}\left[P(T = 1|R, X)\mathbb{I}(X = x)\right]}{\mathbb{E}_{R,X}\left[\mathbb{I}(X = x)\right]} = e(x). \tag{8}$$

NATURAL OI. Similarly inspired by the OI estimator in equation 3, we have for $t \in \{0, 1\}$,

$$P(Y = 1 \mid T = t, X = x) = \frac{\mathbb{E}_{R,X,T}\left[P(Y = 1|R, X, T)\mathbb{I}(X = x, T = t)\right]}{\mathbb{E}_{R,X,T}\left[\mathbb{I}(X = x, T = t)\right]} \tag{9}$$

Thus, for our hybrid OI estimator, we augment the data $\{R_i\}_{i=1}^n$ by sampling from $P(X, T|R_i)$ independently for each report R_i. This gives us a dataset $\{(R_i, X_i, T_i)\}_{i=1}^n$ drawn i.i.d. from $P(R, X, T)$ by Assumption 4. Then, our consistent

outcome predictor is given by

$$\hat{\tau}_{\text{N-OI}}(x,t) = \frac{\sum_{i=1}^{n} P(Y=1|R_i,X_i,T_i)\mathbb{I}(X_i=x,T_i=t)}{\sum_{i=1}^{n}\mathbb{I}(X_i=x,T_i=t)}, \tag{10}$$

and the final estimator is given by:

$$\hat{\tau}_{\text{N-OI}} = \frac{1}{n}\sum_{i=1}^{n} \hat{\tau}_{\text{N-OI}}(X_i,1) - \hat{\tau}_{\text{N-OI}}(X_i,0) \tag{11}$$

NATURAL Monte Carlo. Further in the direction of sampling more variables, we can obtain samples (X_i, T_i, Y_i) from the entire joint conditioned on R_i and compute a Monte Carlo estimate, $\hat{\tau}_{\text{N-MC}}$. The set of samples $\{(X_i, T_i, Y_i)\}_{i=1}^{n}$ constitute a tabular dataset which can be plugged into a standard ATE estimator like IPW or OI, as described in appendix B. We refer to these sample-only estimators as N-MC IPW and N-MC OI, respectively.

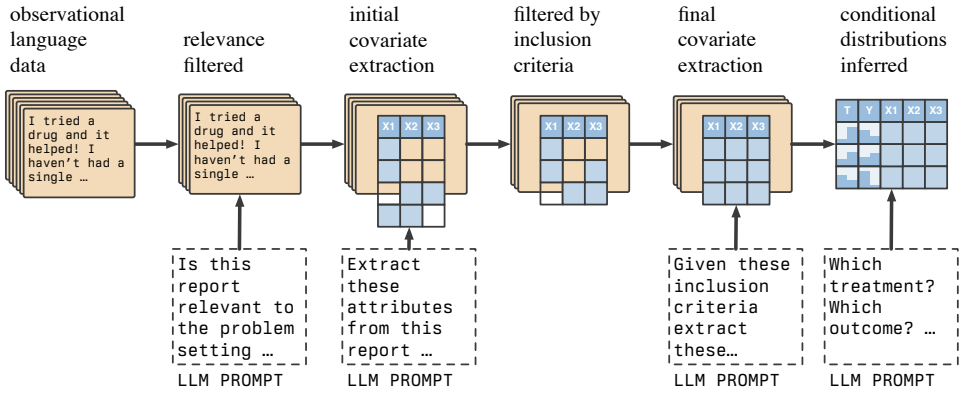

Figure 2: Our pipeline leverages LLMs to curate data that can be plugged into natural language conditioned estimators for average treatment effects.

D. Implementing NATURAL estimators with Large Language Models

LLMs are trained on vast datasets of real-world data, *e.g.*, (AI@Meta, 2024), which likely contain records of data generated by processes that are consistent with Assumption 3. Because LLMs can learn well-calibrated conditionals (Kadavath et al., 2022), our hypothesis is that LLMs can be prompted to approximate the conditionals required by Assumption 4 for real-world causal effect questions of interest. Our LLM implementation of NATURAL estimators is built on this hypothesis to try to satisfy Assumptions 3 and 4 (Assumptions 1 and 2 must be guaranteed by a domain expert). We defer exact prompts to appendix G and a discussion of the limitations to the next section. Figure 2 summarizes our pipeline.

Filtering to match Assumption 3. Our first goal is to produce a dataset of i.i.d. reports R_i that are very likely to be jointly distributed with the random variables (X_i, T_i, Y_i) of a specific observational study of interest. Given a study of interest and a dataset of real-world reports that are potentially relevant to the study, we pass it through a sequence of filters with increasing detail and strictness:

(i) **Initial filter.** Inspired by other work with social media data (Adiwardana et al., 2020; Roller et al., 2020), we first use deterministic rules to filter out uninformative reports: posts that were removed, are too short, have "bot" in the author's name, have no mention of any keyword related to the study, etc.

(ii) **Filter by relevance.** We prompt an LLM to determine whether each report contains information that would make it relevant to the study. We remove reports that are deemed irrelevant.

(iii) **Filter by treatment-outcome.** We ensure that each report pertains specifically to the treatments and outcomes of interest. We do so by prompting an LLM to extract only treatment and outcome information, and retaining only the posts that are deemed to both mention one of the treatments in question and also contain outcome information.

(iv) **Filter known covariates by inclusion criteria.** We are sometimes interested in ATEs over populations defined by constraints on pre-treatment covariates X_i known as *inclusion criteria*. In such cases, we included a filter to enforce inclusion criteria. Managing inclusion criteria is complicated by the fact that many reports R_i contain no information about covariates that are required to verify inclusion. So, in this filtering step, our goal was to ensure that the final set of reports have non-zero probability of matching the inclusion criteria. We begin by prompting an LLM to extract the full set of covariates X_i, following constraints on the possible values each attribute can take, but we allow the LLM to extract `Unknown` if it is impossible for the LLM to determine the value of a covariate. We then remove reports, if any of the non-`Unknown` covariates are determined to fail their inclusion criteria. We found the JSON-mode made available for generation by certain LLM APIs like GPT-4 to suffice for this task; however more involved strategies for constrained generation are also possible (Willard and Louf, 2023; Zheng et al., 2023).

Sampling from and computing conditional probabilities to match Assumption 4. Given a set of reports $\{R_i\}_{i=1}^n$ that pass the filtering stage above, our next steps use LLMs to extract the samples and conditionals $P_{\text{LLM}}(X, T, Y \mid R)$, required to compute NATURAL estimators. For each R_i, we:

(v) **Extract covariates, both known and unknown.** We run a final covariate extraction by prompting an LLM to determine the full set of covariates X_i from the report R_i, subject to the constraint that X_i satisfies the inclusion

criteria. In contrast to (iv), we ask the LLM to guess the values of `Unknown` covariates. We verified that this second extraction agreed exactly with the first extraction (iv) on the known covariates (*i.e.*, the ones that were not extracted as `Unknown` in the first extraction). We contrast the empirical distributions of these known and unknown/guessed covariates for our experiments in appendix I.4.

