Norwegian
University of
Life Sciences

Master's Thesis 2026 **60 ECTS**
Faculty of Chemistry, Biotechnology and Food Science

Using LLM Knowledge for Improved Out-of-Distribution Tabular Prediction

A Cross-Validated Regularization Approach

Sølve Langseth
Bioinformatics and Applied Statistics

Abstract

Out-of-distribution (OOD) prediction is a well-known problem in applied statistics and machine learning: models trained on data from one population can lose performance when deployed on a different one. Standard L2 regularization controls variance by shrinking coefficients toward zero, but offers no mechanism for incorporating outside information about plausible effects, even though such information could help when the source data alone do not characterize the deployment population well. This thesis investigates whether prior coefficient information elicited from large language models (LLMs) can be incorporated into regularized linear and logistic regression to improve OOD prediction on tabular data.

Three prior-informed estimators are developed, each corresponding to a different parameterization of a Gaussian prior inside the maximum a posteriori (MAP) framework: a Prior-Mean estimator that shrinks toward an LLM-estimated mean, an Independent-Prior estimator that adds feature-specific variances estimated from repeated LLM samples, and a Correlated-Prior estimator that uses the full empirical covariance across samples. Fifty coefficient vectors are sampled from each of twelve frontier LLM variants (three providers $\times$ two model sizes $\times$ two modes: reasoning and non-reasoning); the LLM variant and prior strength λ are then selected jointly by 5-fold cross-validation on source-domain training data. The pipeline is evaluated on five tabular datasets (Boston Housing, Cancer, Heart Disease, Diabetes, and Adult), each split into a source domain and an OOD evaluation set on a structural variable.

Prior-informed estimators outperform a regularized baseline on four of the five datasets. The regression tasks show the largest gains: -15.8% OOD MSE on Boston Housing and -3.9% on Cancer, with paired-bootstrap confidence intervals excluding zero for every prior family. On Heart Disease and Diabetes, OOD AUC is essentially unchanged, but F1, sensitivity, and probability calibration improve significantly. Adult, the largest dataset with the mildest shift, exhibits the expected safe-failure behavior: cross-validation drives the prior strength toward zero and the prior-informed estimators recover the baseline. The cross-provider selection pipeline consistently lands close to the per-dataset OOD-best ceiling, suggesting that LLM-derived coefficient priors, filtered through cross-validation and provider-family selection, are a viable source of structured domain information for classical statistical models under distribution shift.

Sammendrag

Prediksjon utenfor treningsfordelingen (out-of-distribution, OOD) er et velkjent problem i anvendt statistikk og maskinlæring: modeller som trenes på data fra én populasjon, kan få redusert ytelse når de tas i bruk på en annen. Standard L2-regularisering kontrollerer varians ved å krympe koeffisienter mot null, men har ingen mekanisme for å innlemme ekstern informasjon om plausible effekter, selv om slik informasjon kan være nyttig når kildedataene alene ikke gir et godt bilde av populasjonen modellen skal brukes på. Denne oppgaven undersøker om koeffisientprior-informasjon hentet fra store språkmodeller (LLM-er) kan integreres i regularisert lineær og logistisk regresjon for å forbedre OOD-prediksjon på tabulære data.

Tre prior-informerte estimatorer utvikles, hver knyttet til en ulik parametrisering av en gaussisk prior innenfor maximum a posteriori (MAP)-rammeverket: en Prior-Mean-estimator som krymper mot en LLM-estimert gjennomsnittsvektor, en Independent-Prior-estimator som legger til variabelspesifikke varianser estimert fra gjentatte LLM-genereringer, og en Correlated-Prior-estimator som bruker den fulle empiriske kovariansen på tvers av genereringene. Femti koeffisientvektorer trekkes fra hver av tolv ledende LLM-varianter (tre leverandører $\times$ to modellstørrelser $\times$ to moduser: resonnerende og ikke-resonnerende); LLM-varianten og prior-styrken λ velges deretter samtidig ved hjelp av 5-fold kryssvalidering på treningsdata fra kildedomenet. Prosedyren evalueres på fem tabulære datasett (Boston Housing, Cancer, Heart Disease, Diabetes og Adult), der hvert datasett deles i et kildedomene og et OOD-evalueringssett basert på en strukturell variabel.

De prior-informerte estimatorene gjør det bedre enn en regularisert basislinje på fire av de fem datasettene. Regresjonsoppgavene viser størst forbedring: $-15{,}8\%$ OOD-MSE på Boston Housing og $-3{,}9\%$ på Cancer, med parede bootstrap-konfidensintervaller som ekskluderer null for alle prior-familiene. På Heart Disease og Diabetes er OOD-AUC i praksis uendret, men F1, sensitivitet og sannsynlighetskalibrering forbedres signifikant. Adult, det største datasettet med det mildeste skiftet, viser den forventede sikkerhetsmekanismen: kryssvalidering driver prior-styrken mot null, og de prior-informerte estimatorene gjenfinner basislinjen. Seleksjonsprosedyren på tvers av leverandører havner konsekvent nær det beste OOD-resultatet per datasett, noe som tyder på at LLM-baserte koeffisient-priorer, filtrert gjennom kryssvalidering og leverandørfamilie-seleksjon, kan være en lovende kilde til strukturert domeneinformasjon for klassiske statistiske modeller under distribusjonsskift.

Acknowledgements

I am grateful to my supervisors Hugo Lewi Hammer and Kristoffer Herland Hellton of OsloMet for proposing this project and helping me shape its direction.

I would like to thank my main supervisor, Hugo, for his knowledge and guidance throughout the past year. His contributions have been crucial to the making of this thesis.

I would also like to thank my co-supervisor, Kristoffer, for the help and expertise he provided throughout the year.

Thank you also to my internal supervisor Hilde Vinje for her comments on the thesis and for her support throughout.

Finally, I want to thank my family for their continued support, and my girlfriend Matilde for her patience and support. Thanks also to Simula for providing a workspace throughout the year.

Contents

List of Abbreviations xiii

1 Introduction **1**
1.1 Prior Elicitation and Large Language Models 2
1.2 Related Work on LLMs in Predictive Modeling and Prior Elicitation 4

2 Underlying Concepts **7**
2.1 Linear Regression 7
2.2 Logistic Regression 8
2.3 L2 Regularization and Ridge Regression 9
2.4 Bayesian Regression and MAP Estimation 10

3 Methods **14**
3.1 Study Design 14
3.2 Domain Splits and Preprocessing 15
3.3 LLM Prior Elicitation 16
3.3.1 Prompt Construction 16
3.3.2 Sampling and Validation 17
3.3.3 Estimation of Prior Quantities from Repeated LLM Samples . 18
3.4 Models 18
3.4.1 Baseline Models 19
3.4.2 Prior-Mean Estimator 19
3.4.3 Independent-Prior Estimator 20
3.4.4 Correlated-Prior Estimator 20
3.4.5 Mixed Model 21
3.5 Training and Evaluation 21
3.5.1 Cross-Validation and λ Tuning 21
3.5.2 Provider Selection 22
3.5.3 Evaluation Metrics 22
3.5.4 Uncertainty Estimation 22
3.6 Use of Generative AI 23

4 Datasets **24**
4.1 Boston Housing 24
4.2 Cancer 26

4.3 Heart Disease 27
4.4 Diabetes 28
4.5 Adult 29

5 Results **31**
5.1 Overview 31
5.2 Per-Dataset Results 31
5.2.1 Boston Housing 31
5.2.2 Cancer 35
5.2.3 Heart Disease 39
5.2.4 Diabetes 44
5.2.5 Adult 48
5.3 LLM Dynamics 51
5.4 Summary 54

6 Discussion **55**
6.1 Predictive Performance 55
6.2 The Role of Model Selection 56
6.3 What the LLMs Actually Know 57
6.4 Limitations and Future Work 58
6.5 Future Work 60

7 Conclusion **61**

Bibliography **62**

A Prompt and Reasoning-Mode Response Example **65**
A.1 System prompt 65
A.2 User prompt 66
A.3 Response: reasoning 67
A.4 Response: coefficients 70

B Adult Prior Diagnostics **71**

List of Figures

3.1 Pipeline overview. The data branch (top) defines the ID/OOD split, fits preprocessing on the ID set only, and produces an ID training pool, an ID test set, and an OOD set. The elicitation branch (middle) feeds dataset context into prompt construction, draws repeated LLM samples per provider, validates them against the JSON schema, and estimates the prior quantities $\widehat{\mu}$, $\widehat{\sigma}^2$, $\widehat{\Sigma}$. The modeling branch (bottom) fits each family using the prior parameters and ID training data, tunes λ by 5-fold CV on the ID training pool, selects the winning LLM variant per family by best mean-CV score, and evaluates on the held-out ID test and OOD sets with paired-Δ-vs-baseline bootstrap CIs as the inferential test against the regularized baseline. 15

4.1 Boston Housing: distribution shift between younger-housing-stock source (AGE $\leq$ 70th percentile) and older-housing-stock OOD (AGE $>$ 70th percentile). Left: scatter on the two top-shifted features, INDUS (proportion of non-retail business acres) and NOX (nitric-oxide concentration); the two domains separate cleanly with the OOD set concentrated at high INDUS and high NOX. Right: per-feature Cohen's d, all 12 features sorted descending; nine of twelve features show $d > 0.5$ and seven of twelve show $d > 0.8$, indicating that the AGE-percentile split induces a coordinated multi-feature shift rather than a single-feature one. Vertical guides mark $d = 0.2$, 0.5, 0.8. 25

4.2 Cancer: distribution shift between U.S. South-region counties (source) and non-South counties (OOD). Left: scatter on the two top-shifted features, with each county colored by region. Right: per-feature Cohen's d, all 13 features sorted descending; orange bars highlight the two features chosen for the scatter. Vertical guides mark $d = 0.2$, 0.5, 0.8. 27

4.3 Heart Disease: distribution shift between Cleveland source and Hungary + Switzerland OOD. Left: scatter on the two top-shifted non-binary features. Right: per-feature Cohen's d, all 18 features sorted descending; orange bars highlight the two scatter features. Vertical guides mark $d = 0.2$, 0.5, 0.8. 28

4.4 Diabetes: distribution shift between younger-patient source (Age $\leq$ median) and older-patient OOD (Age $>$ median). Left: scatter on the two top-shifted features, colored by age cohort. Right: per-feature Cohen's d, all 7 features sorted descending. Vertical guides mark $d = 0.2, 0.5, 0.8$. 29
4.5 Adult: distribution shift between source (private-sector workers) and OOD (non-private). Left: scatter on the two top-shifted non-binary features, with each point colored by domain. Right: per-feature Cohen's d, all 20 features sorted descending; orange bars highlight the two features chosen for the scatter. Vertical guides mark the conventional small / medium / large effect-size landmarks ($d = 0.2, 0.5, 0.8$). . . . 30

5.1 Paired-Δ-vs-baseline on OOD test MSE, per family. Points are the deployment model's Δ relative to baseline; whiskers are the 95% percentile interval of the paired difference from 2,000 bootstrap resamples of the OOD test set. Vertical line at zero marks no improvement; whiskers fully on the negative side indicate significance. 32
5.2 Validation MSE as a function of prior weight λ (log scale), one curve per family at its deployment-winning LLM. Dashed lines mark each family's interior optimum. 33
5.3 Per-feature comparison of the LLM prior and the OOD-oracle coefficient. Blue circles: OOD-oracle coefficient. Open red circles with whiskers: prior mean $\pm 1\sigma$ across 50 LLM samples from the deployment-winning LLM. Gray segments connect the two markers per feature as a visual shrinkage axis. Both axes in source-standardized feature space. 34
5.4 Coefficient shrinkage on Boston Housing. Black filled circle: unregularized reference fit on the source split (anchor). Open red circle with whiskers: LLM prior mean $\pm 1\sigma$ (target). Colored filled circles: each model family's deployed coefficient at its tuned λ. Gray segment per feature shows the reference$\rightarrow$prior shrinkage axis. All in source-standardized feature space. 35
5.5 Paired-Δ-vs-baseline on OOD test MSE, per family. Points are the deployment model's Δ relative to baseline; whiskers are the 95% percentile interval of the paired difference from 2,000 bootstrap resamples of the OOD test set. Vertical line at zero marks no improvement; whiskers fully on the negative side indicate significance. 36
5.6 Validation MSE as a function of prior weight λ (log scale), one curve per family at its deployment-winning LLM. Dashed lines mark each family's interior optimum. 37

5.7 Per-feature comparison of the LLM prior and the OOD-oracle coefficient. Blue circles: OOD-oracle coefficient. Open red circles with whiskers: prior mean $\pm 1\sigma$ across 50 LLM samples from the deployment-winning LLM. Gray segments connect the two markers per feature as a visual shrinkage axis. Both axes in source-standardized feature space. 38
5.8 Coefficient shrinkage on Cancer. Black filled circle: unregularized reference fit on the source split. Open red circle with whiskers: LLM prior mean $\pm 1\sigma$. Colored filled circles: each model family's deployed coefficient at its tuned λ. Gray segment per feature shows the reference→prior shrinkage axis. All in source-standardized feature space. 39
5.9 Paired-Δ-vs-baseline on OOD test AUC, per family. Points are the deployment model's Δ relative to baseline; whiskers are the 95% percentile interval of the paired difference from 2,000 bootstrap resamples of the OOD test set. Vertical line at zero marks no improvement; whiskers fully on the positive side indicate significance. 41
5.10 Validation AUC as a function of prior weight λ (log scale), one curve per family at its deployment-winning LLM. Dashed lines mark each family's interior optimum. 41
5.11 Per-feature comparison of the LLM prior and the OOD-oracle coefficient. Blue circles: OOD-oracle coefficient. Open red circles with whiskers: prior mean $\pm 1\sigma$ across 50 LLM samples from the deployment-winning LLM. Gray segments connect the two markers per feature as a visual shrinkage axis. Both axes in source-standardized feature space. 42
5.12 Coefficient shrinkage on Heart Disease. Black filled circle: unregularized reference fit on the source split. Open red circle with whiskers: LLM prior mean $\pm 1\sigma$. Colored filled circles: each model family's deployed coefficient at its tuned λ. Gray segment per feature shows the reference→prior shrinkage axis. All in source-standardized feature space. 43
5.13 OOD calibration on Heart Disease. Top: 10-quantile reliability curves for the deployed L2-Logistic baseline and the deployed best-prior family on AUC (Correlated-Prior with Opus 4.5). Dotted diagonal is perfect calibration. Bottom: histograms of predicted probabilities under each model. Brier and LogLoss in the legend. 44
5.14 Paired-Δ-vs-baseline on OOD test AUC, per family. Points are the deployment model's Δ relative to baseline; whiskers are the 95% percentile interval of the paired difference from 2,000 bootstrap resamples of the OOD test set. Vertical line at zero marks no improvement; whiskers fully on the positive side indicate significance. 46
5.15 Validation AUC as a function of prior weight λ (log scale), one curve per family at its deployment-winning LLM. Dashed lines mark each family's interior optimum. 46

5.16 Per-feature comparison of the LLM prior and the OOD-oracle coefficient. Blue circles: OOD-oracle coefficient. Open red circles with whiskers: prior mean $\pm 1\sigma$ across 50 LLM samples from the deployment-winning LLM. Gray segments connect the two markers per feature as a visual shrinkage axis. Both axes in source-standardized feature space. 47
5.17 Coefficient shrinkage on Diabetes. Black filled circle: unregularized reference fit on the source split. Open red circle with whiskers: LLM prior mean $\pm 1\sigma$. Colored filled circles: each model family's deployed coefficient at its tuned λ. Gray segment per feature shows the reference→prior shrinkage axis. All in source-standardized feature space. 47
5.18 OOD calibration on Diabetes. Top: 10-quantile reliability curves for the deployed L2-Logistic baseline and the deployed best-prior family on AUC (Independent-Prior with Opus 4.5). Dotted diagonal is perfect calibration. Bottom: histograms of predicted probabilities under each model. Brier and LogLoss in the legend. 48
5.19 Paired-Δ-vs-baseline on OOD test AUC, per family. Points are the deployment model's Δ relative to baseline; whiskers are the 95% percentile interval of the paired difference from 2,000 bootstrap resamples of the OOD test set. Vertical line at zero marks no improvement; whiskers fully on the positive side indicate significance. 50
5.20 Validation AUC as a function of prior weight λ (log scale), one curve per family at its deployment-winning LLM. Dashed lines mark each family's interior optimum. 50
5.21 Per-LLM mean reasoning-vs-non-reasoning lift, expressed as the per-dataset normalised lift in reasoning mode minus the same in non-reasoning mode, averaged across the four prior-helps datasets (Adult excluded; see body text). Positive bars indicate that reasoning mode produces a higher mean OOD lift than non-reasoning mode for that LLM; negative bars indicate the opposite. 51
5.22 Per-dataset OOD lift over baseline: worst achievable across all 48 (variant, family) deployments at their CV-tuned λ (lower tick of each whisker, vermillion), best achievable across the same pool (upper tick, bluish green), and the lift produced by the deployed pipeline (black diamond marker). 52
5.23 Cross-LLM coefficient agreement on the two regression datasets (Cancer and Boston Housing). Each panel shows the prior-mean coefficient for all twelve frontier LLM variants (three providers $\times$ two model sizes $\times$ {non-reasoning, reasoning}) per feature, in source-standardized feature space. Filled circles: non-reasoning. Open circles: reasoning. Gray vertical tick: across-variant mean. Off-panel points are annotated with their value at the boundary. 53

5.24 Cross-LLM coefficient agreement on the two classification datasets (Heart Disease and Diabetes). Same conventions as Figure 5.23. Heart Disease shows numeric features only. 53

B.0.1 Per-feature comparison of the LLM prior and the OOD-oracle coefficient on Adult. Blue circles: OOD-oracle coefficient. Open red circles with whiskers: prior mean $\pm 1\sigma$ across 50 LLM samples from the deployment-winning LLM. Gray segments connect the two markers per feature as a visual shrinkage axis. Both axes in source-standardized feature space. 72

B.0.2 Coefficient shrinkage on Adult. Black filled circle: unregularized reference fit on the source split. Open red circle with whiskers: LLM prior mean $\pm 1\sigma$. Colored filled circles: each model family's deployed coefficient at its tuned λ. Gray segment per feature shows the reference$\rightarrow$prior shrinkage axis. All in source-standardized feature space. 73

List of Tables

5.1 Boston Housing deployment-model comparison. Per-family CV-selected variant and λ, with ID and OOD test MSE, OOD 95% bootstrap CI, and paired-Δ-vs-baseline 95% CI. **Bold** marks paired-Δ CIs that exclude zero. 32

5.2 Cancer deployment-model comparison. Per-family CV-selected variant and λ, with ID and OOD test MSE, OOD 95% bootstrap CI, and paired-Δ-vs-baseline 95% CI. **Bold** marks paired-Δ CIs that exclude zero. 36

5.3 Heart Disease deployment-model comparison. Per-family CV-selected variant and λ, with ID and OOD test AUC, OOD 95% bootstrap CI, and paired-Δ-vs-baseline 95% CI. **Bold** marks paired-Δ CIs that exclude zero. 40

5.4 Heart Disease OOD threshold-0.5 metrics. F1 with 95% bootstrap CI, plus Sens (sensitivity) and Spec (specificity) point estimates for the deployment models in Table 5.3, and the paired-Δ F1 vs baseline 95% CI. **Bold** marks paired-Δ CIs that exclude zero; bold metric values mark the per-column best (excluding the Unreg. ref. row). 40

5.5 Diabetes deployment-model comparison. Per-family CV-selected variant and λ, with ID and OOD test AUC, OOD 95% bootstrap CI, and paired-Δ-vs-baseline 95% CI. **Bold** marks paired-Δ CIs that exclude zero. 45

5.6 Diabetes OOD threshold-0.5 metrics. F1 with 95% bootstrap CI, plus Sens and Spec point estimates for the deployment models in Table 5.5, and the paired-Δ F1 vs baseline 95% CI. **Bold** marks paired-Δ CIs that exclude zero; bold metric values mark the per-column best (excluding the Unreg. ref. row). 45

5.7 Adult deployment-model comparison. Per-family CV-selected variant and λ, with ID and OOD test AUC, OOD 95% bootstrap CI, and paired-Δ-vs-baseline 95% CI. **Bold** marks paired-Δ CIs that exclude zero. 49

5.8 Adult OOD threshold-0.5 metrics. F1 with 95% bootstrap CI, plus Sens and Spec point estimates for the deployment models in Table 5.7, and the paired-Δ F1 vs baseline 95% CI. **Bold** marks paired-Δ CIs that exclude zero; bold metric values mark the per-column best (excluding the Unreg. ref. row). 49

5.9 Headline OOD performance per dataset: the best prior-informed model and its percentage improvement over baseline on the primary metric (OOD MSE for regression, OOD AUC for classification). Positive percentages indicate the prior-informed model beats baseline (lower MSE / higher AUC). 54

List of Abbreviations

API Application Programming Interface

AUC Area Under the ROC Curve

BMI Body Mass Index

CI Confidence Interval

CV Cross-Validation

ECG Electrocardiogram

GPT Generative Pre-trained Transformer

ID In-Distribution

IQR Interquartile Range

JSON JavaScript Object Notation

LLM Large Language Model

MAP Maximum A Posteriori

ML Machine Learning

MSE Mean Squared Error

NMBU Norges miljø- og biovitenskapelige universitet (Norwegian University of Life Sciences)

OLS Ordinary Least Squares

OOD Out-of-Distribution

ROC Receiver Operating Characteristic

UCI University of California, Irvine (Machine Learning Repository)

Chapter 1

Introduction

Prediction models are usually developed under the assumption that the training data and the evaluation data are drawn from the same underlying distribution. Under this independent and identically distributed assumption, performance on held-out test data is taken as a reliable proxy for the real-world data the model will encounter after deployment. However, this assumption can be unrealistic in many applied settings, and performance may deteriorate when the population encountered at prediction time is different from the one used for training (J. Liu et al., 2023). For example, a clinical prediction model trained on patients from one hospital may perform worse when deployed at another hospital, because things like patient demographics, routines, disease prevalence, or treatment practices may differ between institutions.