(vi) **Infer conditionals.** Given $\{R_i, X_i\}_{i=1}^n$ from the previous steps, we compute the probabilities $P_{\text{LLM}}(T = t, Y = y|R_i, X_i)$ by prompting an LLM. Specifically, we ask an LLM to answer questions about T, Y given access to R_i, X_i, and we score every possible answer $T = t, Y = y$ using the LLM log-probabilities. We exponentiate and renormalize these scores across the space of possible realizations to obtain a valid probability distribution.

To manage inclusion criteria, our LLM pipeline requires three additional technical conditions, which we discuss in more detail in Appendix J: (i) inclusion criteria define a box, *i.e.*, a separate criterion is specified for each covariate dimension, (ii) the event that the set of `Unknown` in (iv) satisfy their inclusion criteria is conditionally independent of the report and the known covariates given the event that the knowns satisfy their criteria, and (iii) the set of known and unknown covariates is conditionally independent of the treatment and outcome given the report and the covariate values. We used these to avoid an expensive estimation or a data-wasteful rejection sampling step, but these assumptions are not required in principle by NATURAL estimators.

Nevertheless, while our empirical results are remarkably consistent with the correctness of our pipeline, we cannot formally guarantee that it satisfies Assumptions 3 and 4. The final outcome of this pipeline is a dataset $\{R_i, X_i\}_{i=1}^n$ and a set of conditionals $P(T = t, Y = y|R_i, X_i)$ that can be plugged into the hybrid NATURAL estimators in section 2 to predict ATEs. Therefore, we see this as a first implementation of NATURAL estimators, which we anticipate can be improved.

E. Extended Related Work

Leveraging natural language data (Sridhar and Blei, 2022) to support causal claims is pervasive in applied research (Sridhar and Getoor, 2019; Egami et al., 2022). Our work falls under the broad umbrella of accelerating the identification of real-world evidence (RWE) (Schurman, 2019). For instance, in the context of healthcare, RWE supports not only drug repurposing, but also post-market safety evaluations —its most common application. NATURAL expands the boundaries of how quickly one can obtain and validate such real-world evidence from observational data (Sheldrick, 2023).

The use of natural language data in causal inference comes in different flavors: i) using text to measure confounders (Keith et al., 2020), ii) using text to measure causal effect outcomes (Feder et al., 2022), or iii) producing interpretable causal features from text (Feder et al., 2022; Ban et al., 2023), *e.g.*, what words are more likely to explain the cause of an event.

NATURAL distinguishes itself from these lines of research in two ways: i) NATURAL does not require any curated task-specific training data (it is zero-shot), and ii) NATURAL is not interested in how the text itself, *i.e.*, its words, relate to the causal problem —that is, we are only leveraging the model's ability to predict the distribution of a specified variable conditional on the input text. We highlight that our work lies distinct from research at the intersection of text and causality that has studied the ability of language models to infer *latent* variables (that are implied but not explicitly identified in text data) (Pryzant et al., 2020; Egami et al., 2022). Rather, we require the precise specification of covariates to condition on – we view this as being crucial to creating a more direct way for an end user to verify the validity of information extracted with our approach.

Prior works have also leveraged LLMs in a black-box fashion for causal tasks by querying the model for causal statements. In the context of causal discovery, users directly ask for the existence of cause-and-effect relationships, *e.g.*,"Does changing the age of an abalone causes a change in its length?" (Kıcıman et al., 2023; Naik et al., 2023; Antonucci et al., 2023; Arsenyan and Shahnazaryan, 2023; Tu et al., 2023; Jiralerspong et al., 2024; Ban et al., 2023). Due to the large amount of training data, it is possible that the model learns to apply a causal model described in the training data and answer causal questions with it (Pearl, 2023; Willig et al., 2023). The issue with this approach is i) the user is limited to the causal models observed in training, ii) the user is not aware of *which* causal model they are using, and iii) the queries tend to present high prompt sensitivity (Long et al., 2023).

Finally, we note that a recent work created a benchmark and showed how LLMs struggle to distinguish pairwise correlation from causation (Jin et al., 2023), while another shows that checking causal relationships in a pairwise manner can lead to invalid causal graphs (Vashishtha et al., 2023).

F. Worked Example: Semaglutide vs. Tirzepatide

To make the NATURAL pipeline and its implementation more concrete, we now work through an end-to-end example using the Semaglutide vs. Tirzepatide dataset. This was the setting used to develop our evaluation setup and tune the pipeline. We made choices regarding filtering, covariate extraction and defining variables for causal inference, which were then fixed and used for the test datasets. These considerations are explained below.

Evaluation. Having noticed a large number of posts about the effects of diabetes treatments on weight loss across subreddits (which have eventually been collected in the Pushshift collection), we used `clinicaltrials.gov` as a source for ground truth ATEs that our method could be compared against. Since the available self-reported data we are interested in rarely represents individuals that may constitute a control group, we limited our search to clinical trials that conducted a head-to-head comparison of two treatments, and confirmed that each of these treatments was mentioned in a sufficient number of posts in the data collection. As mentioned in the main paper, we selected Frías et al. (2021) (NCT03987919) which compares Semaglutide to Tirzepatide and measures a variety of endpoints, including whether or not participants achieved a target weight loss of 5% or more. Since NATURAL is developed for binary outcomes, we selected this endpoint and used the reported difference in proportion of participants achieving this target, between the treatment cohorts, as the ground truth ATE.

LLM-supported implementation.

(i) **Initial filter.** We found nine subreddits relevant to this problem setting: r/Mounjaro, r/Ozempic, r/fasting, r/intermittentfasting, r/keto, r/loseit, r/Semaglutide, r/SuperMorbidlyObese, r/PlusSize. From each subreddit, we downloaded all submissions and comments posted upto December 2022 from the Pushshift collection, so as to only use publicly available data. This resulted in a dataset of 577,733 submissions and comments. The initial deterministic, task-agnostic and rule-based filter removed any submission or comment if its content was not a string, if it had no score, if the content was `"[deleted]"` or `"[removed]"`, if it was a comment with fewer than ten space-separated strings (presumably, words), if the author's name contained the string `"bot"`, if there were no spaces in the first 2048 characters, and if less than 50% of all characters were alphabetic. This reduced the dataset size to 380,276. We then formatted this data into dictionary-like datapoints with fields: `subreddit, title, date created, post/comment, author replies`. Comments written by the author as replies to their own post may contain additional relevant information when combined with with original post and other replies. We then passed these through a task-dependent string-matching filters. For this dataset, we listed strings used commonly to refer to the treatments, `["ozempic", "mounjaro", "semaglutide", "tirzepatide", "wegovy", "rybelsus", "zepbound"]`, included common misspellings generated with GPT-4 and Perplexity, and filtered out datapoints that did not contain any of these strings. Similarly, we listed keywords relevant to the outcome of interest, `["kg", "kilo", "lb", "pound", "weigh", "drop", "loss", "lost", "gain", "hb", "a1c", "hemoglobin", "haemoglobin", "glucose", "sugar"]` and filtered out datapoints that did not contain any of these strings. This filtered dataset now contained 50,654 datapoints.

(ii) **Filter by relevance.** Next, we wrote a problem setting description and prompted GPT-3.5-Turbo to determine whether the posts, along with auxiliary information from the formatted dictionaries described above, were relevant to the described setting. The description and instructions for this particular dataset are shown in prompt 2. We manually labeled a handful of datapoints as `Yes` or `No` and included these as incontext examples to improve the LLM's generations. We removed datapoints that were deemed irrelevant, resulting in a "relevant" dataset of 21,229 datapoints.

(iii) **Filter by treatment-outcome.** To further filter the data to points that refer specifically to the treatments and outcome of interest, we prompted GPT-3.5-Turbo to extract only information required to ascertain the treatment and outcome, as shown in prompt 3. Since the outcome for this dataset, achievement of a target weight loss of 5% or more, may be reported in several ways, we attempted to cover all those possibilities. Specifically, we prompted the LLM to extract the user's starting weight, end weight, change in weight and percentage of change in weight. Several combinations of these attributes allow us to programmatically infer the final outcome. We also extracted the units in which weight was reported, converting all extractions to be in lbs. We filtered out any datapoint for which the extracted treatment was not one of the treatments considered for this task or for which it was not possible to infer the outcome using the above-mentioned extracted information. This finally gave us a natural language dataset of 4619 relevant reports, each of which contained treatment and outcome information pertaining to the defined problem setting.

(iv) **Filter known covariates by inclusion criteria.** To fairly evaluate against a real clinical trial, we used the trial design determined before actually conducting the trial to further filter the dataset. In particular, we noted the inclusion criteria

enforced in the clinical trial and aimed for a set of reports with non-zero probability of satisfying these criteria. The criteria for this dataset were: (i). the participant must be diagnosed with type 2 diabetes, (ii). they must already be on a regime of the treatment called Metformin, and (iii). they must have a BMI of 25 or more. Since different treatment dosages can have varying effects, we also included (iv). dosage as a criterion for matching here, *i.e.* 1mg for Semaglutide and 5mg for Trizepatide, as in the clinical trial. As described in item (iv) and motivated in appendix J, we extracted all covariates including ones related to the inclusion criteria and removed datapoints that whose extractions were not `Unknown` but failed to satisfy the criteria above. This resulted in a dataset of 1265 reports.

(v) **Extract known and unknown covariates.** Treating these 1265 reports as the final dataset from which to estimate an ATE, we again used the real trial design as expert guidance for defining covariates, specifically, pre-treatment information ("baseline characteristics") reported in the study. We defined `"age", "sex", "BMI", "start weight", "start HbA1c"]` as the adjustment set for causal inference. We also included the `"duration"` of treatment as a covariate since this information is often reported and is likely to influence the outcome. This extraction step was conditioned on inclusion criteria being satisfied, a description of which was included in the extraction prompt, as in prompt 5.

(vi) **Infer conditionals.** We inferred conditional distributions from LLAMA2-70B for different versions of NATURAL, with the strategy described in item (vi) and LLM imputs of the form shown in prompt 6. Here, "conditioning on covariates" was implemented by adding questions about the covariates and their sampled answers to the input. For instance, for `sex`, the question `"What is the reported sex of the user?"` was followed by its previously extracted answer (`Male` or `Female`). The scoring strategy required enumerated treatments and outcomes for each input, which were `["Semaglutide like Ozempic or Wegovy or Rybelsus", "Tirzepatide like Mounjaro or Zepbound"]` and `["No", "Yes"]`, respectively.

Causal inference. Given all the required extractions and conditionals from LLMs, we required discrete covariates to plug them into our NATURAL estimators. Hence, we converted any continuous covariates into discrete categories. These categories for each dataset are shown in table 4 for all our datasets. Different choices of discretization led to slightly different ATE predictions. We found it most helpful to discretize continuous numerical covariates into intervals such that the number of datapoints were roughly balanced across interval. This avoided covariate strata with too many or too few datapoints and resulted in ATE predictions from all NATURAL estimators that were sufficiently close to the ground truth. Having validated on this "tuned" dataset, we adopted the same principle for the other three test datasets. For the final results reported in table 2, we report ATE mean, standard deviation and root mean squared error over ten trials, each with 80% of the data sampled randomly without replacement.

G. LLM Prompts

Prompt 1: Synthetic report generation (Hillstrom)

```
You are a user who used a website for online purchases in the past one year and want
    to share your background and experience with the purchases on social media.

## Attributes
The following are attributes that you have, along with their descriptions.
> {features}

## Personality Traits
The following dictionary describes your personality with levels (High or Low) of the
    Big Five personality traits.
> {traits}

## Your Instructions
Write a social media post in first-person, accurately describing the information
    provided. Write this post in the tone and style of someone with the given
    personality traits, without simply listing them.
Only return the post that you can broadcast on social media and nothing more.

## Post
>
```

Prompt 2: Relevance filtering (Weight Loss)

```
You are an expert researcher looking around reddit for posts/comments describing the
    effect of a treatment on weight loss or blood sugar level experienced by the
    author.

## Problem Setting
> You are interested in self-reported effects of a treatment on a user who took the
    treatment themselves. You want to be able to answer some or all of the following
    questions from the text of the post or comment:
1. Which treatment did the user take?
2. What change did they observe in their weight due to this treatment, and during
    what duration did they observe this change?
3. What change did they observe in their blood sugar, aka HbA1c levels, due to this
    treatment, and during what duration did they observe this change?
4. What are other attributes they report, e.g. age, sex, country of residence,
    diabetes diagnosis, other treatments they have tried, or side effects?

## Your Instructions
I will show you a post or comment, and contextual information about it. Based on the
    given problem setting and contextual information, you need to judge whether it
    is relevant to the problem setting described above or not. Answer Yes if the
    post is relevant and No otherwise; nothing else.
Here are a few examples:

{incontext examples}

## Subreddit
> This post was found on the subreddit r/{subreddit}.

## Title
> This post was titled: {title}

## Date Created
> This post was created on {date_created}.

## Post
> {post}

The author also replied with the following in the thread:
> {replies}

Answer Yes if the comment is relevant and No otherwise, and nothing more.
## Your Answer
>
```