These distribution changes are often described as a dataset shift or domain shift. There are several different types of shifts, such as covariate shift, which is a change in the distribution of input data from train to test, while the relationship between inputs and outputs remains the same. Label shift (or prior probability shift) is the opposite and occurs when the response variable changes frequencies but the features remain the same. Concept shift refers to a change in the relationship between the features and the outcome itself so that the same feature values can produce different responses across different populations. There is also a subpopulation shift or domain shift, where the whole population shifts from training to test. The example used earlier with the different hospitals is a case of domain shift, and it often contains several of the other shifts at once (J. Liu et al., 2023; Quiñonero-Candela et al., 2009).

Despite extensive work on this problem (Quiñonero-Candela et al., 2009), there is no generally reliable solution for maintaining predictive performance under a domain shift. Standard validation procedures can fail because the validation set is often drawn from the same source distribution as the training data. Several techniques have been developed in response (J. Liu et al., 2023), including domain adaptation methods that try to align source and target representations and ensemble methods that combine multiple models in the hope of producing predictions that are robust across distributional changes. The most well known practical approach to handling predictive uncertainty under distribution shift in modern ML is the use of model

ensembles (Ovadia et al., 2019). Deep ensembles, introduced by Lakshminarayanan et al. (2017), train a number of neural networks independently with different random initialisations and average their probabilistic predictions, with the variance across the ensemble taken as an uncertainty estimate. The paper's main finding is that this simple recipe produces as good or better-calibrated uncertainty than more involved Bayesian methods, and that the ensemble's predictive variance rises on inputs drawn from a distribution different from training, so the deployed model can at least alert when its prediction is unreliable.

However, ensemble methods can have limitations. They require fitting multiple models, can be computationally expensive, and, when based on neural networks or other complex learners, can be difficult to interpret because internal decision-making processes can be hard or impossible to understand. And although ensemble disagreement can indicate predictive uncertainty, it does not by itself improve the underlying predictions. One of the few known ways to actually improve predictions when the training data is limited or unrepresentative of the deployment population is to bring in information from outside the training set. So implementing external domain knowledge about which features should matter, in what direction, or with what magnitude. This then motivates a different approach: using prior knowledge to guide simple, interpretable prediction models in settings where the training data may not fully represent the deployment population. Unlike neural network ensembles, these models remain easy to inspect and inexpensive to fit, with the contributions of both the data and the prior visible in the final estimate.

1.1 Prior Elicitation and Large Language Models

In many applications, useful information can exist outside the training data, for example from previous studies, biological pathways, clinical expertise, or established domain theory (O'Hagan, 2019; Sokolov et al., 2016). When this information is reliable and appropriately encoded, it can improve estimation and prediction, especially in settings where the observed data are limited, noisy, or not fully representative of the target population (van de Schoot et al., 2021). The most established framework for incorporating prior information is in Bayesian modeling, where prior beliefs are expressed through a probability distribution over the model parameters and then updated using the observed data (Bishop, 2006; van de Schoot et al., 2021). A related approach is to incorporate prior information through regularization. In this view, the estimation procedure is penalized in a way that shrinks parameter estimates toward preferred values or structures (Hastie et al., 2009). Standard ridge regression, for example, shrinks coefficients toward zero, which can be interpreted as a generic prior belief that effects should be small unless strongly supported by the data (Bishop, 2006; Hoerl & Kennard, 1970).

This interpretation makes prior-informed regularization a natural framework for combining external knowledge with predictive modeling. Instead of shrinking or

pulling the coefficients toward the same uninformative target of zero, one can pull them toward values suggested by domain knowledge. Under domain shift, this may be useful because the source-domain data alone may overemphasize relationships that do not transfer to the target population, while domain knowledge can encode more stable information about plausible feature effects. Eliciting useful information from experts is a central part of constructing informative priors. O'Hagan (2019) describes expert knowledge elicitation as a structured and demanding process, in which the goal is to translate substantive domain knowledge into probability distributions or numerical quantities that are suitable for statistical analysis. One key limitation is that domain experts are usually not statisticians, they may be medical professionals, biologists, or specialists from another applied field, so the elicitor must guide them through concepts such as uncertainty, probability, and numerical effect sizes. As a result, expert elicitation tends to be time-consuming, difficult to scale, and after all the effort it has no guarantee of producing useful priors (O'Hagan, 2019).

These practical constraints have motivated recent interest in using Large Language Models (LLMs) as a possible alternative. LLMs are neural language models trained on large-scale text data and adapted to generate responses to natural language prompts (Bommasani et al., 2022; Brown et al., 2020). Through this training they can encode broad statistical associations across many domains (Bommasani et al., 2022; Minaee et al., 2024), and they can be queried repeatedly, applied across many domains and feature sets, and constrained to return structured numerical outputs. These properties can make them attractive as proxies for domain expertise in settings where human elicitation would be too costly or too slow. Evidence from medical reasoning suggests that they can sometimes provide better domain-level judgments than humans; Goh et al. (2024) found that ChatGPT alone outperformed physicians alone, and also outperformed physicians using ChatGPT on a diagnostic reasoning task. But LLM-generated priors do come with important limitations. LLMs can hallucinate, are sensitive to prompt wording, and may reflect biases in their training data (Bender et al., 2021; Gallegos et al., 2024; Huang et al., 2025; P. Liu et al., 2023). These limitations are especially important when the output is a numerical coefficient, because errors in scale, sign, or uncertainty can directly affect the fitted model. So this means that the LLMs should not be treated as perfect substitutes for human experts, but it does suggest that they can potentially contain useful domain information that can be elicited and incorporated into predictive models if used correctly.

This thesis builds on the broader idea proposed by Riegler et al. (2025), a recent paper by the supervisors of this project, that LLMs can be used to construct informative Gaussian priors for the coefficients of linear and logistic regression models. The present work takes that idea as inspiration and develops it in three ways. First, the prior is incorporated through MAP estimation rather than full posterior inference, which keeps the prior-informed estimators directly comparable to standard regularized regression so they can be judged directly on prediction performance metrics. Second, three increasingly structured ways of encoding the LLM information

are used: shrinkage toward a prior mean only, toward a mean with feature-specific variances, and toward a mean with the full empirical covariance across repeated LLM samples. Third, the pipeline selects across multiple LLM providers by cross-validation on the source-domain training data, using the same criterion that tunes the absolute prior strength, so that the final result does not depend on any single elicitation source.

1.2 Related Work on LLMs in Predictive Modeling and Prior Elicitation

The use of LLMs as sources of domain information for predictive modeling is still a relatively new research direction, but several notable methods have been developed in the last couple of years. Most existing work either uses LLMs directly as predictors, uses them to construct or modify features, or uses them to provide prior information for statistical models. Only a small part of this literature explicitly studies prediction under distribution shift.

As mentioned above, the present thesis builds on Riegler et al. (2025), in which several LLMs are asked to suggest informative Gaussian prior distributions for coefficients in linear and logistic regression models. The authors find that LLMs are often able to identify the correct direction of feature effects, but that the calibration of prior uncertainty is more difficult. In particular, directly elicited prior widths may be too narrow or poorly aligned with the observed data. So there are two takeaways for this thesis: LLMs may carry useful information about feature-response relationships, but their stated calibration of uncertainty cannot be trusted at face value, which motivates an external control on how strongly the prior is imposed on the fitted model. Riegler et al. (2025) also identify out-of-distribution prediction as a possible future use case for LLM-generated priors, and this thesis takes up that direction as its main empirical focus.

Capstick et al. (2025) propose AutoElicit, a framework for expert prior elicitation in predictive modeling. AutoElicit queries an LLM through many automatically rephrased prompts and combines the resulting per-prompt Gaussians into a mixture-of-Gaussians prior over the coefficients of Bayesian linear and logistic regression models. AutoElicit is the closest methodological precedent to this thesis, it shows that LLM-generated prior information can improve predictive models compared with uninformative priors, and it establishes the general approach of using repeated LLM queries to construct empirical estimates of prior parameters. AutoElicit's main evaluation is on in-distribution data, though its performance under domain shift has subsequently been measured by Rezaeimanesh and Ghassemi (2026) (discussed below) as a comparison for their token-probability method.

Zhang et al. (2025) propose LLM-Lasso, which asks the LLM for feature-specific penalty scores rather than coefficient priors. These scores are then used to guide Lasso regularization. LLM-Lasso illustrates an alternative access pattern: the LLM's

domain knowledge guides the regularization strength on each coefficient rather than the coefficient prior itself.

Zeng et al. (2024) study whether LLM embeddings can improve test-time adaptation under tabular $(Y \mid X)$-shifts. Rather than extracting coefficient priors, they serialize tabular rows into text, obtain LLM embeddings, and train prediction models on those representations. Their experiments cover 7,650 source–target pairs and compare against a large set of tabular baselines. Their main finding is that LLM embeddings alone provide inconsistent robustness, but models trained on LLM embeddings can adapt well to the target domain when even a small number of labeled target examples are available.

Requeima et al. (2024) introduce *LLM Processes*, which treats the LLM itself as an in-context probabilistic regressor. Numerical data and a natural-language description of the prediction problem are passed to the LLM, which then produces a predictive distribution at arbitrary query locations. The authors evaluate LLM Processes on synthetic and real-world regression and forecasting tasks, including image reconstruction and black-box function optimization, and show that it produces well-calibrated uncertainty estimates competitive with Gaussian Processes. Their results support the broader idea that natural-language prior knowledge can usefully shape numerical predictions, and that LLMs encode useful prior information about predictive tasks.

Amin et al. (2026) study heart disease prediction using a hybrid ML–LLM voting-fusion framework, comparing tree-based ensemble models, LLM predictions, and a combined ML–LLM fusion model. Their results show that the ML ensemble performs much better than LLM prediction alone, while the hybrid fusion gives a small additional gain over the ML ensemble. This provides empirical support for the broader principle that LLMs are most useful as structured inputs to classical models rather than as final predictors in their own right.

Rezaeimanesh and Ghassemi (2026) propose LoID, which is highly relevant because it independently explores a very similar method and goal to this thesis by using LLM-derived information as priors for tabular prediction under distribution shift. This paper was first released as a smaller preprint in late January 2026 and then the second version was released in late April 2026, and was not discovered until late in the writing process. It was therefore not used as inspiration for the method developed here. LoID extracts informative priors for Bayesian logistic and linear regression from token-level LLM predictions rather than from generated coefficient values. It compares the model's preference for semantically opposed tokens such as "positive" and "negative" across multiple prompt phrasings, using the average preference as a prior mean and the empirical variation across phrasings as an uncertainty estimate. The authors evaluate LoID on fifteen tabular datasets under synthetic OOD splits constructed by partitioning each dataset on a single feature value or category, and report improvements over both a non-LLM-informed OOD baseline (P_OOD: logistic/linear regression on OOD train) and a weakly-informative Bayesian baseline (N(0,1) prior),

while also outperforming AutoElicit and LLM Processes on most datasets.

Taken together, these papers show that LLMs can support predictive modeling in several ways: as direct predictors, as sources of feature transformations, as guides for regularization, as representation generators, or as sources of prior information. Among these, LoID is the closest in spirit with the method in this thesis; both use LLM-derived priors for regularized regression under distribution shift. This thesis differs in the elicitation mechanism (repeated coefficient sampling rather than token log-probabilities), in the encoding of prior uncertainty (allowing full covariance structure across coefficients rather than only independent variances), and in how the absolute prior strength is set (cross-validation on source-domain data rather than fixed global hyperparameters).

This thesis is guided by three research questions:

1. Can prior means estimated from repeated LLM coefficient samples improve OOD predictive performance for regularized linear and logistic regression, compared with standard regularization toward zero?

2. Does estimating prior uncertainty from the empirical distribution of LLM samples, through feature-specific variances or full covariance structure, improve performance beyond using only the empirical prior mean?

3. Under what conditions do LLM-informed priors help, have little effect, or harm predictive performance?

Chapter 2

Underlying Concepts

2.1 Linear Regression

Linear regression is a foundational predictive model that describes the relationship between a continuous response and one or more features. It plays two roles in this thesis: it is the conceptual building block that all later models extend from, and it is used as one of the unregularized reference models in the experiments, which the regularized and prior-informed models are compared against.

The standard linear regression model is written as

$$y = X\beta + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2 I),$$

where $y \in \mathbb{R}^n$ is the response vector for n observations, $X \in \mathbb{R}^{n \times p}$ is the design matrix containing p features, $\beta \in \mathbb{R}^p$ is the vector of regression coefficients, and $\varepsilon \in \mathbb{R}^n$ is a vector of independent and identically distributed Gaussian noise terms with variance σ^2. The Gaussian noise assumption is what allows the model to be given a probabilistic interpretation later in this chapter.

The classical estimator for β is ordinary least squares (OLS), which is defined as the minimiser of the residual sum of squares:

$$\widehat{\beta}_{\text{OLS}} = \arg\min_{\beta} \|y - X\beta\|^2.$$

This optimization problem has a closed-form solution,

$$\widehat{\beta}_{\text{OLS}} = (X^\top X)^{-1} X^\top y,$$

provided that $X^\top X$ is invertible (Hastie et al., 2009, §3.2).

OLS can equivalently be derived from a probabilistic standpoint. Under the Gaussian noise assumption above, the log-likelihood of β given the observed data is, up to additive constants that do not depend on β,

$$\log p(y \mid X, \beta, \sigma^2) = -\frac{1}{2\sigma^2} \|y - X\beta\|^2 + \text{const.}$$

Maximising this log-likelihood with respect to β is therefore equivalent to minimising the residual sum of squares, so the maximum likelihood estimator coincides with the OLS estimator (Hastie et al., 2009, §2.6). This equivalence will be important in Section 2.4, where introducing a prior distribution on β gives rise to maximum a posteriori (MAP) estimation, of which ridge regression and the prior-informed models in this thesis are special cases.

Despite its appeal, OLS has well-known limitations. When the features are highly correlated or when the number of features p is close to the number of observations n, the matrix $X^\top X$ becomes poorly conditioned, and the resulting coefficient estimates have high variance. The model is then prone to overfitting and generalises poorly to new data. OLS also offers no mechanism to incorporate external information about plausible coefficient values, since the estimator depends only on the observed data (Hastie et al., 2009, §3.2). Both of these limitations motivate the regularized estimators introduced in Section 2.3.

2.2 Logistic Regression

For a binary response, linear regression no longer works. The predicted values $x^\top\beta$ are unbounded and cannot be interpreted as probabilities, and the assumption of Gaussian noise is not compatible with a response that takes only the values 0 and 1. Logistic regression addresses this by modeling the conditional probability of the response rather than the response itself, and its role in this thesis mirrors that of the linear regression model: it serves both as a conceptual building block for the regularized classifiers introduced later and as one of the unregularized reference models in the experiments.

The model assumes that the conditional probability of the positive class is given by

$$P(y = 1 \mid x, \beta) = \sigma(x^\top\beta),$$

where $y \in \{0, 1\}$, $x \in \mathbb{R}^p$ is the feature vector for a single observation, $\beta \in \mathbb{R}^p$ is the vector of regression coefficients, and σ denotes the logistic (sigmoid) function

$$\sigma(z) = \frac{1}{1 + \exp(-z)}.$$

The sigmoid maps the real line into the interval $(0, 1)$, so that the linear combination $x^\top\beta$ (which is itself unbounded) produces a valid probability. The complementary probability of the negative class follows directly: $P(y = 0 \mid x, \beta) = 1 - \sigma(x^\top\beta)$.

Because the response is binary, each observation is modeled as a Bernoulli trial with success probability $\sigma(x_i^\top\beta)$. Assuming the observations are independent, the coefficients are estimated by maximum likelihood. Equivalently, $\widehat{\beta}$ is obtained by minimising the negative log-likelihood,

$$-\log L(\beta) = -\sum_i \left[y_i \log \sigma(x_i^\top\beta) + (1 - y_i) \log(1 - \sigma(x_i^\top\beta)) \right],$$

so that

$$\widehat{\beta}_{\text{MLE}} = \arg\min_{\beta} \left[-\log L(\beta) \right].$$

This objective is the analogue of the residual sum of squares in linear regression: in both cases, the estimator is the maximiser of the likelihood under the model's noise assumption (Hastie et al., 2009, §4.4). Unlike OLS, however, $-\log L(\beta)$ does not admit a closed-form minimiser, since σ is nonlinear in β. The estimate is instead obtained by numerical optimization, with standard methods including gradient descent and iteratively reweighted least squares. The objective is convex in β, so these methods converge to the global minimum.

Logistic regression shares the limitations of OLS and adds one of its own. When features are highly correlated, the estimated coefficients again exhibit high variance and the model is prone to overfitting. In addition, when the two classes are linearly separable in the training data, the maximum likelihood estimate does not exist: the negative log-likelihood can be reduced indefinitely by scaling the coefficients toward infinity, and the optimiser does not converge to a finite solution. As with linear regression, the model also offers no mechanism to incorporate external information about plausible coefficient values. These limitations motivate the regularized estimators introduced in the next section (Hastie et al., 2009, §4.4).

2.3 L2 Regularization and Ridge Regression

L2 regularization addresses limitations shared by the linear and logistic regression models introduced in the previous sections. Both estimators are sensitive to multicollinearity among features and to settings where the number of features p is large relative to the number of observations n, leading to high-variance coefficient estimates and poor generalization. The idea behind L2 regularization is to augment the loss function with a penalty term that shrinks the coefficients toward zero, reducing variance at the cost of introducing some bias. This trade-off can improve predictive performance on new data and is particularly relevant in this thesis, since variance reduction is even more important when the deployment distribution differs from the training distribution: a model that has fit idiosyncratic features of the training data will generalise badly under domain shift. L2-regularized models are widely used in high-dimensional applied problems, including genomic prediction and chemometric analysis, where the number of features typically exceeds the number of observations.

The penalty is defined as the squared Euclidean norm of the coefficient vector,

$$\|\beta\|^2 = \sum_j \beta_j^2,$$

where the sum runs over the components of β. Adding this penalty to the squared-error loss gives the ridge regression objective,

$$\widehat{\beta}_{\text{Ridge}} = \arg\min_{\beta} \left[\|y - X\beta\|^2 + \frac{\lambda}{2}\|\beta\|^2 \right],$$

where $\lambda \geq 0$ is the regularization strength controlling the trade-off between fit and shrinkage.