Prompt 3: Treatment-outcome filtering (Weight Loss)

```
You are a medical assistant, helping a doctor structure posts about weight loss
    treatments found on Reddit. Your task is to use the self-report to interpret
    accurate information about the following fields and store them in a JSON
    dictionary.

## Your Instructions
I will provide a post along with its subreddit name, title and date of creation. You
    must return a valid JSON dictionary containing the following keys along with the
    corresponding accurate information:
"start_weight": Numerical value for the user's starting weight, before starting the
    treatment described, sometimes referred to as SW.
"end_weight": Numerical value for the user's current or final weight, at the end of
    the treatment regime, sometimes referred to as CW.
"weight_unit": Units in which weight is reported: "kg" or "lb".
"weight_change": Numerical value for net change in the user's weight. Use a postive
    sign to indicate weight gain and negative sign for weight loss. Leave blank if
    it is not possible to infer the change in weight.
"percentage_weight_change": Numerical value for percentage reduction in user's
    weight relative to their start weight. Use a postive sign to indicate weight
    gain and negative sign for weight loss. Leave blank if it is not possible to
    infer the percentage.
"drug_type": Treatment taken by the user: "Semaglutide", "Tirzepatide" or "Other".
    Semaglutide includes Ozempic, Wegovy or Rybelsus. Tirzepatide includes Mounjaro
    or Zepbound.

Assign a valid value to each key above. If you can't find the required information
    in the post, assign the value "Unknown". Remember to ONLY return a valid JSON
    with ALL of the above keys and their accurate values.
```

Prompt 4: Covariate extraction (Weight Loss)

```
As a medical assistant aiding a physician, your role involves examining Reddit posts
    discussing weight loss treatments and interpreting self-reported information
    accurately. This data needs to be translated into a well-structured JSON
    dictionary, with the most suitable option chosen from the choices provided.

## Your Instructions
Assume a user shares a post along with related data. Your job will be to create a
    dictionary comprising of the following keys as well as their matching accurate
    data:

{covariate descriptions}

Please ensure you fill all the fields and that you choose a valid value for each key
    from the provided options. Unfilled fields are not allowed. In instances where
    certainty is impossible, make your best educated guess, or provide the "Unknown"
    value. Note that your completed task should ONLY yield a JSON containing ALL the
    listed keys alongside their accurate values.

Here are a few examples:

{incontext examples}

## Input
{report}

## Output
>
```

Prompt 5: Covariate imputation (Weight Loss)

```
You are a medical assistant tasked with creating a profile of a patient who is
    taking a weight loss treatment, and presenting it as a JSON dictionary with
    prespecified keys. Fill in suitable values for ALL the keys. You can use
    information provided about the patient.

## Your Instructions
A patient has Type 2 Diabetes, is known to have taken Metformin for the last 3
    months and has a BMI greater than 25 kg per meter squared.
Dosage for Semaglutide, Ozempic, Wegovy and Rybelsus is 1mg. Dosage for Tirzepatide,
    Mounjaro and Zepbound is 5mg.
Create a possible profile for this patient with the following fields and represent
    it as dictionary:

{covariate descriptions}

Please ensure you fill all the fields with a valid value. Unfilled fields or values
    like "Unknown" are not allowed. Note that your completed task should ONLY yield
    a JSON containing ALL the listed keys alongside their accurate values.

Here is an entry that the patient wrote about themselves, which may be useful for
    your task.
## Input
{report}

## Output
>
```

Prompt 6: Conditional distribution inference (Weight Loss)

```
You are a medical assistant aiding a physician. I am going to ask you a few multiple
    choice questions about some posts I just found online. Please, answer
    accordingly.

## Your Instructions
I will give you a post about an individual's experience with a treatment and its
    effect on their weight, and a few questions with their correct answers, followed
    by additional multiple choice questions and options to choose from. Pick the
    right answer.

## Social Media Post
> {report}

## Questions and their correct answers
Q: {question about covariate X1} A: {X1 sample}.
Q: {question about covariate X2} A: {X2 sample}.
..

## Questions
Q: Which treatment did the user take?
Options: a) {t0} b) {t1}
A: {t0}

Q: Did the user lose 5 or more percent of their initial weight?
Options: a) {y0} b) {y1}
A: {y0}
```

H. Dataset details

We provide further details about the treatments, outcomes and covariates (along with their discrete categories used in our experiments) for each dataset in tables 3 and 4.

Table 3: Treatments, outcomes and synthetic confounders (where applicable) for each dataset.

Dataset	Treatment	Outcome	Synthetic confounder
Hillstrom	email communication	website visit	newbie
Retail Hero	SMS communication	purchase	age
Semaglutide vs. Tirzepatide	corresponding drug	weight loss of 5% or more	NA
Semaglutide vs. Liraglutide	corresponding drug	weight loss of 10% or more	NA
Erenumab vs. Topiramate	corresponding drug	discontinuation due to adverse effects	NA
OnabotulinumtoxinA vs. Topiramate	corresponding drug	discontinuation due to adverse effects	NA

Table 4: Covariate descriptions, corresponding discrete categories and inclusion criteria enforced for each dataset. Intervals for continuous numerical variables were determined from the extracted values such that each discrete category is roughly balanced in terms of its number of datapoints.