Setting the gradient of the objective to zero and solving for β yields the closed-form ridge estimator,

$$\widehat{\beta}_{\text{Ridge}} = \left(X^\top X + \frac{\lambda}{2} I\right)^{-1} X^\top y.$$

For any $\lambda > 0$, the matrix $X^\top X + \frac{\lambda}{2} I$ is invertible regardless of the conditioning of X, which directly addresses the multicollinearity issue affecting the unregularized models in Sections 2.1 and 2.2 (Hoerl & Kennard, 1970). Ridge regression produces uniform shrinkage toward zero: the penalty treats all components of β equivalently, so every coefficient is shrunk by the same factor.

The same L2 penalty can be applied to logistic regression, with the same conceptual effect. Adding the penalty to the negative log-likelihood from Section 2.2 yields the L2-regularized logistic regression objective,

$$\widehat{\beta}_{\text{L2}} = \arg\min_{\beta} \left[-\log L(\beta) + \frac{\lambda}{2} \|\beta\|^2 \right].$$

Just as in the unregularized case, this objective does not admit a closed-form solution and is obtained by numerical optimization. The penalty makes the objective strictly convex, however, which guarantees a unique finite minimum even when the classes are linearly separable, addressing the second limitation noted in Section 2.2.

Across both regression and classification, L2 regularization trades a small amount of bias for a substantial reduction in variance, and this trade-off typically improves predictive performance on unseen data. The shrinkage target of zero is generic in the sense that it does not encode any specific knowledge about the problem; ridge simply expresses a default preference for small coefficients in the absence of other information. Section 2.4 reframes ridge regression as the MAP estimator under a particular Gaussian prior, generalises this to Gaussian priors with arbitrary mean and precision structure, and provides the theoretical foundation for the prior-informed estimators developed in Chapter 3.

2.4 Bayesian Regression and MAP Estimation

The linear and logistic regression models introduced in the previous sections were treated from a frequentist perspective: the coefficient vector β was regarded as a fixed but unknown quantity, and estimation reduced to choosing a single value of β that minimized a loss function. The Bayesian view of regression takes a different starting point. It treats β as an uncertain quantity, described by a probability distribution that represents what is believed about its value before observing the data. This distribution is called the prior. Combining the prior with the likelihood, which describes how the data depend on β, yields the posterior (the updated distribution over β after the data have been observed) through an application of Bayes' rule. This thesis uses the

Bayesian framework to motivate the form of the penalty introduced in Section 2.3, but estimates β as a point rather than retaining the full posterior (Hastie et al., 2009, §8.3). The estimator used is the MAP estimator, which corresponds to the value of β at which the posterior density is maximized.

Bayes' rule expresses the posterior distribution over β as

$$p(\beta \mid y, X) = \frac{p(y \mid X, \beta) \cdot p(\beta)}{p(y \mid X)}.$$

Here, $p(y \mid X, \beta)$ is the likelihood, introduced in Section 2.1 for linear regression and Section 2.2 for logistic regression, $p(\beta)$ is the prior distribution over the regression coefficients, and $p(\beta \mid y, X)$ is the posterior distribution after observing the data. The denominator

$$p(y \mid X) = \int p(y \mid X, \beta)\, p(\beta)\, d\beta$$

is the marginal likelihood, obtained by integrating the joint density over β. It acts as a normalising constant that makes the posterior integrate to one. Because it does not depend on β, it does not affect the location of the posterior maximum, and Bayes' rule is therefore often written in the proportional form

$$p(\beta \mid y, X) \propto p(y \mid X, \beta) \cdot p(\beta).$$

This is the form used throughout the remainder of this section, since MAP estimation only requires the value of β at which the posterior is maximised, not the full normalised density.

The MAP estimator is then the value of β that maximises the posterior $p(\beta \mid y, X)$. Because the logarithm is monotonic, this is equivalent to maximising the log-posterior, which by Bayes' rule decomposes into the sum of the log-likelihood and the log-prior up to a constant in β:

$$\widehat{\beta}_{\text{MAP}} = \arg\max_{\beta} \log p(\beta \mid y, X) = \arg\max_{\beta} \left[\log p(y \mid X, \beta) + \log p(\beta)\right].$$

Equivalently, this can be written as a minimisation problem by changing the sign:

$$\widehat{\beta}_{\text{MAP}} = \arg\min_{\beta} \left[-\log p(y \mid X, \beta) - \log p(\beta)\right].$$

This has exactly the form of a penalized loss function: the negative log-likelihood acts as the data-fitting loss, while the negative log-prior acts as the penalty term.

To make the connection to ridge regression concrete, consider linear regression with Gaussian noise (Section 2.1) together with a zero-mean Gaussian prior on β with covariance $\tau^2 I$,

$$\beta \sim \mathcal{N}(0, \tau^2 I).$$

The posterior under this prior takes the proportional form

$$p(\beta \mid y, X) \propto p(y \mid X, \beta) \cdot p(\beta).$$

The negative log-prior is, up to additive constants that do not depend on β,

$$-\log p(\beta) = \frac{1}{2\tau^2}\|\beta\|^2 + \text{const.}$$

Setting $\lambda = 1/\tau^2$ gives exactly the ridge penalty $(\lambda/2)\|\beta\|^2$ introduced in Section 2.3, and the MAP estimator is therefore the minimiser of the resulting penalized objective,

$$\widehat{\beta}_{\text{MAP}} = \arg\min_{\beta} \left[\|y - X\beta\|^2 + \frac{\lambda}{2}\|\beta\|^2 \right],$$

which is exactly the ridge objective from Section 2.3. Ridge regression can therefore be interpreted as MAP estimation under the prior $\beta \sim \mathcal{N}(0, (1/\lambda)I)$ (Hastie et al., 2009, §3.4), and the two estimators coincide:

$$\widehat{\beta}_{\text{Ridge}} = \widehat{\beta}_{\text{MAP}}.$$

This interpretation makes the role of λ explicit: λ controls the inverse of the prior variance, called the precision. A large value of λ corresponds to a tight prior concentrated around zero, producing stronger shrinkage, while a small value of λ corresponds to a weaker prior and therefore less shrinkage. In the limit $\lambda \to 0$, the prior becomes increasingly diffuse and the estimator recovers the unregularized OLS estimator from Section 2.1.

In this thesis, the Bayesian framework is used to motivate the estimator rather than to perform full Bayesian inference. MAP estimation returns a single point estimate, namely the posterior mode, rather than retaining the full posterior distribution. This is appropriate here because the evaluation metrics are based on point predictions, the research question concerns predictive lift rather than uncertainty quantification, and MAP keeps the prior-informed models directly comparable to the regularized baselines from Section 2.3: only the prior specification changes, while the inference procedure remains the same.

The Gaussian prior used so far has had zero mean and covariance $\tau^2 I$, but the same MAP framework applies to any *multivariate* Gaussian prior. The multivariate Gaussian extends the univariate Gaussian to a vector-valued quantity: it is parameterized by a mean vector $\mu \in \mathbb{R}^p$, which locates the distribution, and a (symmetric, positive-definite) covariance matrix $\Sigma \in \mathbb{R}^{p \times p}$, whose diagonal entries give the variance of each component of β and whose off-diagonal entries give the covariances between components.

Consider a prior on β of this form, with arbitrary mean μ and arbitrary covariance Σ,

$$\beta \sim \mathcal{N}(\mu, \Sigma).$$

Combining this prior with the likelihood gives the posterior

$$p(\beta \mid y, X) \propto p(y \mid X, \beta) \cdot p(\beta).$$

The negative log-prior is, up to additive constants that do not depend on β,

$$-\log p(\beta) = \frac{1}{2}(\beta - \mu)^\top \Sigma^{-1}(\beta - \mu) + \text{const.}$$

The inverse covariance $P = \Sigma^{-1}$ is called the *precision matrix*, and is the matrix that naturally appears in the penalty. Writing $\Sigma = (1/\lambda)P^{-1}$ factors out a global regularization strength λ, giving the negative log-prior the same form as in the ridge case,

$$-\log p(\beta) = \frac{\lambda}{2}(\beta - \mu)^\top P(\beta - \mu) + \text{const.}$$

The MAP estimator is then the minimiser of the resulting penalized objective,

$$\widehat{\beta}_{\text{MAP}} = \arg\min_{\beta} \left[-\log p(y \mid X, \beta) + \frac{\lambda}{2}(\beta - \mu)^\top P(\beta - \mu) \right].$$

Ridge regression corresponds to the special case $\mu = 0$ and $P = I$ (equivalently, $\Sigma = (1/\lambda)I$), which recovers the ridge penalty derived above. More generally, μ specifies the shrinkage target and P (equivalently Σ^{-1}) specifies the structure of the shrinkage, encoding which coefficients are more uncertain than others and how they covary. The prior-informed estimators used in this thesis correspond to specific choices of μ and P derived from LLM samples, and are introduced in Chapter 3.

Chapter 3

Methods

3.1 Study Design

This thesis investigates whether prior information elicited from an LLM can improve the predictive performance in OOD prediction tasks; the ID setting is examined briefly but it is not the central focus. The goal is to test different ways to encode information from the LLM into prediction models and to compare them against a baseline of ridge regression / L2-regularized logistic regression, whose penalty corresponds to regularization towards zero.

To study this, a set of regularized regression models is used, all based on the same form of MAP estimation but with different parameters and information encoded from the LLM samples: informative prior means, feature-specific variances, and a full covariance matrix, all estimated from the empirical distribution of repeated LLM samples. The same standard pipeline is applied across five datasets from different domains and sample sizes, with identical preprocessing and splits, and the LLM provider per family is selected based on mean cross-validation performance on the source training pool.

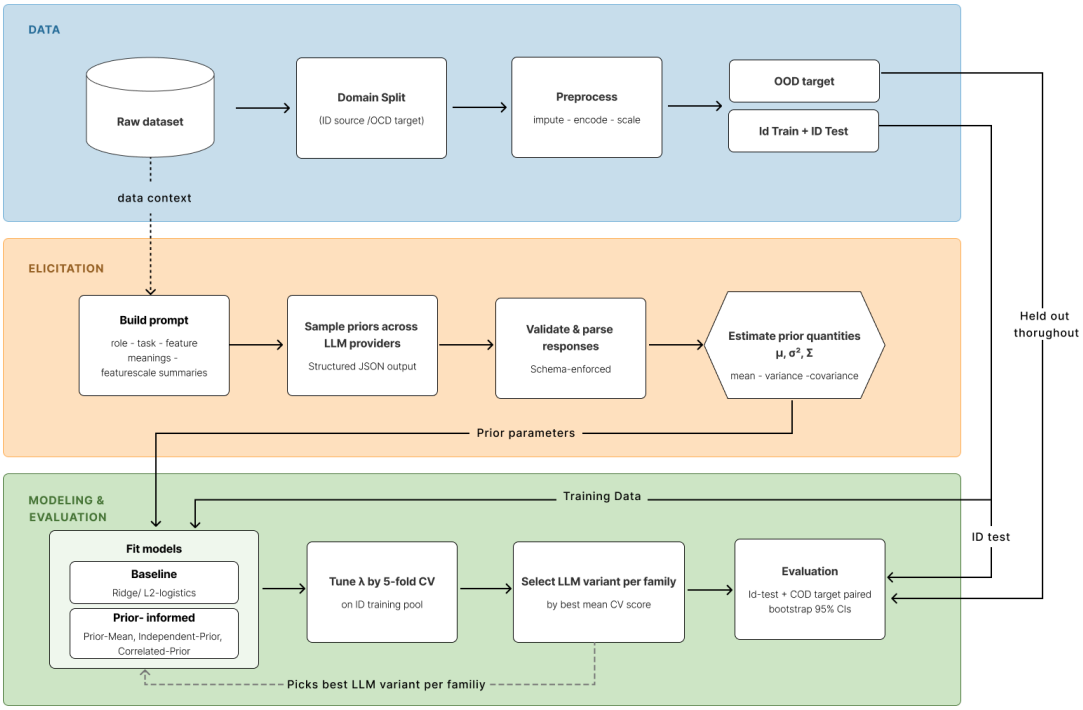


Figure 3.1: *Pipeline overview. The data branch (top) defines the ID/OOD split, fits preprocessing on the ID set only, and produces an ID training pool, an ID test set, and an OOD set. The elicitation branch (middle) feeds dataset context into prompt construction, draws repeated LLM samples per provider, validates them against the JSON schema, and estimates the prior quantities $\widehat{\mu}$, $\widehat{\sigma}^2$, $\widehat{\Sigma}$. The modeling branch (bottom) fits each family using the prior parameters and ID training data, tunes λ by 5-fold CV on the ID training pool, selects the winning LLM variant per family by best mean-CV score, and evaluates on the held-out ID test and OOD sets with paired-Δ-vs-baseline bootstrap CIs as the inferential test against the regularized baseline.*

3.2 Domain Splits and Preprocessing

Every dataset in the study is run through the same source/OOD pipeline. A domain-defining variable is chosen for each dataset (e.g. a categorical role indicator or a percentile cut on a continuous feature), the data are divided into an in-domain (ID) source set and an out-of-domain (OOD) set on that variable, and the splitter itself is excluded from the feature matrix to prevent leakage from domain identity into the model. The ID set is then divided into an 80% training pool and a 20% held-out test set. The 80% training pool is the only data used for hyperparameter tuning and model fitting; the 20% ID test set and the OOD set are held out throughout tuning and only used at evaluation time.

Preprocessing is standardized across datasets. Features with high source-domain

missingness (typically a 50% threshold) are removed, and remaining missing values are imputed with the source-domain median. All numeric features are then scaled to zero mean and unit variance. Categorical features are one-hot encoded with categories aligned across source and OOD. The imputer and scaler are fitted on the training fold during cross-validation (and on the full 80% pool at deployment) and then applied to the held-out splits.

3.3 LLM Prior Elicitation

The LLM was used as a structured source of domain-informed prior beliefs based on the prediction task, the feature set, and the dataset context. Repeated elicitation calls were submitted through each provider's API, and the returned outputs were parsed and validated before being used to compute the prior mean $\widehat{\mu}$ and covariance $\widehat{\Sigma}$ that feed the prior-informed models in Section 3.4. Priors were elicited from three providers with two different models from each, one larger and one smaller: OpenAI GPT-5.1 and GPT-5-mini, Anthropic Opus 4.5 and Haiku 4.5, and Google Gemini 2.5 Pro and 2.5 Flash, each of them in two different modes: non-reasoning and reasoning (twelve variants in total). The LLM-side cross-variant selection rule for picking which variant's prior to deploy is described in Section 3.5.2.

3.3.1 Prompt Construction

Each prompt consisted of two parts: a *system prompt* containing stable elicitation instructions, and a *user prompt* containing dataset-specific content. This separation ensured that the core elicitation methodology remained consistent across datasets while allowing the domain context and feature information to vary.

The system prompt assigned the LLM a domain-appropriate role (e.g., epidemiologist, cardiologist, demographic specialist), specified that coefficients should be given in raw feature units, that no intercept should be included, and that the output must conform to a fixed JSON schema. The intercept was excluded because it is determined by the baseline rate of the response and the feature centering, neither of which requires domain expertise about feature–response relationships. For the sampling mode used to generate repeated draws, the system prompt additionally instructed the model to return one plausible draw from its belief distribution rather than collapsing to a single average.

The user prompt supplied the modeling task (e.g., logistic regression for heart disease), domain context describing the characteristics of the training population (including a request that priors capture broad, transferable structural effects rather than local artifacts), feature meanings with explicit descriptions, and quantitative feature-scale summaries (median, interquartile range, and central 95% range) computed from the ID training pool only, never from the held-out ID test set or the OOD set. These summaries enabled the LLM to reason about plausible coefficient

magnitudes rather than guessing from feature names alone.

A deliberate design choice was made regarding domain context: prompts described the characteristics of the source (training) domain but did not reveal the specific out-of-domain target population. For example, the Cancer prompt stated that the source domain consisted of U.S. South-region counties, without specifying that the OOD evaluation would use non-Southern counties. This avoided leaking information about the evaluation setting into the prior, ensuring that any OOD improvements reflect genuine domain knowledge rather than targeted optimization for a known shift. Since models were fitted on standardized features, the raw priors were converted to standardized space by multiplying each coefficient by the corresponding feature standard deviation.

Across all providers and modes, the JSON output is constrained by both an explicit format instruction in the prompt and by a JSON-schema constraint passed in the API call: OpenAI's `json_schema` strict mode, Anthropic's tool input schema with forced `tool_choice`, and Google's `response_schema` together with an `application/json` response MIME type. The one exception is Anthropic's extended-thinking (reasoning) mode, where the API forbids combining extended thinking with forced tool use; for that path the schema is enforced via the prompt only and the parser recovers the JSON from the text output.

Some informal testing was conducted on prompt phrasing and structure during development, but this was deliberately kept limited to avoid overfitting the elicitation pipeline to the specific datasets and domain shifts used in this study. The final pipeline uses a single prompt style for every LLM and every dataset.

3.3.2 Sampling and Validation

For each (dataset, provider, mode) cell, $M = 50$ prior samples were drawn at temperature 0.8. The temperature parameter controls the randomness of the LLM's token sampling: a value of 0 produces near-deterministic output, while higher values increase variability. A temperature of 0.8 was chosen to encourage meaningful diversity across samples while avoiding the incoherence that can arise at very high temperatures. Using 50 samples provides stable estimates of both the prior mean and the covariance structure, which is particularly important for the Correlated-Prior model where the number of samples should comfortably exceed the number of features. Even more samples would be ideal but, because of cost and time, the count was limited to 50.

Returned responses were required to follow a fixed JSON structure and were validated against the feature set. Outputs with missing keys, extra keys, or invalid numeric values were rejected, and additional calls were submitted in their place; the elicitation loop continued until $M = 50$ valid samples had been accumulated, subject to a fixed upper bound on the total number of attempts to guard against runaway costs from a misbehaving LLM.

3.3.3 Estimation of Prior Quantities from Repeated LLM Samples

The $M = 50$ validated samples are used to estimate the prior quantities that feed the prior-informed estimators in Section 3.4. The samples are denoted

$$\beta^{(1)}, \beta^{(2)}, \ldots, \beta^{(M)} \in \mathbb{R}^p,$$

and are treated as draws from the LLM's implicit belief distribution over the regression coefficients. From these samples, the prior mean vector is estimated as

$$\widehat{\mu} = \frac{1}{M} \sum_{m=1}^{M} \beta^{(m)}.$$

The feature-specific prior variance is estimated for each feature j as

$$\widehat{\sigma}_j^2 = \frac{1}{M-1} \sum_{m=1}^{M} \left(\beta_j^{(m)} - \widehat{\mu}_j \right)^2.$$

The full empirical covariance matrix is estimated as

$$\widehat{\Sigma} = \frac{1}{M-1} \sum_{m=1}^{M} \left(\beta^{(m)} - \widehat{\mu} \right) \left(\beta^{(m)} - \widehat{\mu} \right)^\top.$$

Its inverse,

$$\widehat{P} = \widehat{\Sigma}^{-1},$$

defines the precision matrix used in the most general version of the model.

3.4 Models

The models in this thesis follow a common penalized estimation framework. With the exception of the unregularized baselines, all fitted models are special cases of the general MAP framework derived in Section 2.4: the coefficients are shrunk toward a target μ with structure P scaled globally by λ, where specific choices of (μ, P) are determined by the LLM elicitation procedure. LLM-derived prior information is therefore translated into structured shrinkage within linear and logistic regression rather than being used in a full posterior Bayesian analysis. The four model families correspond to specific (μ, P) instances of this framework: ridge / L_2-Logistic with $(0, I)$, Prior-Mean with $(\widehat{\mu}, I)$, Independent-Prior with $(\widehat{\mu}, \widehat{P})$ and diagonal $\widehat{P}$, and Correlated-Prior with $(\widehat{\mu}, \widehat{\Sigma}^{-1})$.

Only the *shape* of $\widehat{\Sigma}$ (the relative uncertainty across coefficients and the correlations between them) is taken directly from the elicitation. The overall scale of the prior is controlled by λ, which is tuned by cross-validation rather than fixed at unity. This empirical-Bayes treatment is motivated by the findings of Riegler et al. (2025) where they evaluated LLM-elicited Gaussian priors across two distinct domains and

found that while the LLMs can reliably identify the correct direction of regression coefficients, they had substantial difficulty calibrating the magnitude, exhibiting both overconfidence and underconfidence depending on the variable and provider. So trusting the relative uncertainty structure encoded in $\widehat{\Sigma}$ while tuning the absolute prior strength via λ is therefore a safeguard against miscalibration of the LLM's confidence, and lets the prior contribute its more reliable directional information without committing to its less reliable magnitude estimates.

Each prior-informed estimator is presented below in both its regression form (with squared-error loss) and its logistic form (with negative log-likelihood loss). The prior structure is identical across the two; only the data-fitting term differs.

3.4.1 Baseline Models

The unregularized baselines are ordinary least squares (regression) and unregularized logistic regression (classification). The regularized baselines are ridge regression and L_2-regularized logistic regression, both corresponding to the choice $\mu = 0$ and $P = I$ in the general framework. These baselines are described in detail in Sections 2.1–2.3.

3.4.2 Prior-Mean Estimator

The first prior-informed model replaces the zero shrinkage target with the LLM-estimated prior mean $\widehat{\mu}$, while keeping the precision matrix $P = I$. The corresponding Gaussian prior is

$$\beta \sim \mathcal{N}(\widehat{\mu},\, (1/\lambda)I)\,,$$

and combining this prior with the likelihood gives the posterior

$$p(\beta \mid y, X) \propto p(y \mid X, \beta) \cdot p(\beta).$$

For the regression case the likelihood is the Gaussian model of Section 2.1, and the negative log-posterior reduces, up to additive constants that do not depend on β, to the penalized objective

$$\widehat{\beta}_{\mathrm{PM}} = \arg\min_{\beta} \left[\|y - X\beta\|^2 + \frac{\lambda}{2}\|\beta - \widehat{\mu}\|^2 \right].$$

For the logistic case the likelihood is the Bernoulli model $L(\beta)$ of Section 2.2, and the same construction gives

$$\widehat{\beta}_{\mathrm{PM}} = \arg\min_{\beta} \left[-\log L(\beta) + \frac{\lambda}{2}\|\beta - \widehat{\mu}\|^2 \right].$$

Relative to ridge regression, the model no longer shrinks coefficients toward zero, but toward the average coefficient values suggested by the LLM across repeated prior draws.

3.4.3 Independent-Prior Estimator

The next model allows the shrinkage strength to vary by feature according to the LLM-estimated prior variance. The precision matrix is defined as

$$\widehat{P} = \operatorname{diag}\left(\frac{1}{\widehat{\sigma}_1^2 + \varepsilon}, \ldots, \frac{1}{\widehat{\sigma}_p^2 + \varepsilon}\right),$$

which is scaled globally by λ in the penalized objective. The corresponding Gaussian prior is

$$\beta \sim \mathcal{N}\left(\widehat{\mu},\, (1/\lambda)\widehat{P}^{-1}\right),$$

and combining this prior with the likelihood gives the posterior

$$p(\beta \mid y, X) \propto p(y \mid X, \beta) \cdot p(\beta).$$

Taking the negative log of the posterior under the Gaussian likelihood of Section 2.1 yields, up to additive constants that do not depend on β, the regression objective

$$\widehat{\beta}_{\mathrm{IP}} = \arg\min_{\beta} \left[\|y - X\beta\|^2 + \frac{\lambda}{2}(\beta - \widehat{\mu})^\top \widehat{P}(\beta - \widehat{\mu})\right],$$

and replacing the Gaussian likelihood with the Bernoulli likelihood $L(\beta)$ of Section 2.2 gives the logistic version

$$\widehat{\beta}_{\mathrm{IP}} = \arg\min_{\beta} \left[-\log L(\beta) + \frac{\lambda}{2}(\beta - \widehat{\mu})^\top \widehat{P}(\beta - \widehat{\mu})\right].$$

This model captures feature-specific uncertainty. Features with low estimated variance, meaning that the LLM was relatively consistent across samples, receive stronger shrinkage toward the prior mean. Features with high estimated variance are shrunk more weakly, allowing the observed data to have more influence.

3.4.4 Correlated-Prior Estimator

The most general regression model uses the full empirical covariance matrix $\widehat{\Sigma}$, with corresponding precision matrix $\widehat{P} = \widehat{\Sigma}^{-1}$. The corresponding Gaussian prior is

$$\beta \sim \mathcal{N}\left(\widehat{\mu},\, (1/\lambda)\widehat{\Sigma}\right),$$

and combining this prior with the likelihood gives the posterior

$$p(\beta \mid y, X) \propto p(y \mid X, \beta) \cdot p(\beta).$$

The same derivation as in Section 3.4.3 applies, with $\widehat{P}$ replaced by $\widehat{\Sigma}^{-1}$. For regression, the penalized objective becomes

$$\widehat{\beta}_{\mathrm{CP}} = \arg\min_{\beta} \left[\|y - X\beta\|^2 + \frac{\lambda}{2}(\beta - \widehat{\mu})^\top \widehat{\Sigma}^{-1}(\beta - \widehat{\mu})\right],$$

with the corresponding logistic version

$$\widehat{\beta}_{\mathrm{CP}} = \arg\min_{\beta} \left[-\log L(\beta) + \frac{\lambda}{2}(\beta - \widehat{\mu})^{\top} \widehat{\Sigma}^{-1} (\beta - \widehat{\mu}) \right].$$

Unlike the Independent-Prior model, this formulation allows the shrinkage to depend not only on the uncertainty of each individual coefficient, but also on the covariance structure between coefficients. In other words, it incorporates the possibility that the LLM systematically assigns related coefficients together across repeated prior draws.

3.4.5 Mixed Model

A potential weakness of the prior-informed models introduced above is that they shrink toward the LLM-estimated mean $\widehat{\mu}$ rather than toward zero, so individual coefficients can remain large whenever the LLM places mass away from the origin. This removes the implicit shrinkage-toward-zero that ridge inherits from its zero prior and may leave the prior-informed models more prone to overfitting, particularly if the LLM is miscalibrated. To address this, an extended estimator was tested that blends the *predictions* of the standard regularized baseline and the Prior-Mean model (not their coefficients):

$$\widehat{y}_{\mathrm{mix}} = (1 - \gamma)\widehat{y}_{\mathrm{baseline}} + \gamma \widehat{y}_{\mathrm{PM}},$$

where $\gamma \in [0, 1]$ is tuned by 5-fold CV using plain mean-CV (the same selection rule used for λ). At the endpoints, $\gamma = 0$ gives the regularized baseline and $\gamma = 1$ gives the Prior-Mean model, with intermediate values blending the LLM-informed shrinkage of Prior-Mean and the zero-target shrinkage of the baseline.

If a training set spanning multiple domains were available, γ could in principle be tuned on cross-domain folds (or on synthetic domain shifts of the training set), making the resulting mixture an explicit OOD-robust estimator. In the present setup only single-domain training data is available, so γ is tuned on in-domain folds, and the resulting mixture is not guaranteed to be optimal under OOD shift.

3.5 Training and Evaluation

3.5.1 Cross-Validation and λ Tuning

Hyperparameters were selected by 5-fold cross-validation (Stone, 1974) on the source-domain training pool defined in Section 3.2: the regularization strength λ for the regularized models, and for the Mixed model additionally the prediction-mixing weight γ. The per-fold validation score for each candidate was averaged across folds, and the value with the best mean was selected: lowest mean-CV MSE for regression, highest mean-CV AUC for classification.

After tuning, every model was refit on the full training pool at its chosen hyperparameter and evaluated on the held-out ID test and OOD sets. The same procedure

was applied across all five datasets and across every prior family, so any difference in reported performance between families reflects the prior specification rather than differences in tuning, preprocessing, or evaluation protocol.

The λ grid is `np.logspace(-4, 6, 51)`: 51 log-spaced values from 10^{-4} to 10^{6} in increments of $10^{0.2}$ (factor ≈ 1.58 per step). Log-spaced grids are standard for L2 regularization tuning because the effect of λ on the fitted coefficients is multiplicative; the same convention is used by glmnet and by sklearn's regularized estimators. The γ grid for the Mixed model is `np.linspace(0, 1, 101)`.

3.5.2 Provider Selection

The best provider per family per dataset was selected automatically. For each of the twelve LLM variants (three providers $\times$ two model modes $\times$ {non-reasoning, reasoning}) and all model types (Prior-Mean, Independent-Prior, Correlated-Prior, Mixed) were fitted using the standard training and CV pipeline. For each prior family, the variant whose family-model achieved the best mean-CV score on the training pool was selected as the winner for that family on that dataset. The selection metric is the same one used to pick λ within each cell: the strict-best mean of the per-fold validation scores. The held-out 20% ID test set is a clean out-of-fold evaluation set used only for reporting.

This multi-provider design avoids tying the results to a single LLM, which would make the findings difficult to interpret if that provider happened to be unusually strong or weak for a given domain. Crossing each provider with the reasoning / non-reasoning elicitation modes additionally isolates the effect of chain-of-thought on the resulting prior.

3.5.3 Evaluation Metrics

For regression tasks, mean squared error (MSE) was used. For classification tasks, the primary metric was AUC (area under the ROC curve), with sensitivity, specificity, and F1 score as secondary metrics. Brier score and log loss are also reported as scoring rules that capture probability calibration in addition to ranking performance.

3.5.4 Uncertainty Estimation

Uncertainty was quantified using bootstrap resampling (Efron & Tibshirani, 1993) of the evaluation data. For each evaluation set of size n, B=2,000 bootstrap replicates were generated by sampling with replacement (stratification by class for classification metrics). The performance metric was recomputed for each replicate, yielding an empirical distribution. Percentile-based 95% confidence intervals were obtained from the 2.5th and 97.5th percentiles of the bootstrap distribution.

The same resample indices were applied to every fitted model inside one bootstrap call, which makes the per-resample draws paired across models. This yields two

complementary CIs: a per-model marginal CI on each model's metric, and a paired-Δ-vs-baseline CI on the per-resample difference $\Delta = \text{model} - \text{baseline}$. Comparison-vs-baseline rows in Chapter 5 use the paired form because it accounts for evaluation-set covariance between the prior-enhanced model and its baseline; the marginal CIs are reported alongside for the model's own metric. The procedure estimates sensitivity of the reported performance to the composition of the evaluation sample, but does not account for variability in model training, hyperparameter selection, or prior elicitation.

3.6 Use of Generative AI

This thesis was prepared in accordance with NMBU's guidelines on the use of generative artificial intelligence. AI tools were used throughout the project as a learning aid, a writing aid, and a programming aid.

For writing support, Claude Chat (Anthropic) was used for proofreading and language polish (språkvask), and for suggestions on wording and flow during revision. I also used it as a thinking partner to refine how ideas were expressed and to talk through phrasing choices while editing.

For programming and document engineering, Claude Code (Anthropic) was used in the final weeks of the project to convert the thesis from .docx to LaTeX, to set up automated pipelines that regenerated all figures from a shared styling configuration so that styling changes could be applied across datasets efficiently, and to clean up and suggest optimizations to the analysis-pipeline code.

All AI-generated suggestions were reviewed and approved by the author before incorporation. The research design, methodology, experimental implementation, analysis, and interpretation of results are the author's own work, and the author bears full responsibility for the content, claims, and conclusions of this thesis.

Chapter 4

Datasets

This section documents the five datasets used in the study. For each dataset the description covers the source of the data, the features and response variable, the choice of OOD splitter, and the resulting source/OOD partitions. The distribution shift induced by the chosen split is also visualized: a 2D scatter on the two features with the largest source-vs-OOD mean shift (in pooled-σ units, Cohen's-d style (Cohen, 1988)) colored by domain, alongside a per-feature representativeness bar that ranks every feature by the magnitude of its source-vs-OOD mean shift. The two panels together let the reader see both the dimensionality of the shift (a single shifted feature versus a coordinated multi-feature shift) and what part of feature space the OOD set actually occupies relative to the source.

A sixth dataset, a small in vitro ruminant fermentation experiment with $n_{\text{train}} = 124$ (Wang et al., 2017) was also run through the same pipeline but is not reported here. Its sample size proved too small for the cross-validation procedure to identify a useful regularization strength: the per-fold validation curve was essentially flat across the entire λ grid, the strict-best mean rule landed at $\lambda \approx 0$ for every prior family, and the resulting deployed estimators either tied OLS or produced marginally worse OOD MSE. The experience is consistent with the general expectation that CV-tuned prior-informed estimators need a validation surface with enough signal to distinguish λ values from each other, and is noted here as a practical lower-bound condition on when this methodology can be expected to deploy a meaningful prior.

4.1 Boston Housing

The Boston Housing dataset (Harrison & Rubinfeld, 1978) contains 506 observations on residential housing in census tracts of the Boston Standard Metropolitan Statistical Area, originally collected to study how air quality and other neighborhood characteristics relate to property values. Each row represents one tract; the response is the median owner-occupied home value ($MEDV$), in thousands of 1970 dollars. The 12 retained features are tract-level: per-capita crime rate (CRIM), proportion of land zoned for large residential lots (ZN), proportion of non-retail business acres (INDUS),

Charles River dummy (CHAS), nitric-oxide concentration (NOX), average number of rooms per dwelling (RM), weighted distance to five Boston employment centers (DIS), accessibility-to-radial-highways index (RAD), full-value property tax rate (TAX), pupil-teacher ratio (PTRATIO), the historical Black-population transformation B, and percent lower-status population (LSTAT). The B feature has a well-documented racial-coding history that makes its inclusion ethically questionable; it is retained for parity with prior comparison studies on this dataset, but its sign and magnitude are not given a causal interpretation.

The OOD splitter is *AGE*, the proportion of pre-1940 owner-occupied units in each tract. A 70th-percentile cut on AGE produces a source pool of newer-housing-stock tracts and an OOD pool of older-housing-stock tracts. The cut location was chosen because the source and OOD AGE ranges have negligible overlap above this percentile. AGE itself is excluded from the feature matrix to prevent leakage of the splitter back into the model. The resulting partition has 284 ID training observations, 71 ID test observations, and 151 OOD observations.

Nine of twelve features show $d > 0.5$ and seven show $d > 0.8$; the AGE-percentile split induces a coordinated multi-feature shift, not a single-feature one.

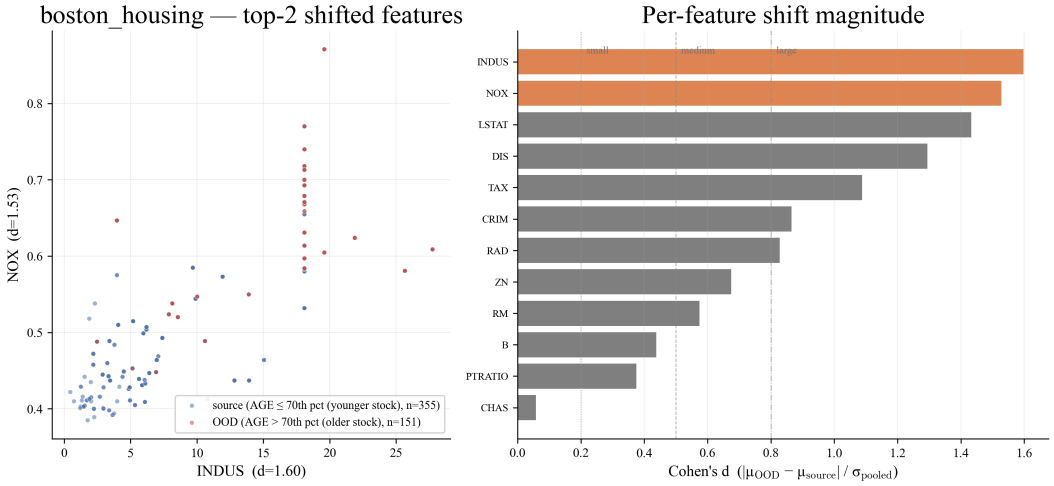


Figure 4.1: *Boston Housing: distribution shift between younger-housing-stock source (AGE $\leq$ 70th percentile) and older-housing-stock OOD (AGE $>$ 70th percentile). Left: scatter on the two top-shifted features, INDUS (proportion of non-retail business acres) and NOX (nitric-oxide concentration); the two domains separate cleanly with the OOD set concentrated at high INDUS and high NOX. Right: per-feature Cohen's d, all 12 features sorted descending; nine of twelve features show $d > 0.5$ and seven of twelve show $d > 0.8$, indicating that the AGE-percentile split induces a coordinated multi-feature shift rather than a single-feature one. Vertical guides mark $d = 0.2, 0.5, 0.8$.*

4.2 Cancer

The Cancer dataset (Rippner, 2017) is a US county-level cancer-mortality dataset hosted on Kaggle, compiled from cancer.gov county-level mortality summaries, ClinicalTrials.gov registrations, and US Census American Community Survey 5-year estimates covering roughly 2010–2016. Each row is one US county; the response is *TARGET_deathRate*, the annual age-adjusted cancer death rate per 100,000 residents. The full source file has 34 columns including identifiers, raw counts, and several proxy or duplicated coverage variables that are highly collinear with each other; the feature set is trimmed to 13 features that capture the core epidemiological structure: cancer incidence (*incidenceRate*), income (*medIncome*), poverty (*povertyPercent*), age (*MedianAge*), educational attainment (*PctHS25_Over*, *PctBachDeg25_Over*), employment (*PctEmployed16_Over*, *PctUnemployed16_Over*), insurance coverage (*PctPrivateCoverage*, *PctPublicCoverage*), racial composition (*PctBlack*, *PctOtherRace*), and household structure (*PctMarriedHouseholds*). The trim removes population-size proxies (*avgAnnCount*, *avgDeathsPerYear*, *popEst2015*), redundant coverage variables (*PctPrivateCoverageAlone*, *PctEmpPrivCoverage*, *PctPublicCoverageAlone*), and the granular education-by-young-age categories that overlap with the retained adult education variables. Because this is an ecological dataset, all coefficients should be read as population-level associations between county-aggregated features and county-level mortality, not as individual-level clinical effects.

The OOD splitter is US Census region: the source pool consists of counties in the South (AL, AR, DE, DC, FL, GA, KY, LA, MD, MS, NC, OK, SC, TN, TX, VA, WV) and the OOD pool consists of counties in the rest of the country. Region itself is not in the feature matrix. The resulting partition is 1,106 ID training counties, 277 ID test counties, and 1,664 OOD counties.

All thirteen features shift between South and non-South counties; *PctBlack* and *PctMarriedHouseholds* dominate, with several others above $d = 0.5$.

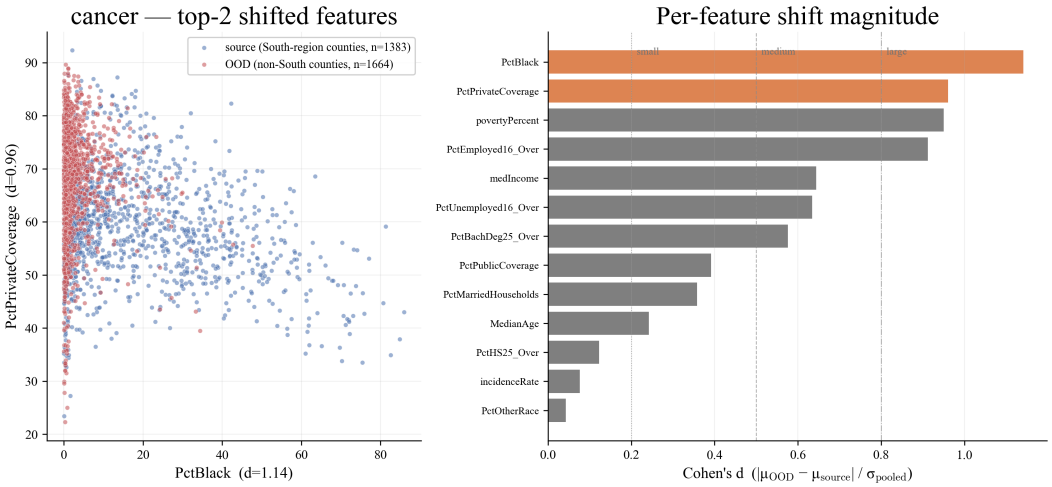


Figure 4.2: *Cancer: distribution shift between U.S. South-region counties (source) and non-South counties (OOD). Left: scatter on the two top-shifted features, with each county colored by region. Right: per-feature Cohen's d, all 13 features sorted descending; orange bars highlight the two features chosen for the scatter. Vertical guides mark* $d = 0.2, 0.5, 0.8$.