Covariate	Description	Discrete categories	Inclusion criteria
Hillstrom			
recency	number of months since last purchase	$[1-4, 5-8, 9-12]$	
history	dollar value of previous purchase	$[0-100, 100-200, \ldots, >1000]$	
mens	purchase of men's merchandise	[True,False]	
womens	purchase of women's merchandise	[True,False]	NA
zip_code	type of area of residence	[Suburban area,Rural area,Urban area]	
newbie	new customer	[True,False]	
channel	channel used for purchases	[Phone,Web,Multichannel]	
Retail Hero			
avg. purchase	avg. purchase value per transaction	$[1-263, 264-396, 397-611, >612]$	
avg. product quantity	avg. number of products bought	$[\leq 7, >7]$	
avg. points received	avg. number of points received	$[\leq 5, >5]$	NA
num transactions	total number of transactions so far	$[\leq 8, 9-15, 16-27, >28]$	
age	age of user	$[\leq 45, >45]$	
Semaglutide vs. Tirzepatide			
age	age of user	$[\leq 45, >45]$	
sex	sex of user	[Male,Female]	(t2 diabetes==True)
bmi	body mass index of user	$[\leq 28.5, >28.5]$	& (7 $\leq$ start HbA1c $\leq$ 10.5))
start HbA1c	initial glycated haemoglobin value	$[\leq 7.5, >7.5]$	& (metformin==True)
start weight	initial weight in lbs	$[\leq 220, >220]$	& (bmi $\geq$ 25)
duration (days)	number of days treatment was taken for	$[\leq 90, >90]$	
Semaglutide vs. Liraglutide			
age	age of user	$[\leq 45, >45]$	
sex	sex of user	[Male,Female]	
bmi	body mass index of user	$[\leq 28.5, >28.5]$	(t2 diabetes==True)
start HbA1c	initial glycated haemoglobin value	$[\leq 7.5, >7]$	& (7 $\leq$ start HbA1c $\leq$ 11))
start weight	initial weight in lbs	$[\leq 220, >220]$	& (metformin/other==True)
duration (days)	number of days treatment was taken for	$[\leq 120, >120]$	
Erenumab vs. Topiramate			
age	age of user	$[\leq 32, >32]$	
sex	sex of user	[Male,Female]	(18 $\leq$ age $\leq$ 65)
country	country of residence	[United States,Canada,...]	& (pregnant==False)
baseline MMD	initial number of monthly migraine days	$[\leq 6, >6]$	& (baseline MMD $\geq$ 4)
duration (days)	number of days treatment was taken for	$[\leq 30, >30]$	
OnabotulinumtoxinA vs. Topiramate			
age	age of user	$[\leq 25, >25]$	
sex	sex of user	[Male,Female]	
country	country of residence	[United States,Canada,...]	(18 $\leq$ age $\leq$ 65)
baseline MMD	initial number of monthly migraine days	$[\leq 15, >15]$	& (baseline MMD $\geq$ 15)
duration (days)	number of days treatment was taken for	$[\leq 30, >30]$	

I. Further experimental results

I.1. How well does NATURAL estimate observational distributions from self-reported data?

Our synthetic datasets give us access to the true joint distributions $P(X, T, Y)$ and true propensity scores $P(T = 1|X)$. The top row of fig. 3 shows the KL divergence between these distributions and those estimated by NATURAL Full, for Hillstrom (left) and Retail Hero (right). We find that these KL divergences decrease steadily as the number of reports used in the estimation increases. The bottom row shows corresponding root-mean-squared error (RMSE) between NATURAL and the true ATE. This corroborates the insight that as the joint distribution and propensity scores are estimated more accurately, the predicted ATE gets closer to its true value. In particular, we observe a clear correlation between the quality of estimated propensity scores and estimated ATEs.

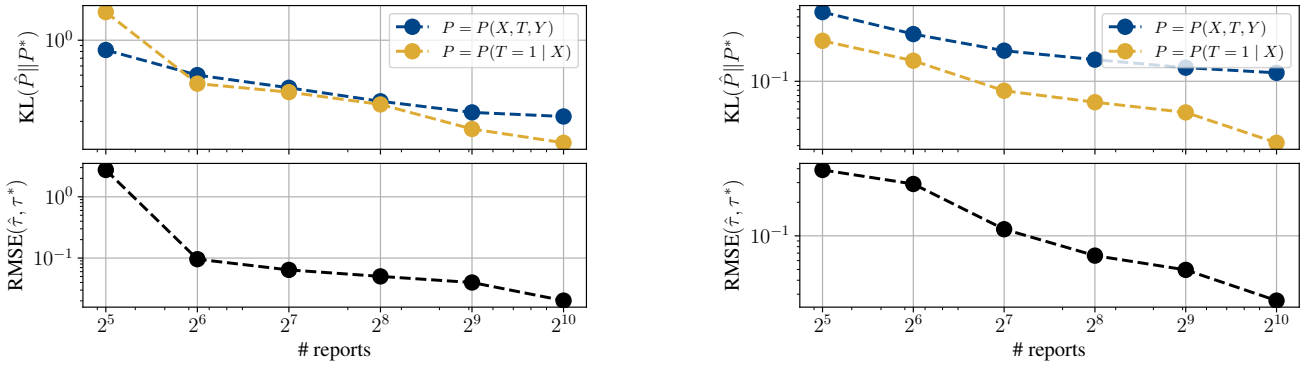

Figure 3: For Hillstrom (left) and Retail Hero (right), the KL divergence between estimated joint and propensity distributions and their true counterparts reduces with increasing number of posts (top), as does the RMSE between the NATURAL Full estimate and true ATE (bottom).

I.2. How do different choices in the NATURAL pipeline effect ATE prediction?

We assess the impact of key choices in our pipeline described in appendix D, by ablating them one-by-one. We investigated and selected these choices on the Semaglutide vs. Tirzepatide experiment. Appendix I.2 compares the RMSE of predicted ATEs when data is not filtered according to inclusion criteria and LLM imputations are replaced with samples from an uniform distribution. It shows that both inclusion-based filtering and imputations from a pretrained LLM are crucial for the performance of NATURAL. We also compared performance of our method when the conditional probabilities in eq. (7) are evaluated using models of different scales in appendix I.2, and found that performance improves at larger scales and with greater quantity of data.

I.3. How well do different estimates of propensity score balance covariates?

A property of accurate propensity score estimates is that they balance covariates across treatment cohorts (see (Ding, 2023) for details and proofs), *i.e.* the average treatment effect on each covariate, corrected using propensity scores, is close to zero. fig. 5 visualizes this quantity for different covariates of the Semaglutide vs. Tirzepatide experiment and shows that propensity scores estimated using LLAMA conditional distributions balance the covariates far better than a uniform distribution does, with the 70B model consistently estimating the treatment effect on each covariate as close to zero.

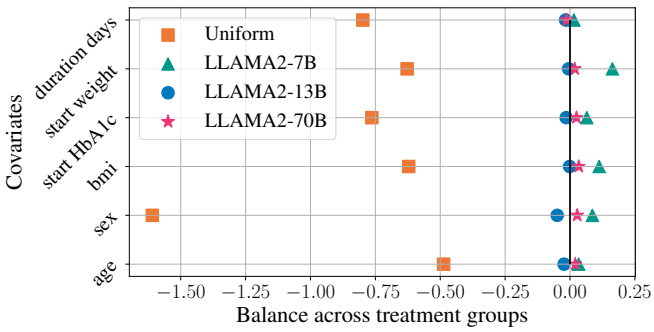

Figure 5: NATURAL propensity scores balance the Semaglutide vs. Tirzepatide covariates better than uniform scores.

We refer the reader to fig. 6 for visualizations of the propensity score corrected average treatment effect on covariates for all test clinical settings. For each setting, our estimated propensity score balances each covariate, far better than a uniform propensity distribution would.