4.3 Heart Disease

The Heart Disease dataset is the multi-cohort UCI Heart Disease database (Detrano et al., 1989), which aggregates diagnostic records from four cardiology cohorts: the Cleveland Clinic Foundation, the Hungarian Institute of Cardiology, the University Hospital of Zurich and Basel (Switzerland), and the V.A. Medical Center, Long Beach. The response is binary (column *target* in the source file): presence of heart disease ($target > 0$) versus absence ($target = 0$). After an audit-time trim that removes the three columns with prohibitive missingness in the OOD cohorts (*ca*, *thal*, and *slope*, each $\geq 50\%$ missing in Hungary and Switzerland), the feature set comprises 5 numeric features (age, *trestbps*, *chol*, *thalach*, *oldpeak*) and 13 one-hot indicator levels expanded from the categorical features sex, *cp* (chest pain type), *fbs* (fasting blood sugar), *restecg* (resting electrocardiographic results), and *exang* (exercise-induced angina). The retained *fbs* feature is approximately 20% missing in the OOD cohorts; it is kept under the source-domain median imputation policy described in §3.2 because it is a clinically meaningful feature and the missingness fraction is below the 50% drop threshold.

The OOD splitter is study cohort, which also corresponds to a continental shift from North America to Europe: the ID source pool is Cleveland and the OOD pool is Hungary plus Switzerland. The V.A. Long Beach cohort is excluded entirely because it has severe missingness across most clinical features. Cohort identity is not in the feature matrix. The resulting partition is 242 ID training patients, 61 ID test patients, and 417 OOD patients.

The Cleveland $\rightarrow$ Hungary + Switzerland shift is driven primarily by chest-pain

composition and resting-ECG categories; numeric features (age, cholesterol) shift only modestly. The scatter panel plots the top two numeric features.

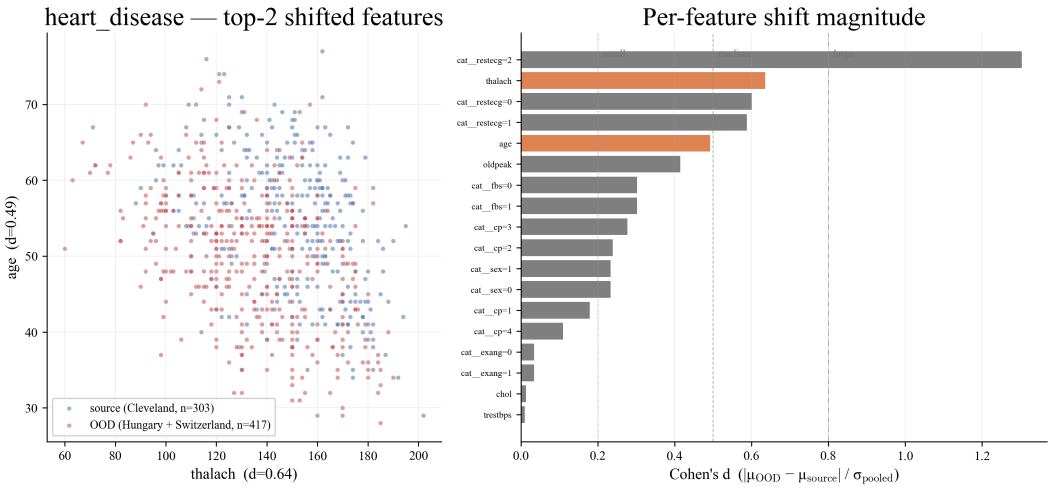


Figure 4.3: *Heart Disease: distribution shift between Cleveland source and Hungary + Switzerland OOD. Left: scatter on the two top-shifted non-binary features. Right: per-feature Cohen's d, all 18 features sorted descending; orange bars highlight the two scatter features. Vertical guides mark* $d = 0.2, 0.5, 0.8$.

4.4 Diabetes

The Diabetes dataset is the Pima Indians Diabetes Database (Smith et al., 1988), distributed by the National Institute of Diabetes and Digestive and Kidney Diseases. It contains 768 records of women of Pima Indian heritage aged 21 or older, drawn from a long-running diabetes screening study. The response is *Outcome*, a binary indicator of whether the patient developed diabetes within five years of the visit. The 7 retained features are clinical or metabolic measurements: number of pregnancies (Pregnancies), plasma glucose concentration in a 2-hour oral glucose tolerance test (Glucose), diastolic blood pressure (BloodPressure), triceps skinfold thickness (SkinThickness), 2-hour serum insulin (Insulin), body mass index (BMI), and a family-history score (DiabetesPedigreeFunction). The dataset has a well-documented data-quality issue: zeros in Glucose, BloodPressure, SkinThickness, Insulin, and BMI are not true zero measurements but encode missing values; these are treated as missing and imputed with the source-domain median during preprocessing.

The OOD splitter is age, with a median cut at 29 years: the ID source pool consists of patients aged 29 or younger and the OOD pool consists of patients older than 29. Age itself is excluded from the feature matrix to prevent leakage of the splitter back into the model. The resulting partition is 316 ID training patients, 80 ID test patients, and 372 OOD patients.

All seven clinical features shift with age; Glucose and BMI dominate but every

retained feature has $d \geq 0.2$.

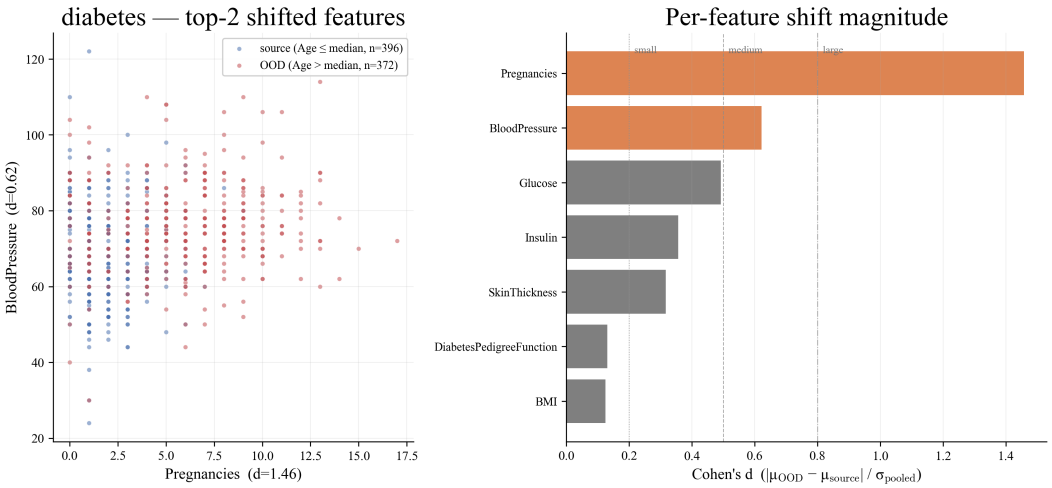


Figure 4.4: *Diabetes: distribution shift between younger-patient source (Age ≤ median) and older-patient OOD (Age > median). Left: scatter on the two top-shifted features, colored by age cohort. Right: per-feature Cohen's d, all 7 features sorted descending. Vertical guides mark* $d = 0.2, 0.5, 0.8$.

4.5 Adult

The Adult dataset is the UCI Adult Income / Census Income dataset (Kohavi, 1996), extracted from the 1994 US Census Bureau database. Each row is one individual; the response is a binary indicator of whether the individual earned more than $50,000 in 1994. The retained feature set contains six numeric variables (age; *fnlwgt*, the inverse-probability census sampling weight; *education-num*; *capital-gain*; *capital-loss*; *hours-per-week*) and one-hot indicator levels expanded from the categorical features marital-status and occupation; the occupation feature is grouped to the top-six source-domain categories with the remainder collapsed into a single *Other_ occupation* indicator. The *sex* feature is excluded from the feature matrix to avoid embedding a demographic protected attribute in the model's coefficient structure; the elicited Adult priors were generated under the same exclusion.

The OOD splitter is *workclass*: the ID source pool is private-sector workers and the OOD pool is non-private workers (federal, state, and local government employees, the self-employed, and unpaid family workers). Workclass itself is excluded from the feature matrix. The resulting partition is 27,124 ID training records, 6,782 ID test records, and 14,936 OOD records, making Adult the largest dataset in the study by a huge margin.

Capital-gain and hours-per-week dominate the shift; most marital-status and occupation indicators move modestly. Adult is the dataset with the highest feature count and the smallest per-feature shift on average.

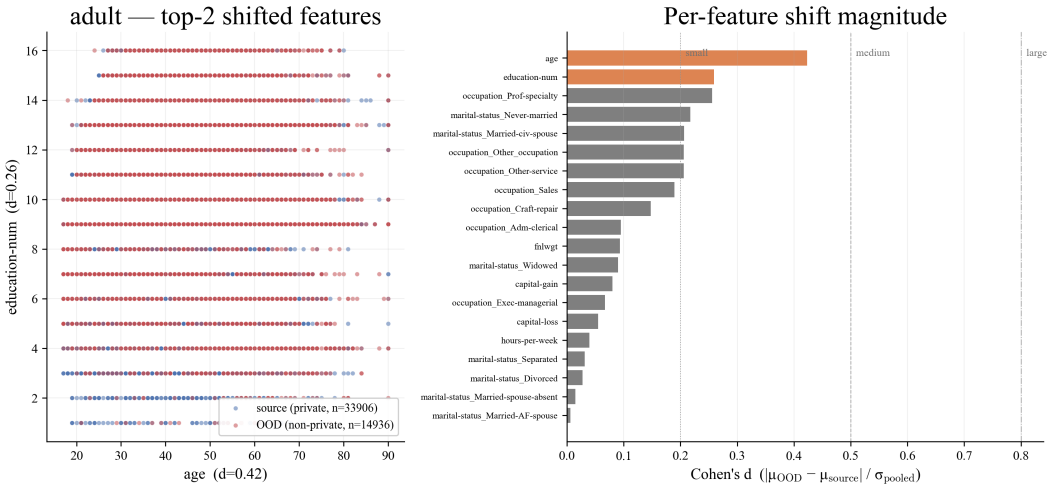


Figure 4.5: *Adult: distribution shift between source (private-sector workers) and OOD (non-private). Left: scatter on the two top-shifted non-binary features, with each point colored by domain. Right: per-feature Cohen's d, all 20 features sorted descending; orange bars highlight the two features chosen for the scatter. Vertical guides mark the conventional small / medium / large effect-size landmarks ($d = 0.2, 0.5, 0.8$).*

Chapter 5

Results

5.1 Overview

This chapter evaluates whether LLM-elicited priors improve OOD predictive performance, and examines the conditions under which improvements appear.

For each dataset and each prior family, the (LLM, λ) pair with the best 5-fold-CV mean is selected on the source-domain training data and reported as the deployment model. The deployment model's ID-test and OOD-test metrics are reported with marginal 95% bootstrap CIs, alongside the paired-Δ-vs-baseline 95% CI as the inferential test (significance = paired CI excludes zero). Baselines are Ridge (regression) and L2-Logistic (classification), the unregularized reference is also reported. Throughout the per-dataset tables and figures, "Unreg. ref." denotes the unregularized reference model (ordinary least squares for the two regression datasets, unregularized logistic regression for the three classification datasets) fitted on the source training pool. The "OOD-oracle" shown in the per-feature comparison plots is the same family of unregularized model fitted instead on the OOD partition.

Datasets are presented one at a time, each subsection has roughly the same structure: a per-family deployment table, a paired-Δ-vs-baseline OOD bootstrap CI plot, a λ tuning curve, a per-feature comparison of the LLM prior to the OOD-oracle coefficient, and a coefficient-shrinkage plot showing where each model family lands between the unregularized reference and the prior. Heart Disease and Diabetes additionally include an OOD calibration plot. Cross-cutting analyses follow: §5.3 examines per-LLM behavior across datasets, and §5.4 compiles the headline numbers.

5.2 Per-Dataset Results

5.2.1 Boston Housing

Table 5.1 shows the deployment-model comparison for the Boston Housing dataset. Independent-Prior with Opus 4.5 achieves the best OOD performance, with an OOD MSE of 39.12 compared with 46.45 for the regularized baseline. This corresponds

to a 15.8% reduction in OOD MSE relative to the baseline. Correlated-Prior and Prior-Mean also improve on the baseline, although by smaller margins. All three prior families have paired-Δ 95% CIs versus the baseline that exclude zero.

Table 5.1: *Boston Housing deployment-model comparison. Per-family CV-selected variant and λ, with ID and OOD test MSE, OOD 95% bootstrap CI, and paired-Δ-vs-baseline 95% CI.* ***Bold*** *marks paired-Δ CIs that exclude zero.*

Family	Selected variant	λ	ID MSE	OOD MSE	OOD 95% CI	Paired Δ 95% CI
Unreg. ref.	—	—	24.34	47.17	[31.37, 67.82]	—
Baseline	—	25.1	24.40	46.45	[30.53, 67.40]	—
Prior-Mean	Gemini Pro	158	22.84	42.58	[27.69, 61.80]	**[−7.07, −0.69]**
Independent	Opus 4.5	6.31	21.59	**39.12**	[25.28, 57.04]	**[−13.61, −1.51]**
Correlated	Gemini Pro (R)	39.8	23.37	42.94	[28.36, 62.48]	**[−6.65, −0.61]**
Mixed	Gemini Pro	$\gamma = 1.00$	22.84	42.58	[27.69, 61.80]	**[−7.07, −0.69]**

Figure 5.1 shows the paired-Δ-vs-baseline OOD MSE per family. All four prior-enhanced rows sit below the Ridge line, Independent furthest left with noticeably wider whiskers.

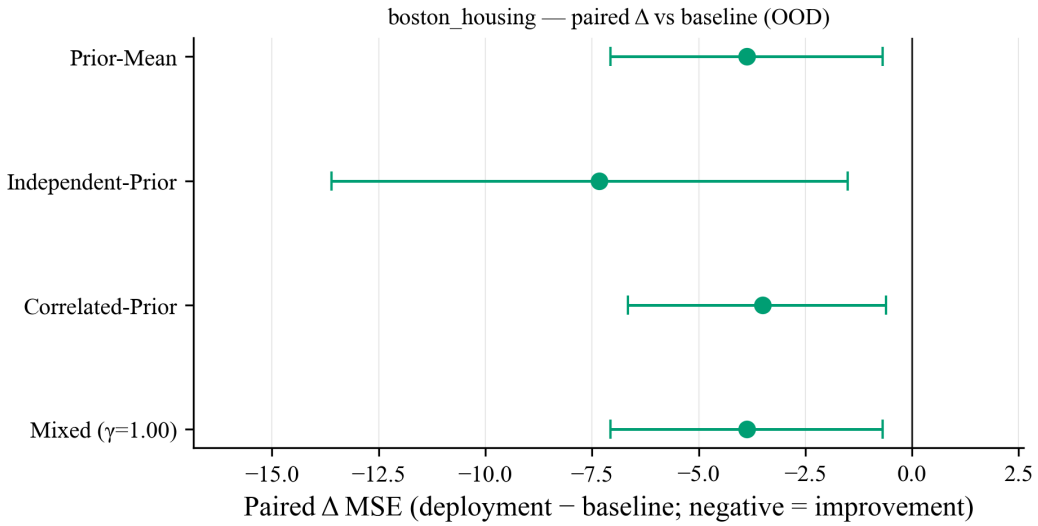


Figure 5.1: *Paired-Δ-vs-baseline on OOD test MSE, per family. Points are the deployment model's Δ relative to baseline; whiskers are the 95% percentile interval of the paired difference from 2,000 bootstrap resamples of the OOD test set. Vertical line at zero marks no improvement; whiskers fully on the negative side indicate significance.*

Figure 5.2 shows the validation MSE as a function of λ for each family. Each prior family has a shallow interior minimum at moderate λ; Ridge climbs sharply once λ exceeds the prior families' shrinkage scale.

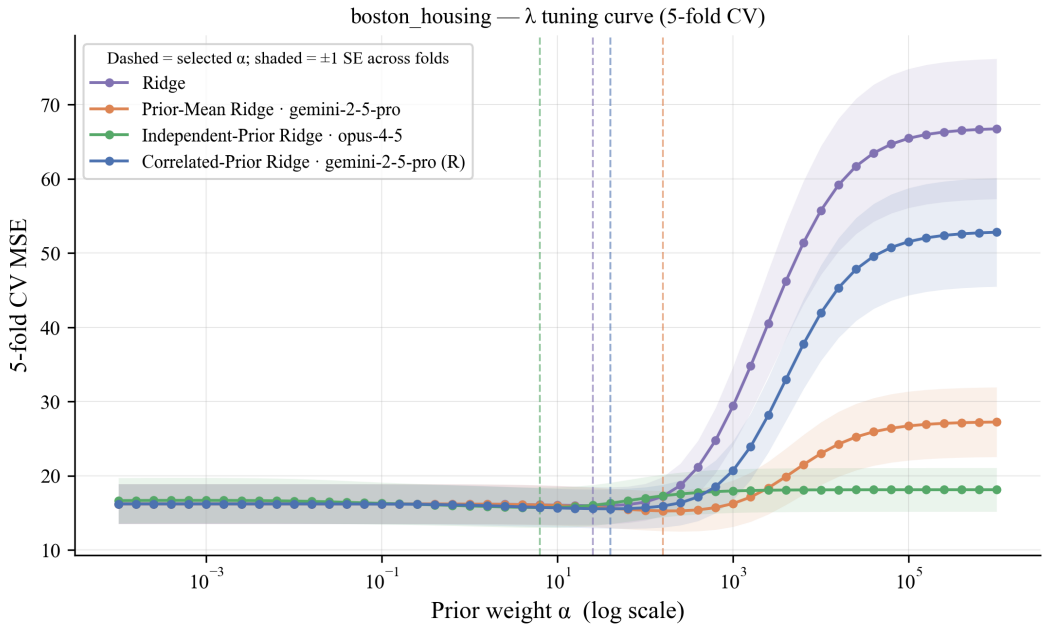


Figure 5.2: *Validation MSE as a function of prior weight λ (log scale), one curve per family at its deployment-winning LLM. Dashed lines mark each family's interior optimum.*

Figure 5.3 shows the per-feature LLM prior alongside the OOD-oracle coefficient. Tight prior intervals on most features with sign agreement on most. Avg rooms per dwelling and Distance to employment are the largest magnitude disagreements. But there are no sign disagreements.

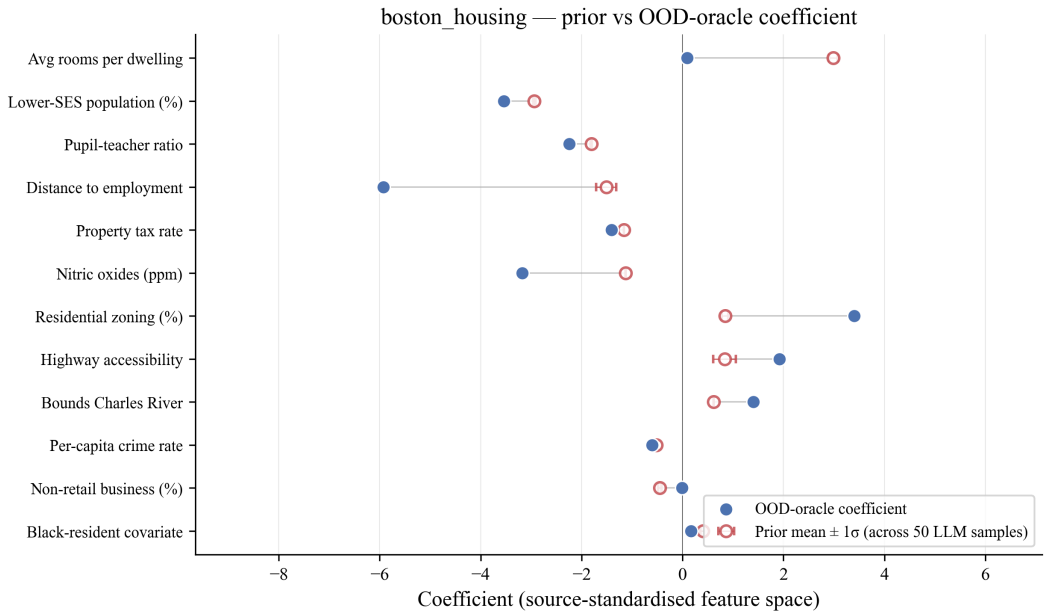


Figure 5.3: *Per-feature comparison of the LLM prior and the OOD-oracle coefficient. Blue circles: OOD-oracle coefficient. Open red circles with whiskers: prior mean $\pm 1\sigma$ across 50 LLM samples from the deployment-winning LLM. Gray segments connect the two markers per feature as a visual shrinkage axis. Both axes in source-standardized feature space.*

Figure 5.4 shows the coefficient shrinkage from the unregularized reference toward the LLM prior across families. Avg rooms per dwelling, Highway accessibility, and Property tax rate are pulled noticeably from the unregularized reference toward the prior. The Bounds Charles River prior is closer to the oracle coefficient than unreg. ref. is and pulls the prior models from negative to positive bounds.

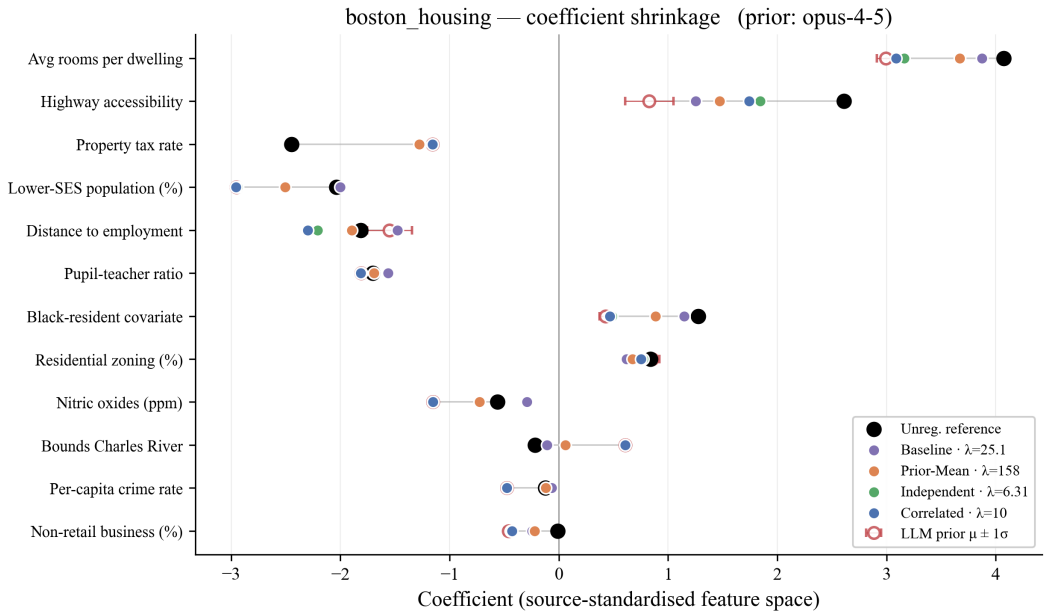


Figure 5.4: *Coefficient shrinkage on Boston Housing. Black filled circle: unregularized reference fit on the source split (anchor). Open red circle with whiskers: LLM prior mean $\pm 1\sigma$ (target). Colored filled circles: each model family's deployed coefficient at its tuned λ. Gray segment per feature shows the reference→prior shrinkage axis. All in source-standardized feature space.*

Boston Housing produces the largest single OOD lift in the study. Independent-Prior with Opus 4.5 reduces OOD MSE by 15.8% relative to the Ridge baseline, and all three prior families' paired-Δ CIs exclude zero. The improvement is consistent with the LLM picking up the coordinated multi-feature shift induced by the AGE-percentile split, where shrinkage toward zero is less informative than shrinkage toward a domain-anchored target.

5.2.2 Cancer

Table 5.2 shows the comparison of the deployed models for the Cancer dataset. Prior-Mean with Opus 4.5 is best OOD MSE 481.67 vs 501.22 baseline, which corresponds to a -3.90% improvement. Independent and Correlated also beat baseline. All three prior families' paired-Δ 95% CIs exclude zero and the mixed model has a very balanced blend, although the prior mean by itself is better than the mix.

Table 5.2: *Cancer deployment-model comparison. Per-family CV-selected variant and λ, with ID and OOD test MSE, OOD 95% bootstrap CI, and paired-Δ-vs-baseline 95% CI.* ***Bold*** *marks paired-Δ CIs that exclude zero.*

Family	Selected variant	λ	ID MSE	OOD MSE	OOD 95% CI	Paired Δ 95% CI
Unreg. ref.	—	—	372.90	524.17	[487.86, 562.44]	—
Baseline	—	100	374.35	501.22	[464.75, 538.48]	—
Prior-Mean	Opus 4.5	158	374.00	**481.67**	[446.17, 517.87]	**[−21.65, −17.58]**
Independent	GPT-5.1	158	372.05	487.27	[451.93, 523.89]	**[−16.89, −10.94]**
Correlated	Gemini Pro (R)	158	375.69	490.60	[455.36, 527.21]	**[−12.93, −8.20]**
Mixed	Opus 4.5	$\gamma = 0.68$	374.01	487.69	[451.82, 524.01]	**[−14.96, −12.18]**

Figure 5.5 shows the paired-Δ-vs-baseline OOD MSE per family. Prior-Mean lowest, then Independent/Mixed/Correlated. CIs for all except prior mean which is significantly better than the rest.

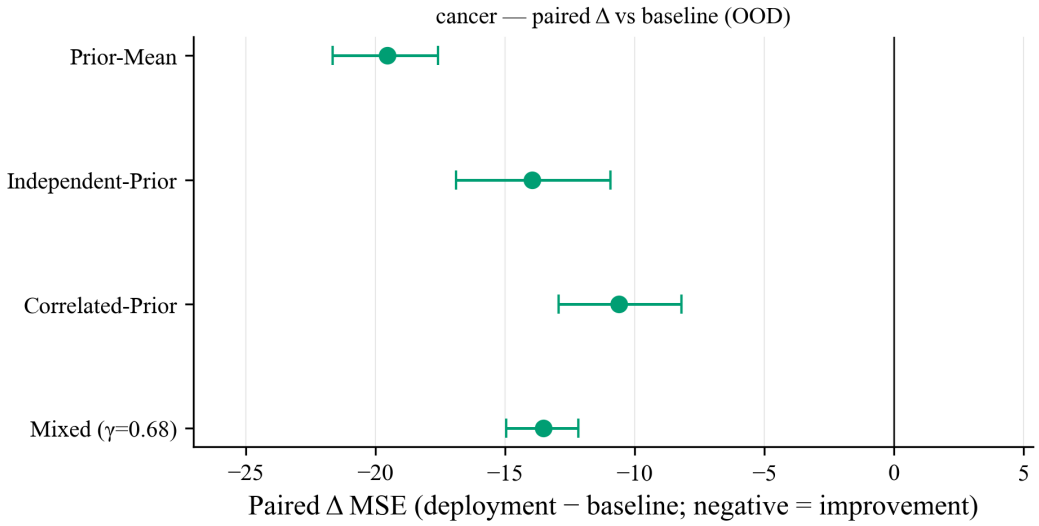


Figure 5.5: *Paired-Δ-vs-baseline on OOD test MSE, per family. Points are the deployment model's Δ relative to baseline; whiskers are the 95% percentile interval of the paired difference from 2,000 bootstrap resamples of the OOD test set. Vertical line at zero marks no improvement; whiskers fully on the negative side indicate significance.*

Figure 5.6 shows the validation MSE as a function of λ for each family. Prior families flat near 420 at low λ then dip into a broad interior minimum near $\lambda \approx 100$–200; baseline Ridge rises fastest as λ grows. Prior mean plateaus at lower MSE than the rest.

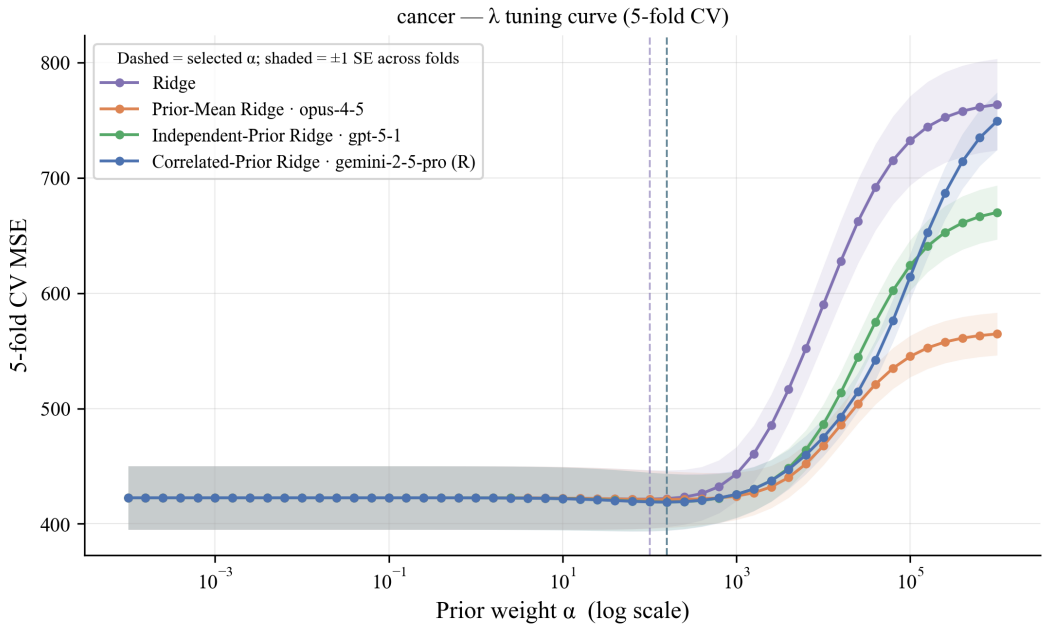


Figure 5.6: *Validation MSE as a function of prior weight λ (log scale), one curve per family at its deployment-winning LLM. Dashed lines mark each family's interior optimum.*

Figure 5.7 shows the per-feature LLM prior alongside the OOD-oracle coefficient. Prior tracks oracle on the *incidenceRate*; sign disagreements on *medIncome*, and *PctPublicCoverage*. The Cancer priors carry substantially more uncertainty than Boston Housing's.

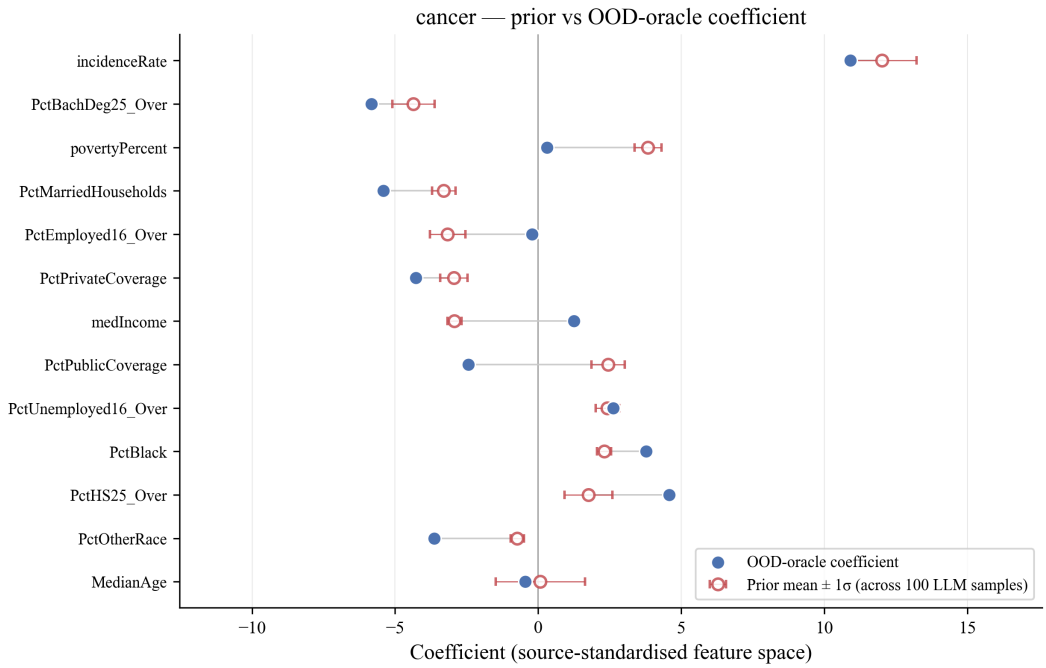


Figure 5.7: *Per-feature comparison of the LLM prior and the OOD-oracle coefficient. Blue circles: OOD-oracle coefficient. Open red circles with whiskers: prior mean $\pm 1\sigma$ across 50 LLM samples from the deployment-winning LLM. Gray segments connect the two markers per feature as a visual shrinkage axis. Both axes in source-standardized feature space.*

Figure 5.8 shows the coefficient shrinkage from the unregularized reference toward the LLM prior across families. *incidenceRate* stays close to the unregularized reference for every family; *PctPublicCoverage* is pulled noticeably toward the prior by prior mean but the two others stay behind which highlights the variance and covariance effects. The independent model is clearly being heavily affected by the priors with a low variance, for example on *medIncome* and PctBlack where it is the one being most heavily pulled toward the priors. On a couple of features (povertyPercent and PctMarriedHouseholds most visibly) the deployed coefficient lands on the zero side of OLS, even though both OLS and the LLM prior agree it should be far from zero. None of the penalties targets zero, so this is not a zero-shrinkage effect. The explanation is that the regression fits all 13 coefficients jointly. Several of the strongly correlated socio-economic features have priors that disagree with OLS on sign, so the regularizer pulls them hard toward those priors, and the remaining correlated features have to shift in the opposite direction to keep the joint fit consistent with the data. That compensating shift can carry a coefficient past the OLS-to-prior interval, sometimes landing nearer zero than either OLS or the prior alone would suggest.

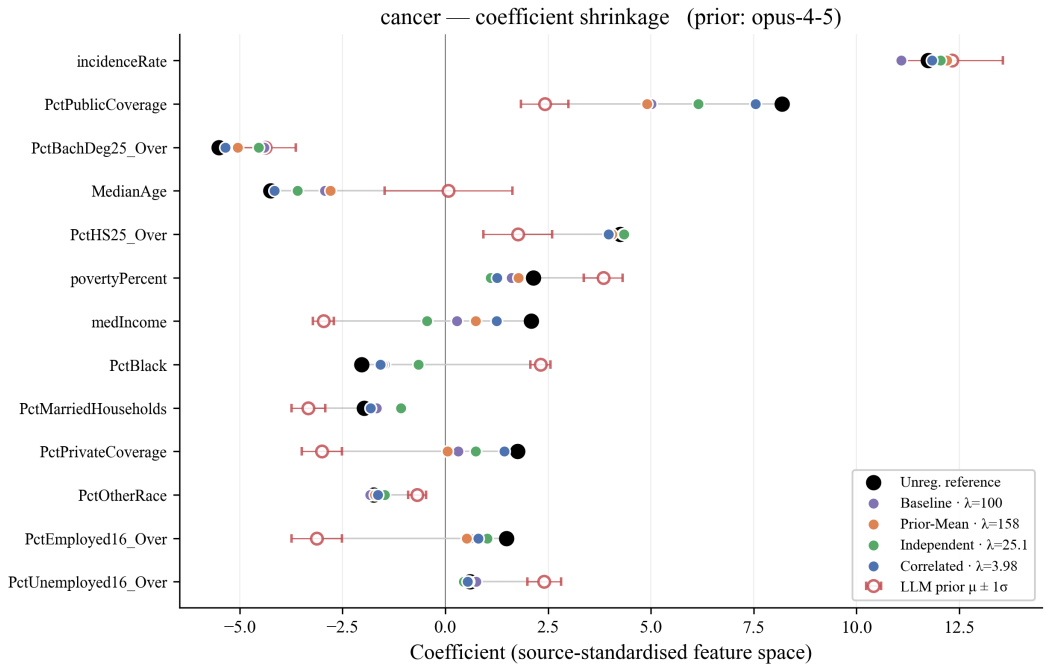


Figure 5.8: *Coefficient shrinkage on Cancer. Black filled circle: unregularized reference fit on the source split. Open red circle with whiskers: LLM prior mean $\pm 1\sigma$. Colored filled circles: each model family's deployed coefficient at its tuned λ. Gray segment per feature shows the reference→prior shrinkage axis. All in source-standardized feature space.*

Cancer is a very consistent positive in the study: every prior family's paired-Δ excludes zero on OOD MSE. Prior-Mean with Opus 4.5 wins at -3.9%, with Independent and Correlated trailing closely behind. The lift is smaller in absolute percentage than Boston Housing, but paired-Δ is more substantial for this dataset.

5.2.3 Heart Disease

Table 5.3 shows the deployment-model comparison for the Heart Disease dataset. By OOD AUC the three prior families and mixed sit above the baseline of 0.878; Correlated-Prior (Opus 4.5) has the highest single-run OOD AUC (0.8862) but all paired-Δ CIs cross zero, so discrimination is essentially unchanged. The threshold-0.5 metrics improve: the prior families lift F1 (Prior-Mean +0.048, Correlated +0.037, Independent +0.032, all paired-Δ CIs exclude zero) and sensitivity (0.6878 → 0.7692 for Prior-Mean), with a small drop in specificity (0.8878 → 0.8724).

Table 5.3: *Heart Disease deployment-model comparison. Per-family CV-selected variant and λ, with ID and OOD test AUC, OOD 95% bootstrap CI, and paired-Δ-vs-baseline 95% CI.* ***Bold*** *marks paired-Δ CIs that exclude zero.*

Family	Selected variant	λ	ID AUC	OOD AUC	OOD 95% CI	Paired Δ 95% CI
Unreg. ref.	—	—	0.8561	0.7509	[0.704, 0.799]	—
Baseline	—	25.1	0.9437	0.8779	[0.844, 0.911]	—
Prior-Mean	Opus 4.5	1,000	0.9535	0.8860	[0.852, 0.918]	[−0.008, +0.024]
Independent	Gemini Pro (R)	0.251	0.9524	0.8845	[0.851, 0.916]	[−0.009, +0.022]
Correlated	Opus 4.5	0.01	**0.9600**	**0.8862**	[0.853, 0.918]	[−0.008, +0.024]
Mixed	Opus 4.5	$\gamma = 1.00$	0.9535	0.8860	[0.852, 0.918]	[−0.008, +0.024]

Table 5.4: *Heart Disease OOD threshold-0.5 metrics. F1 with 95% bootstrap CI, plus Sens (sensitivity) and Spec (specificity) point estimates for the deployment models in Table 5.3, and the paired-Δ F1 vs baseline 95% CI.* ***Bold*** *marks paired-Δ CIs that exclude zero; bold metric values mark the per-column best (excluding the Unreg. ref. row).*

Family	OOD F1	OOD F1 CI	OOD Sens	OOD Spec	Paired Δ F1 95% CI
Unreg. ref.	0.6947	[0.640, 0.747]	0.5973	0.8622	—
Baseline	0.7696	[0.722, 0.815]	0.6878	**0.8878**	—
Prior-Mean	**0.8173**	[0.776, 0.856]	**0.7692**	0.8724	**[+0.018, +0.077]**
Independent	0.8020	[0.757, 0.844]	0.7421	0.8776	**[+0.005, +0.060]**
Correlated	0.8067	[0.765, 0.846]	0.7647	0.8520	**[+0.011, +0.065]**
Mixed	**0.8173**	[0.776, 0.856]	**0.7692**	0.8724	**[+0.018, +0.077]**

Figure 5.9 shows the paired-Δ-vs-baseline OOD AUC per family. Three prior families and Mixed cluster within 0.002 of each other slightly above the L2-Logistic line.