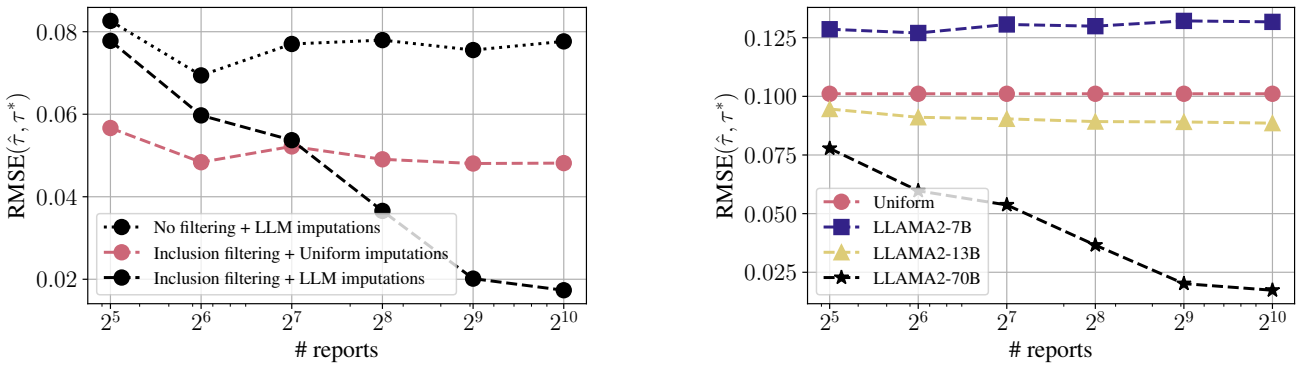

Figure 4: Ablation study on Semaglutide vs. Tirzepatide, to tease apart the effect of data filtering and imputation (left) as well as LLM scale for conditionals (right) on NATURAL performance.

Since covariates may take values at different scales, we computed the standard mean difference (SMD) across cohorts for each covariate $X^{(i)}$ (Flury and Riedwyl, 1986), given by:

$$SMD = \frac{X^{(i)}(1) - X^{(i)}(0)}{\sqrt{0.5 * (\texttt{var}(X^{(i)}(1)) + \texttt{var}(X^{(i)}(0)))}}, \tag{12}$$

where $X^{(i)}(1) - X^{(i)}(0)$ estimates the average treatment effect on $X^{(i)}$, using propensity score weighting, and $\texttt{var}(\cdot)$ denotes sample variance.

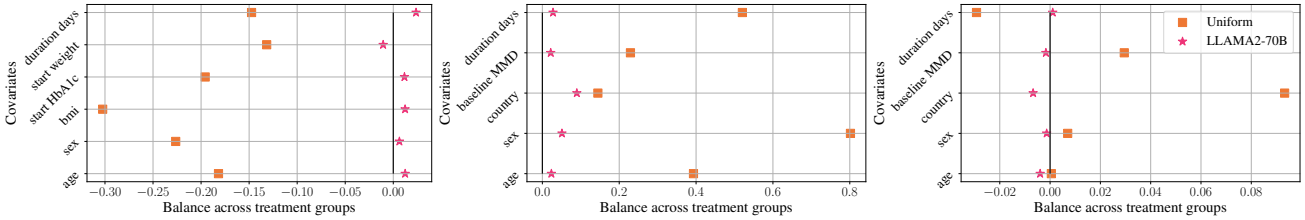

Figure 6: Propensity scores estimated with LLAMA2-70B balance covariates far better than uniform scores do, for the real clinical settings, Semaglutide vs. Liraglutide (left), Erenumab vs. Topiramate (center), and OnabotulinumtoxinA vs. Topiramate (right).

I.4. Known and Unknown/Imputed covariates for real data experiments

We refer the reader to figs. 7 to 10 for empirical distributions of covariates extracted by an LLM in its first extraction as well as those imputed in its second imputataion conditioned on inclusion criteria.

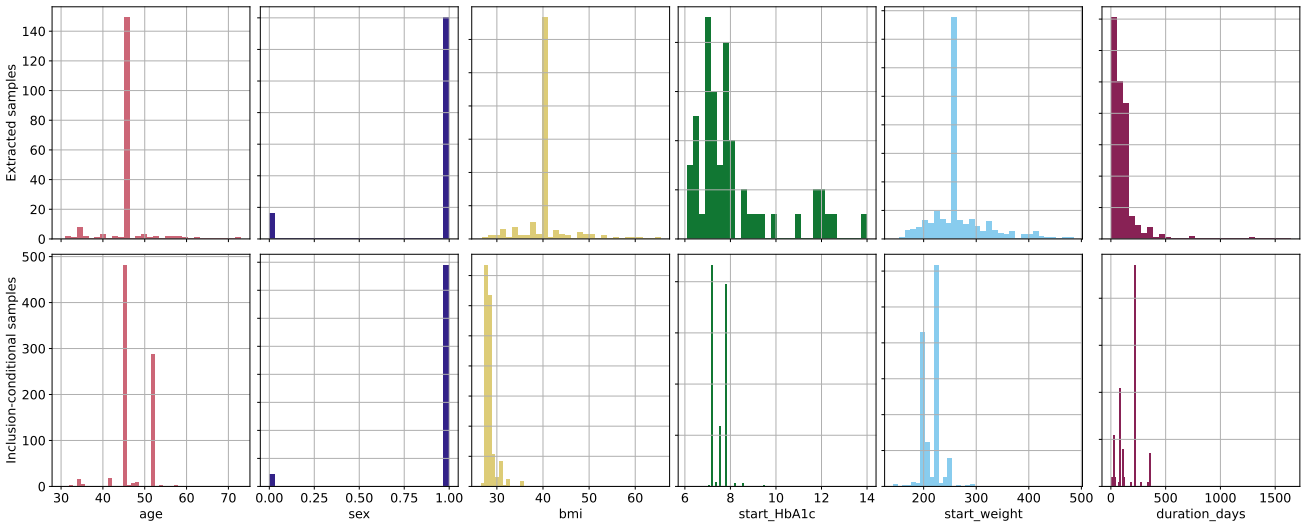

Figure 7: Distributions of "known" (top) vs "unknown" and imputed (bottom) covariates for Semaglutide vs. Tirzepatide.