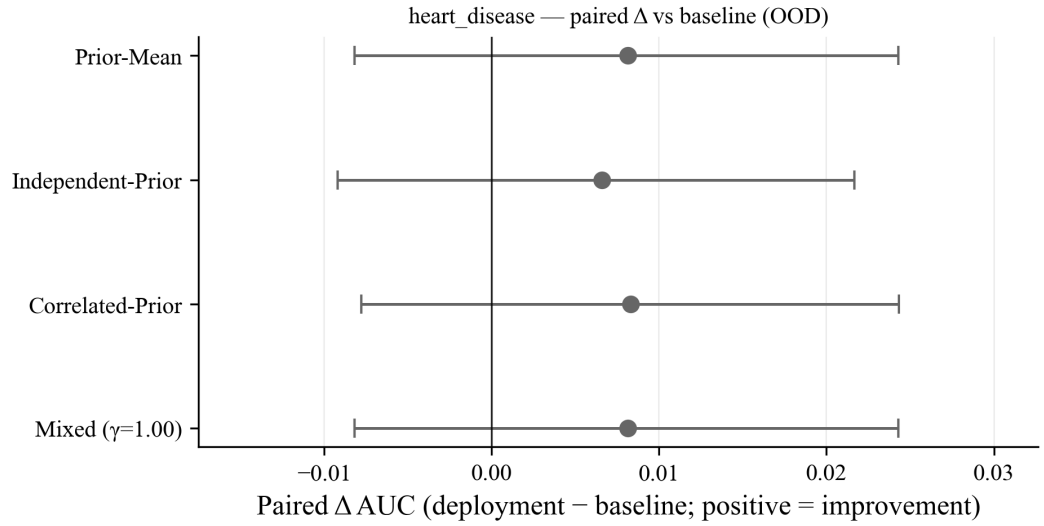


Figure 5.9: *Paired-Δ-vs-baseline on OOD test AUC, per family. Points are the deployment model's Δ relative to baseline; whiskers are the 95% percentile interval of the paired difference from 2,000 bootstrap resamples of the OOD test set. Vertical line at zero marks no improvement; whiskers fully on the positive side indicate significance.*

Figure 5.10 shows the validation AUC as a function of λ for each family. Correlated peaks at very small λ, Independent at intermediate λ, Prior-Mean at large λ; all three peaks land at very similar AUC scores.

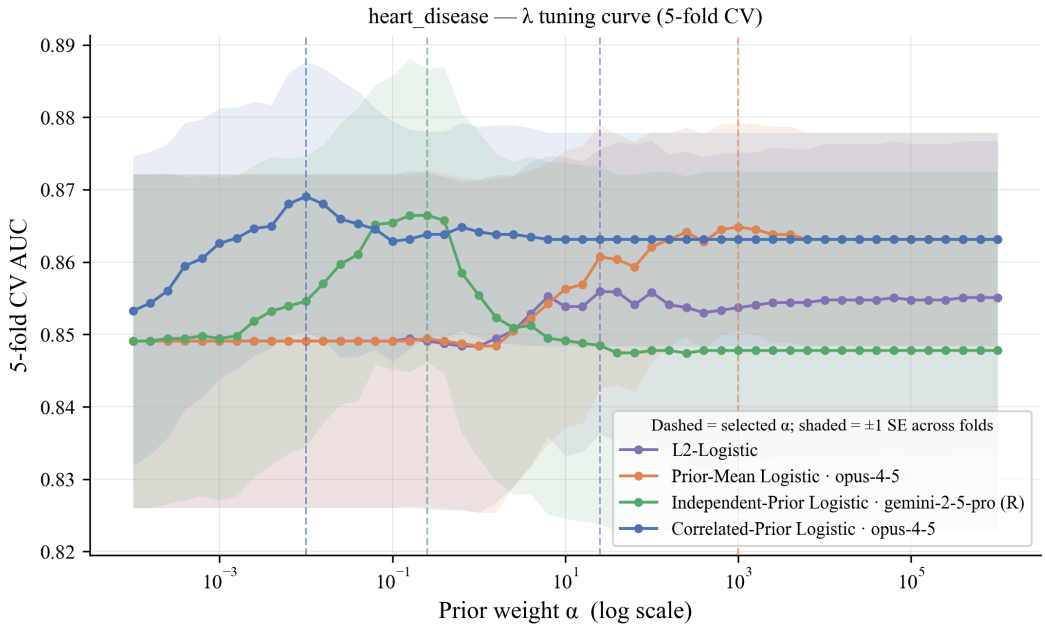


Figure 5.10: *Validation AUC as a function of prior weight λ (log scale), one curve per family at its deployment-winning LLM. Dashed lines mark each family's interior optimum.*

Figure 5.11 shows the per-feature LLM prior alongside the OOD-oracle coefficient.

Tight prior intervals across all 18 features and sign agreement on most; biggest disagreement is on *Chest pain: atypical angina* and *Chest pain: typical angina*, where the prior and oracle disagree on the sign. Other disagreements include Resting BP, Resting ECG: LV hypertrophy, and Resting ECG: normal. Some of the larger coefficients show very close agreement between prior and oracle, with low prior uncertainty.

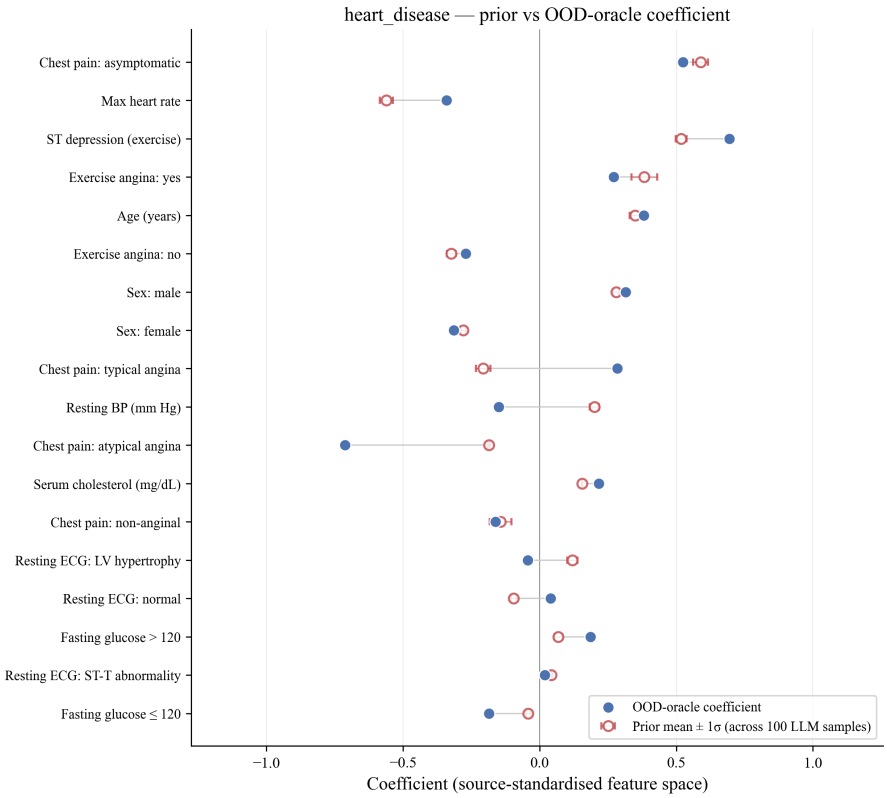


Figure 5.11: *Per-feature comparison of the LLM prior and the OOD-oracle coefficient. Blue circles: OOD-oracle coefficient. Open red circles with whiskers: prior mean $\pm 1\sigma$ across 50 LLM samples from the deployment-winning LLM. Gray segments connect the two markers per feature as a visual shrinkage axis. Both axes in source-standardized feature space.*

Figure 5.12 shows the coefficient shrinkage from the unregularized reference toward the LLM prior across families. Prior families pull strongly toward the priors for almost all features. We can also see the same multivariate shift as on Cancer for the correlated model on four of the features.

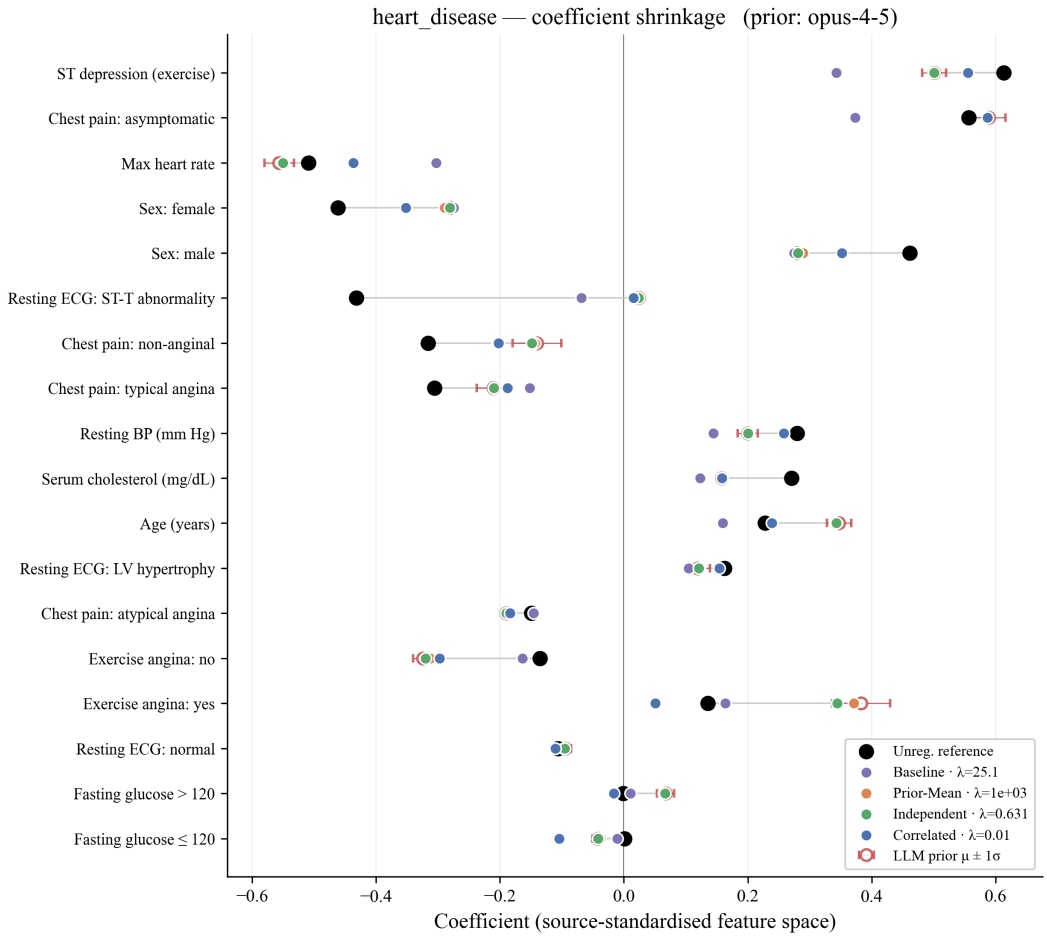


Figure 5.12: *Coefficient shrinkage on Heart Disease. Black filled circle: unregularized reference fit on the source split. Open red circle with whiskers: LLM prior mean* $\pm 1\sigma$*. Colored filled circles: each model family's deployed coefficient at its tuned* λ*. Gray segment per feature shows the reference→prior shrinkage axis. All in source-standardized feature space.*

Figure 5.13 shows OOD reliability curves for the baseline and the best-prior family. AUC barely moves and both curves track the diagonal across most of the probability range. Correlated-Prior sits slightly closer to the diagonal in the mid-to-high probability region and shows less over-prediction at higher predicted probabilities, producing lower Brier (0.138 vs 0.152) and LogLoss (0.433 vs 0.471).

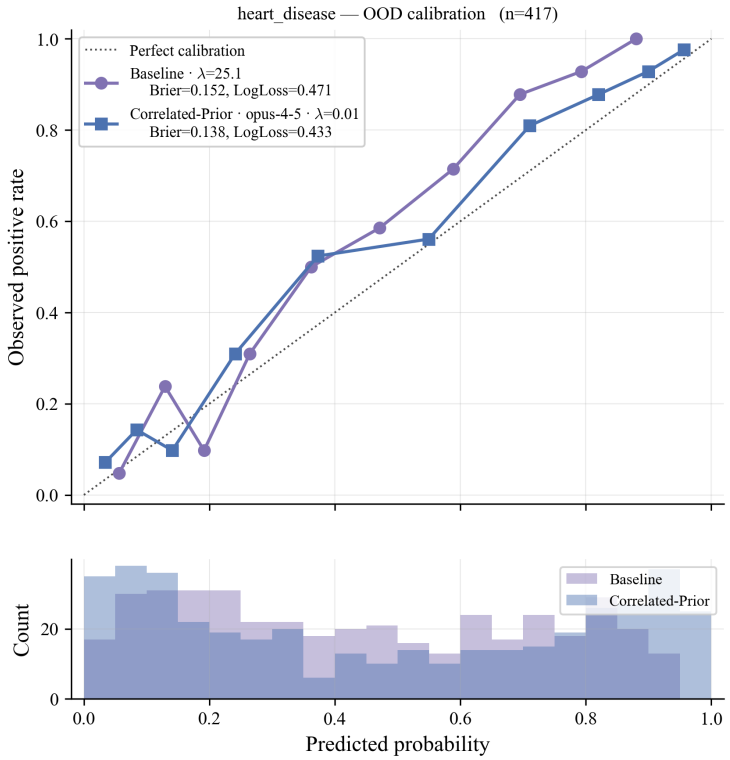


Figure 5.13: *OOD calibration on Heart Disease. Top: 10-quantile reliability curves for the deployed L2-Logistic baseline and the deployed best-prior family on AUC (Correlated-Prior with Opus 4.5). Dotted diagonal is perfect calibration. Bottom: histograms of predicted probabilities under each model. Brier and LogLoss in the legend.*

On Heart Disease the LLM priors leave OOD AUC essentially unchanged (paired-Δ CIs cross zero) but produce significant gains on threshold-0.5 metrics (Prior-Mean F1 $+0.048$ $[+0.018, +0.077]$, sensitivity $+0.08$) and on calibration (Brier 0.138 vs 0.152, LogLoss 0.433 vs 0.471). The lift is calibration-and-decision-quality rather than discrimination ranking.

5.2.4 Diabetes

Table 5.5 shows the deployment-model comparison for the Diabetes dataset. All four prior families pick Opus 4.5. OOD AUC sits at approx. 0.78 across the four prior families vs baseline 0.773; paired-Δ CIs cross zero. The same pattern appears here: on threshold metrics, the prior families gain performance: the prior families lift F1 (Prior-Mean $+0.089$, Independent $+0.089$, Correlated $+0.080$, Mixed $+0.085$, all paired-Δ CIs exclude zero) and sensitivity ($0.4185 \rightarrow 0.5435$ for Prior-Mean), with a drop in specificity ($0.9202 \rightarrow 0.8670$). The unregularized reference posts the highest F1 of any row (0.6667) despite the lowest AUC.

Table 5.5: *Diabetes deployment-model comparison. Per-family CV-selected variant and λ, with ID and OOD test AUC, OOD 95% bootstrap CI, and paired-Δ-vs-baseline 95% CI.* ***Bold*** *marks paired-Δ CIs that exclude zero.*

Family	Selected variant	λ	ID AUC	OOD AUC	OOD 95% CI	Paired Δ 95% CI
Unreg. ref.	—	—	0.7946	0.7676	[0.722, 0.813]	—
Baseline	—	25.1	0.7722	0.7730	[0.727, 0.817]	—
Prior-Mean	Opus 4.5	398	0.7983	0.7796	[0.736, 0.824]	[−0.006, +0.019]
Independent	Opus 4.5	0.158	0.7983	**0.7813**	[0.737, 0.826]	[−0.004, +0.020]
Correlated	Opus 4.5	0.631	**0.8011**	0.7805	[0.737, 0.825]	[−0.005, +0.020]
Mixed	Opus 4.5	$\gamma = 0.97$	0.7993	0.7797	[0.736, 0.824]	[−0.005, +0.019]

Table 5.6: *Diabetes OOD threshold-0.5 metrics. F1 with 95% bootstrap CI, plus Sens and Spec point estimates for the deployment models in Table 5.5, and the paired-Δ F1 vs baseline 95% CI.* ***Bold*** *marks paired-Δ CIs that exclude zero; bold metric values mark the per-column best (excluding the Unreg. ref. row).*

Family	OOD F1	OOD F1 CI	OOD Sens	OOD Spec	Paired Δ F1 95% CI
Unreg. ref.	0.6667	[0.609, 0.721]	0.5978	0.8085	—
Baseline	0.5580	[0.490, 0.625]	0.4185	**0.9202**	—
Prior-Mean	**0.6472**	[0.587, 0.703]	**0.5435**	0.8670	**[+0.039, +0.141]**
Independent	0.6471	[0.587, 0.704]	0.5380	0.8777	**[+0.039, +0.139]**
Correlated	0.6384	[0.576, 0.698]	0.5326	0.8670	**[+0.030, +0.130]**
Mixed	0.6429	[0.580, 0.700]	0.5380	0.8670	**[+0.036, +0.135]**

Figure 5.14 shows the paired-Δ-vs-baseline OOD AUC per family. All four prior families post small positive paired-Δ AUC point estimates of 0.007–0.008, with intervals spanning roughly $[-0.005, +0.020]$, all four cross zero.

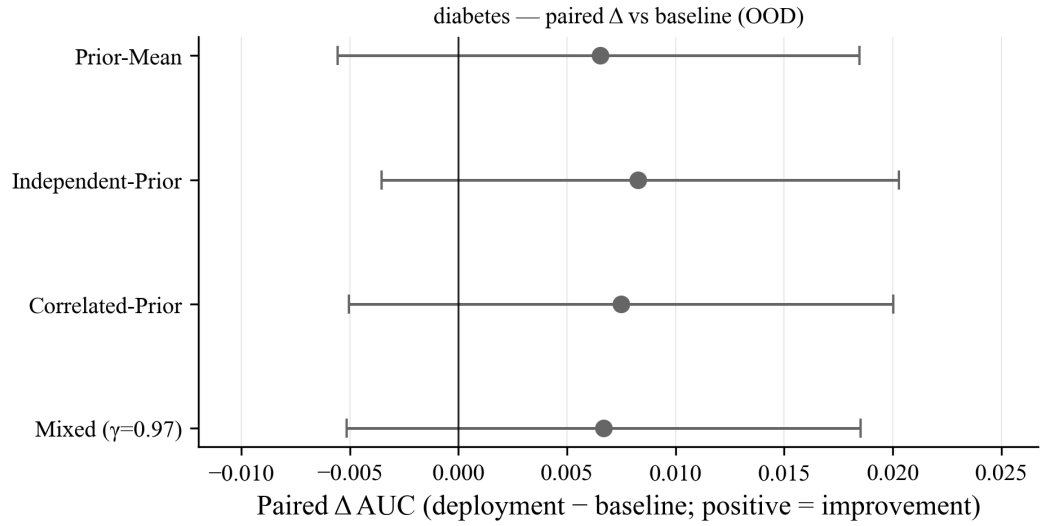


Figure 5.14: *Paired-Δ-vs-baseline on OOD test AUC, per family. Points are the deployment model's Δ relative to baseline; whiskers are the 95% percentile interval of the paired difference from 2,000 bootstrap resamples of the OOD test set. Vertical line at zero marks no improvement; whiskers fully on the positive side indicate significance.*

Figure 5.15 shows the validation AUC as a function of λ for each family. Independent and Correlated start higher than the rest, most likely because the strength of the covariance and precision matrices rises fastest to a plateau at small λ; Prior-Mean catches up at $\lambda = 398$; baseline plateaus below the prior families.