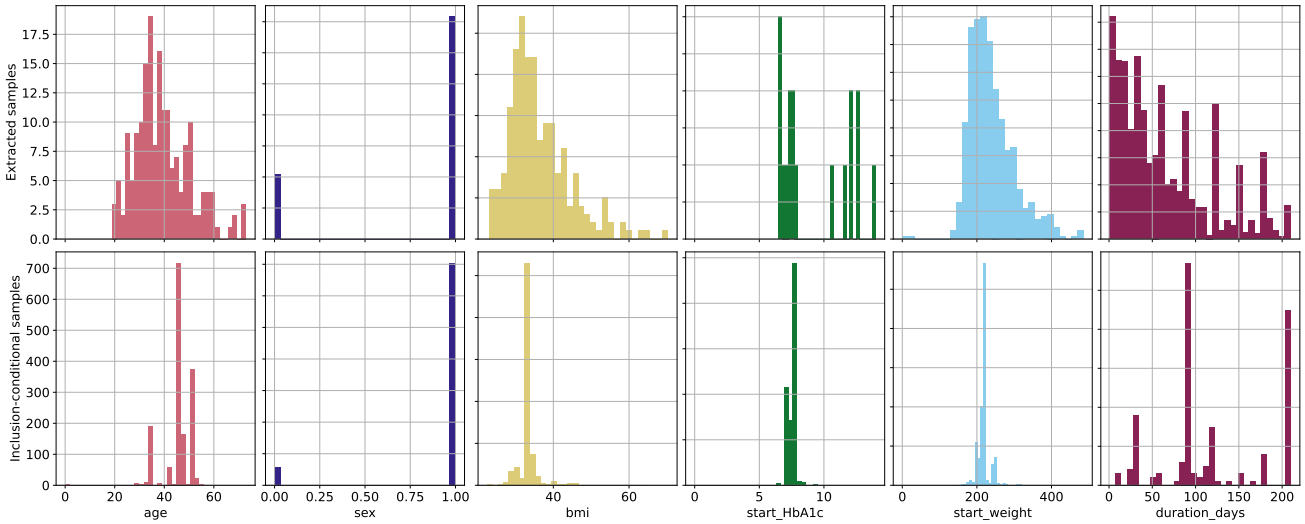

Figure 8: Distributions of "known" (top) vs "unknown" and imputed (bottom) covariates for Semaglutide vs. Liraglutide.

J. Inclusion Criteria conditioned Estimator

We are interested in an ATE conditioned on inclusion criteria denoted I,

$$\tau(I) = \mathbb{E}[Y(1) - Y(0) \mid X \in I]. \tag{13}$$

Let $\tau(X, T, Y)$ be a function such that

$$\tau(I) = \mathbb{E}_{X,T,Y}[\tau(X, T, Y) \mid X \in I]. \tag{14}$$

For example, if $\tau(I)$ can estimated by the IPW estimator,

$$\tau(X, T, Y) = \frac{TY}{e(X)} - \frac{(1-T)Y}{1-e(X)},$$

because the $P(T = 1|X = x, X \in I) = P(T = 1|X = x)$. Throughout this section, we operate under Assumptions 3 and 4 and assume that the LLM gives us access to the true data-generating conditionals.

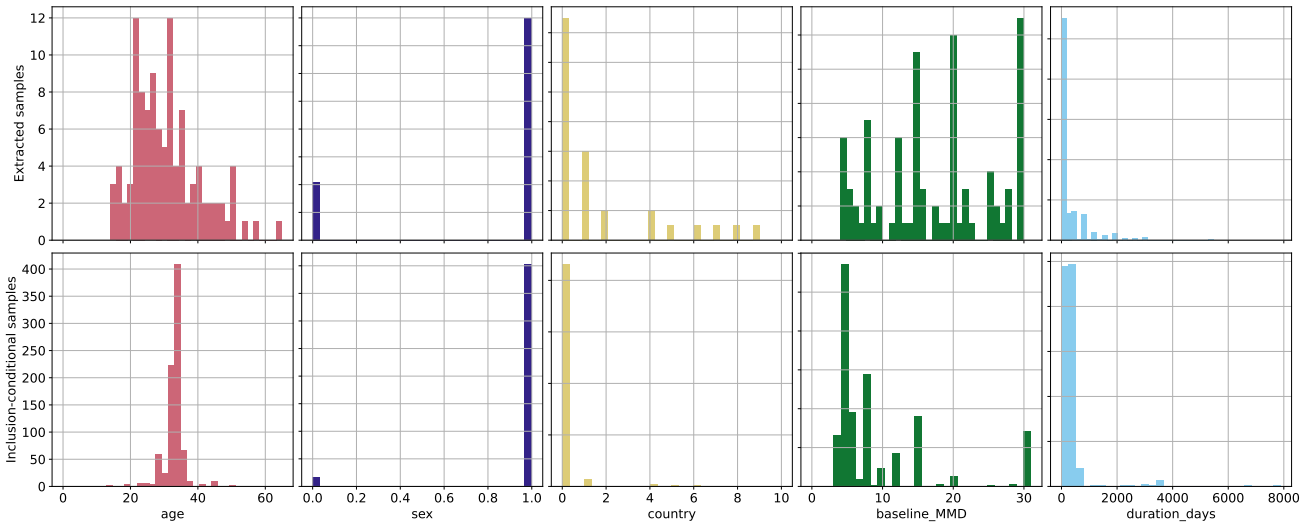

Figure 9: Distributions of "known" (top) vs "unknown" and imputed (bottom) covariates for Erenumab vs. Topiramate.

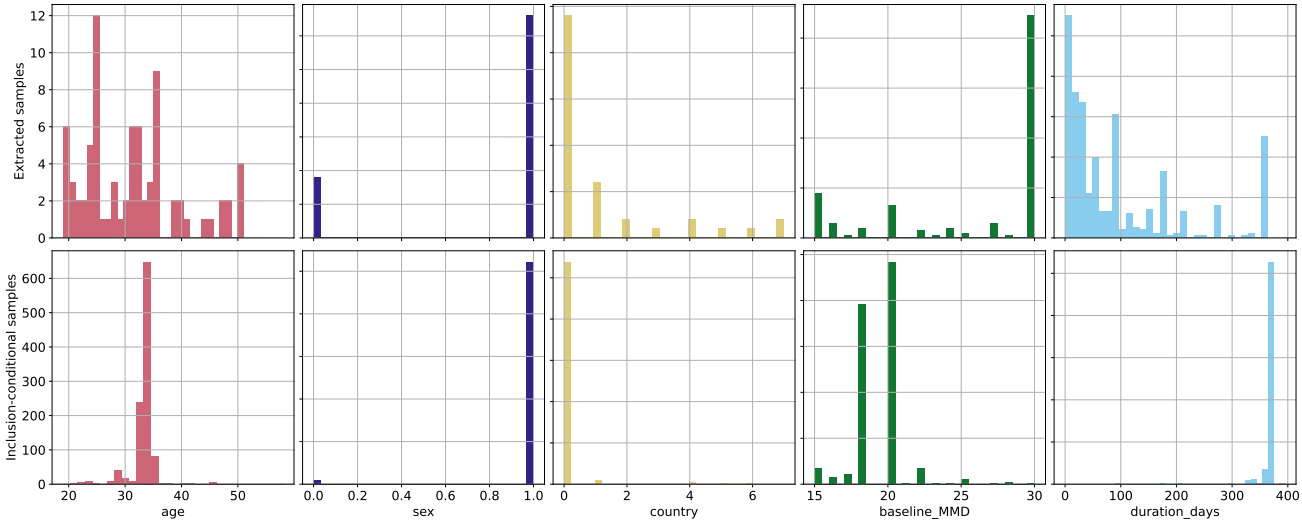

Figure 10: Distributions of "known" (top) vs "unknown" and imputed (bottom) covariates for OnabotulinumtoxinA vs. Topiramate.

The law of total expectation gives us an estimator that can operate on samples of reports R:

$$
\begin{aligned}
\tau(I) &= \mathbb{E}_{X,T,Y}[\tau(X,T,Y) \mid X \in I] \\
&= \mathbb{E}_{R|X\in I}[\mathbb{E}_{X,T,Y}[\tau(X,T,Y) \mid X \in I, R]] \\
&= \sum_r P(R=r|X \in I)\mathbb{E}_{X,T,Y}[\tau(X,T,Y) \mid X \in I, R] \\
&= \sum_r P(R=r)\frac{P(X \in I|R=r)}{P(X \in I)}\mathbb{E}_{X,T,Y}[\tau(X,T,Y) \mid X \in I, R] \\
&= \mathbb{E}_R\left[\frac{P(X \in I|R)}{P(X \in I)}\mathbb{E}_{X,T,Y}[\tau(X,T,Y) \mid X \in I, R]\right].
\end{aligned}
$$

To summarize, we have the identities:

$$\tau(I) = \mathbb{E}_{R|X\in I}[\mathbb{E}_{X,T,Y}[\tau(X,T,Y) \mid X \in I, R]] \tag{15}$$

$$= \mathbb{E}_R\left[\frac{P(X \in I|R)}{P(X \in I)}\mathbb{E}_{X,T,Y}[\tau(X,T,Y) \mid X \in I, R]\right]. \tag{16}$$

It is possible to directly estimate eq. (16) with Monte Carlo samples from $P(R)$ and an approximation of $P(X \in I|R)/P(X \in I)$ using an LLM, but the latter is computationally very expensive. Alternatively, one can use rejection sampling to simulate $R|X \in I$ by sampling $X_i \sim P(X|R_i)$ and rejecting R_i if $X_i \notin I$. This step could be very wasteful for precious data. In fact, in our experiments with the Semaglutide vs. Tirzepatide dataset, we found that using GPT-4-Turbo for sampling $X_i \sim P(X|R_i)$ resulted in a very peaky distribution with little diversity in the extractions, even after increasing its temperature argument as much as possible without sacrificing generation quality. As a result, after rejecting R_i if $X_i \notin I$, we were left with a very small number of reports, that would be infeasible to plug into any estimator. Therefore, we devise a few assumptions that allow us to estimate eq. (14) without these approaches.

Let $X \in \mathbb{R}^D$ and let us make the following assumption on the inclusion criteria:

Assumption 5 (Inclusion criteria specification)**.** *The inclusion criterion I defines a box, i.e., it is specified separately for each covariate dimension $I^d, d \in \{1,\dots,D\}$ and the set of covariates satisfying every inclusion criteria is given by the product of individual criteria over the covariate dimensions, i.e., $\{X \in I\} = \prod_{d=1}^{D}\{X^d \in I^d\}$ where X^d is the d-th dimension of $X = (X^d)_{d=1}^D$.*

Recall from appendix D that inclusion-based filtering leaves us with reports whose covariates are either "known" and satisfy their criteria or `Unknown`. We also have the value of the known covariates. Let $K \in \{0,1\}^D$ be the binary vector of variables K^d that indicate whether the covariate X^d is found to be "known" for a random report R. Let $X^K = (X^d : K^d = 1)$ be the vector of length $\sum_d K^d$ holding the values of the known covariates. For ease of notation, define the event that the known covariates satisfy their criteria and the event that the unknown covariates satisfy their criteria:

$$\{X^K \in I^K\} = \{X^d \in I^d, \forall d\colon K^d = 1\} \tag{17}$$

$$\{X^{1-K} \in I^{1-K}\} = \{X^d \in I^d, \forall d\colon K^d = 0\} \tag{18}$$

Notice that $\{X^K \in I^K\} \cap \{X^{1-K} \in I^{1-K}\} = \{X \in I\}$. Thus, after the filtering steps we have

$$\{R_i, K_i, X_i^{K_i}\}_{i=1}^n \tag{19}$$

with the guarantee that the knowns satisfy their inclusion criteria, $\{X_i^{K_i} \in I^{K_i}\}$. Formally, assuming that the LLM computes the true conditional distribution of the data-generating process (Assumption 4), this gives us data sampled i.i.d. from $P(R = r, K = k, X^k = x^k|X^K \in I^K)$. Note, that we are assuming the existence of an additional ground-truth random variable K in the data-generating process that describes whether a covariate is knowable from a report. Here, we show how to estimate $\tau(I)$ from this dataset of filtered reports using importance sampling, under the following assumption:

Assumption 6 (Satisfaction of I by `Unknown` covariates)**.** *Satisfaction of inclusion criteria by unknown covariates is conditionally independent of the report and the known covariates given satisfaction of inclusion criteria by known covariates, i.e., for all r, k, x^k:*

$$P(R = r, K = k, X^k = x^k|X^K \in I^K, X^{1-K} \in I^{1-K}) = P(R = r, K = k, X^k = x^k|X^K \in I^K) \tag{20}$$

One can derive the following identity in a similar fashion as eq. (16)

$$\tau(I) = \mathbb{E}_{R,K,X^K|X^K\in I^K}\left[\frac{P(R,K,X^K|X \in I)}{P(R,K,X^K|X^K \in I^K)}\mathbb{E}_{X,T,Y}[\tau(X,T,Y) \mid X \in I, R, K, X^K]\right]. \tag{21}$$

From assumption 6, the fraction above simplifies to 1, leaving us with the following estimator

$$\tau(I) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{X_i,T_i,Y_i}[\tau(X_i,T_i,Y_i) \mid X_i \in I, R_i, K_i, X_i^{K_i}], \tag{22}$$

which can be computed from the information available at the end of filtering. In practice, we do not condition the LLM on K_i in the final inference step (vi), which amounts to an additional conditional independence assumption:

Assumption 7 (Conditional independence of knowable covariates)**.** $K \perp\!\!\!\perp (T, Y) \mid (X, R)$.

Equation (22) above can now be more efficiently estimated by prompting the LLM to extract covariates under the constraints of the inclusion criteria for each report in our filtered dataset, and then following the remaining steps in the pipeline to an ATE estimate.