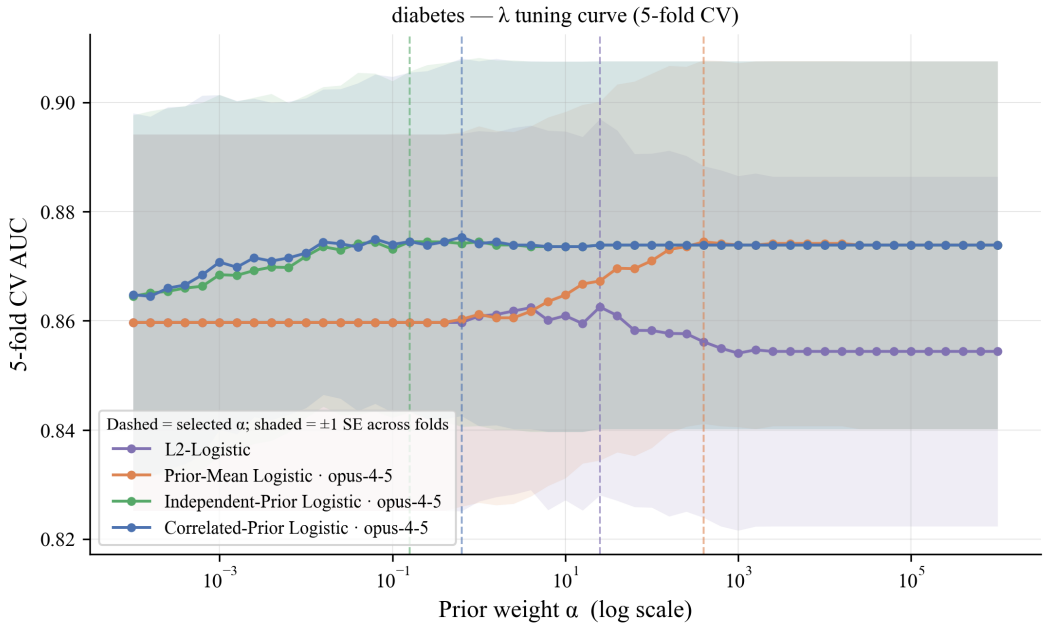


Figure 5.15: *Validation AUC as a function of prior weight λ (log scale), one curve per family at its deployment-winning LLM. Dashed lines mark each family's interior optimum.*

Figure 5.16 shows the per-feature LLM prior alongside the OOD-oracle coefficient. Close sign and magnitude agreement on Plasma glucose and Diabetes pedigree; prior under-shoots BMI; sign disagreement on Triceps skin fold and 2h serum insulin.

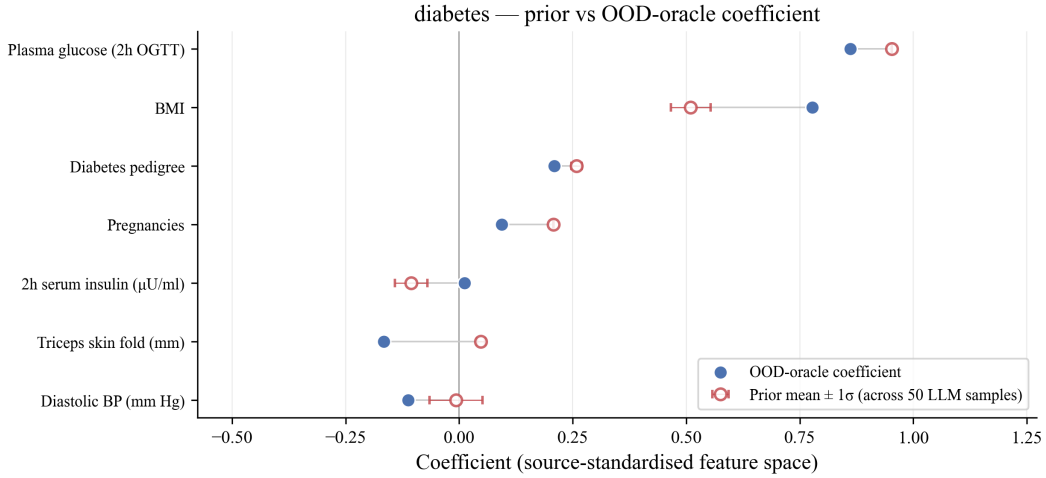


Figure 5.16: *Per-feature comparison of the LLM prior and the OOD-oracle coefficient. Blue circles: OOD-oracle coefficient. Open red circles with whiskers: prior mean $\pm 1\sigma$ across 50 LLM samples from the deployment-winning LLM. Gray segments connect the two markers per feature as a visual shrinkage axis. Both axes in source-standardized feature space.*

Figure 5.17 shows the coefficient shrinkage from the unregularized reference toward the LLM prior across families. Plasma glucose and BMI are pulled toward the prior; smaller features (Triceps skin fold, 2h serum insulin, Pregnancies) approach zero under shrinkage.

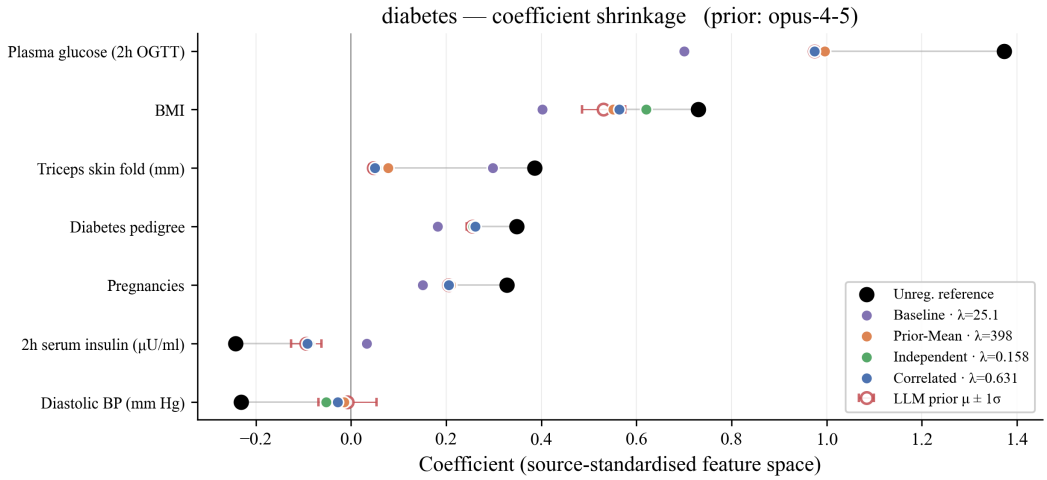


Figure 5.17: *Coefficient shrinkage on Diabetes. Black filled circle: unregularized reference fit on the source split. Open red circle with whiskers: LLM prior mean $\pm 1\sigma$. Colored filled circles: each model family's deployed coefficient at its tuned λ. Gray segment per feature shows the reference→prior shrinkage axis. All in source-standardized feature space.*

Figure 5.18 shows OOD reliability curves for the baseline and the best-prior family. AUC barely moves but the prior model posts lower Brier (0.207 vs 0.220) and LogLoss (0.605 vs 0.633). Both curves still over-predict positives in the low-mid probability range.

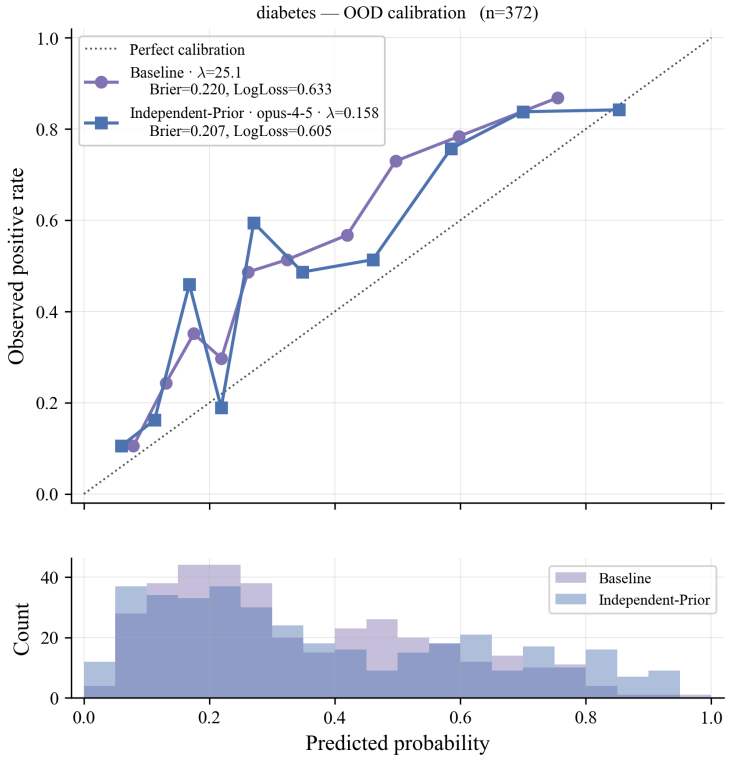


Figure 5.18: *OOD calibration on Diabetes. Top: 10-quantile reliability curves for the deployed L2-Logistic baseline and the deployed best-prior family on AUC (Independent-Prior with Opus 4.5). Dotted diagonal is perfect calibration. Bottom: histograms of predicted probabilities under each model. Brier and LogLoss in the legend.*

Diabetes shows the same calibration-not-discrimination pattern as Heart Disease: paired-Δ AUC CIs cross zero, but F1 lifts significantly (Prior-Mean +0.089 $[+0.039, +0.141]$) and the prior model posts lower Brier (0.207 vs 0.220) and LogLoss (0.605 vs 0.633). All four prior families pick Opus 4.5 as their winning variant, suggesting the directional signal is consistent across the ways of encoding it. Both calibration curves still over-predict positives in the low-mid probability range.

5.2.5 Adult

Table 5.7 shows the deployment-model comparison for the Adult dataset. All families are essentially indistinguishable at OOD AUC 0.877 $\pm$ 0.001, with F1/Sens/Spec all within 0.001. At $n_{\text{train}} \approx 27{,}000$ CV picks tiny λ on the prior families and they collapse to baseline. Because the tiny tuned λ combined with tight prior intervals leaves the prior-vs-OOD-coefficient and coefficient-shrinkage diagrams visually flat

and largely uninformative, those two plots are deferred to Appendix B; only the OOD bootstrap and λ tuning curve are shown in-line below.

Table 5.7: *Adult deployment-model comparison. Per-family CV-selected variant and λ, with ID and OOD test AUC, OOD 95% bootstrap CI, and paired-Δ-vs-baseline 95% CI.* ***Bold*** *marks paired-Δ CIs that exclude zero.*

Family	Selected variant	λ	ID AUC	OOD AUC	OOD 95% CI	Paired Δ 95% CI
Unreg. ref.	—	—	0.9114	0.8771	[0.871, 0.882]	—
Baseline	—	1	0.9114	0.8770	[0.871, 0.882]	—
Prior-Mean	GPT-5-mini	1	0.9114	0.8771	[0.871, 0.882]	[−0.0000, +0.0000]
Independent	Gemini Flash (R)	0.000631	0.9113	0.8768	[0.871, 0.882]	[−0.0004, −0.0001]
Correlated	Opus 4.5 (R)	0.001	0.9114	0.8769	[0.871, 0.882]	[−0.0002, −0.0000]
Mixed	GPT-5-mini (R)	$\gamma = 0.95$	0.9114	0.8771	[0.871, 0.882]	[−0.0000, +0.0000]

Table 5.8: *Adult OOD threshold-0.5 metrics. F1 with 95% bootstrap CI, plus Sens and Spec point estimates for the deployment models in Table 5.7, and the paired-Δ F1 vs baseline 95% CI.* ***Bold*** *marks paired-Δ CIs that exclude zero; bold metric values mark the per-column best (excluding the Unreg. ref. row).*

Family	OOD F1	OOD F1 CI	OOD Sens	OOD Spec	Paired Δ F1 95% CI
Unreg. ref.	0.6685	[0.658, 0.679]	0.6647	0.8691	—
Baseline	0.6685	[0.658, 0.679]	0.6647	0.8691	—
Prior-Mean	0.6687	[0.659, 0.680]	0.6649	0.8691	[+0.000, +0.000]
Independent	0.6695	[0.659, 0.680]	0.6651	0.8699	[−0.001, +0.002]
Correlated	0.6689	[0.659, 0.680]	0.6640	0.8702	[−0.001, +0.002]
Mixed	0.6687	[0.659, 0.680]	0.6649	0.8691	[+0.000, +0.000]

Figure 5.19 shows the paired-Δ-vs-baseline OOD AUC per family. Prior-Mean and Mixed sit on the baseline. Independent-Prior and Correlated-Prior are technically below baseline with paired-Δ CIs that exclude zero, but the magnitudes are negligible (paired-$\Delta \approx -0.0002$ and -0.0001, less than 0.03% of baseline AUC).

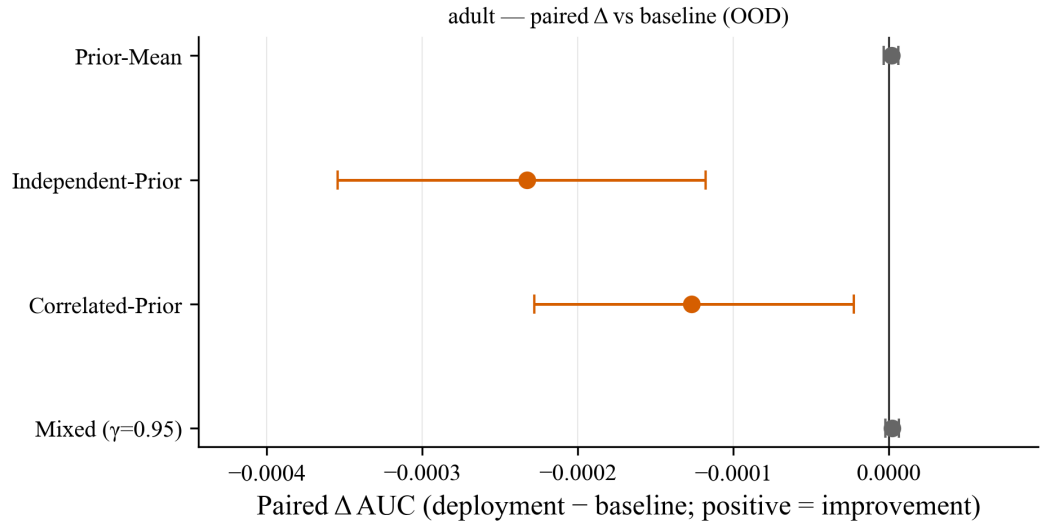


Figure 5.19: *Paired-Δ-vs-baseline on OOD test AUC, per family. Points are the deployment model's Δ relative to baseline; whiskers are the 95% percentile interval of the paired difference from 2,000 bootstrap resamples of the OOD test set. Vertical line at zero marks no improvement; whiskers fully on the positive side indicate significance.*

Figure 5.20 shows the validation AUC as a function of λ for each family. The four families flat at CV AUC ≈ 0.911 across the low-λ end then declining at different rates; Prior-Mean stays flat longest, Independent declines fastest.

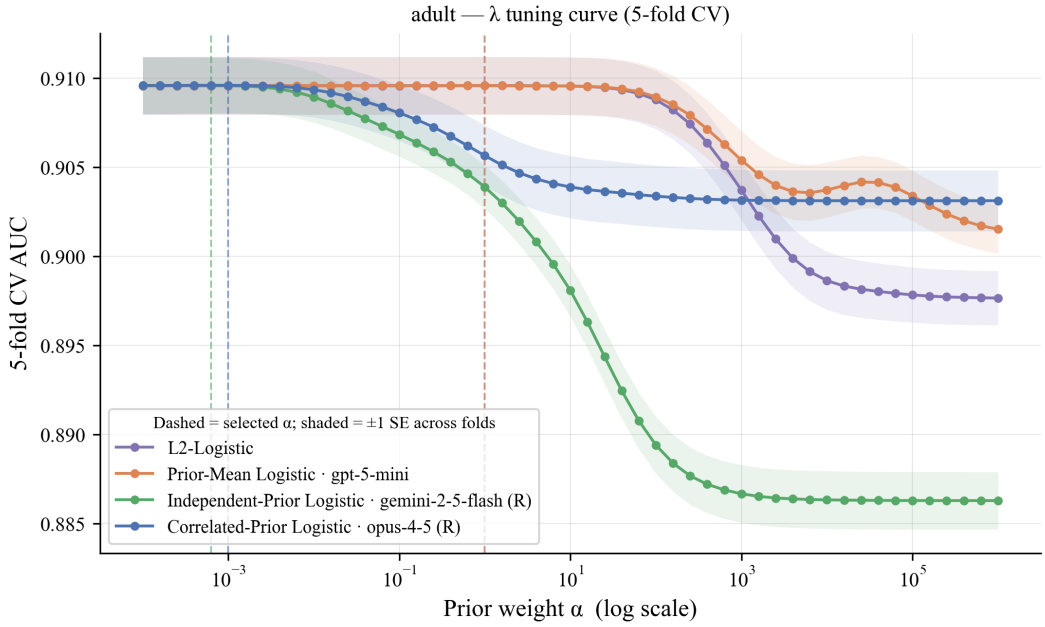


Figure 5.20: *Validation AUC as a function of prior weight λ (log scale), one curve per family at its deployment-winning LLM. Dashed lines mark each family's interior optimum.*

The LLM prior is essentially a zero vector (tight intervals, systematic under-

shooting of the OOD-oracle coefficients on every feature, though signs agree), and the tiny tuned λ leaves Independent and Correlated anchored near the unregularized reference rather than near the prior, because the data signal dominates. Per-feature prior–coefficient and coefficient-shrinkage diagrams are presented in Appendix B.

Adult is the data-dominated negative control of the study. With $n_{\text{train}} \approx 27{,}000$ the L2-Logistic baseline already saturates AUC on the OOD partition and CV picks tiny λ on every prior family, so the prior-informed estimators collapse to baseline predictions and all six rows agree to within ± 0.001 AUC. Priors here neither help nor hurt, which is the expected behavior of a well-tuned regularizer when the data signal is overwhelmingly stronger than the prior signal.

5.3 LLM Dynamics

Figure 5.21 summarises the reasoning-mode effect across datasets, showing the per-LLM mean R−NR lift averaged across the four datasets where prior families produce a lift (Adult is excluded since it is data-dominated and contributes no R−NR signal), normalised to a fraction of the per-dataset baseline so that MSE-unit and AUC-unit datasets contribute on the same scale. The figure shows a clean vendor-based split: all four non-Anthropic providers sit on the positive side of the axis, with reasoning mode producing modestly higher OOD lift on average (GPT-5.1 +1.10%, GPT-5-mini +0.66%, Gemini Pro +0.38%, Gemini Flash +0.29%); both Anthropic providers sit on the negative side (Opus 4.5 −1.35%, Haiku 4.5 −2.00%).

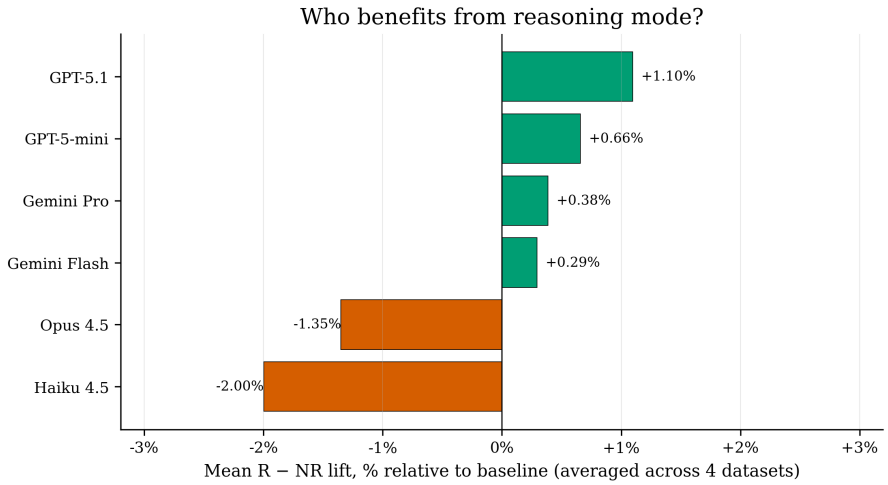


Figure 5.21: *Per-LLM mean reasoning-vs-non-reasoning lift, expressed as the per-dataset normalised lift in reasoning mode minus the same in non-reasoning mode, averaged across the four prior-helps datasets (Adult excluded; see body text). Positive bars indicate that reasoning mode produces a higher mean OOD lift than non-reasoning mode for that LLM; negative bars indicate the opposite.*

Figure 5.22 shows, for each dataset, the worst and best OOD lift achievable across

the 48 (variant, family) deployments at their CV-tuned λ, and the lift produced by the deployed pipeline. The plot motivates the cross-variant and cross-family selection step in the deployment pipeline: the deployed lift sits close to the per-dataset ceiling and well above the floor, which indicates the selection step contributes meaningfully to deployment stability.

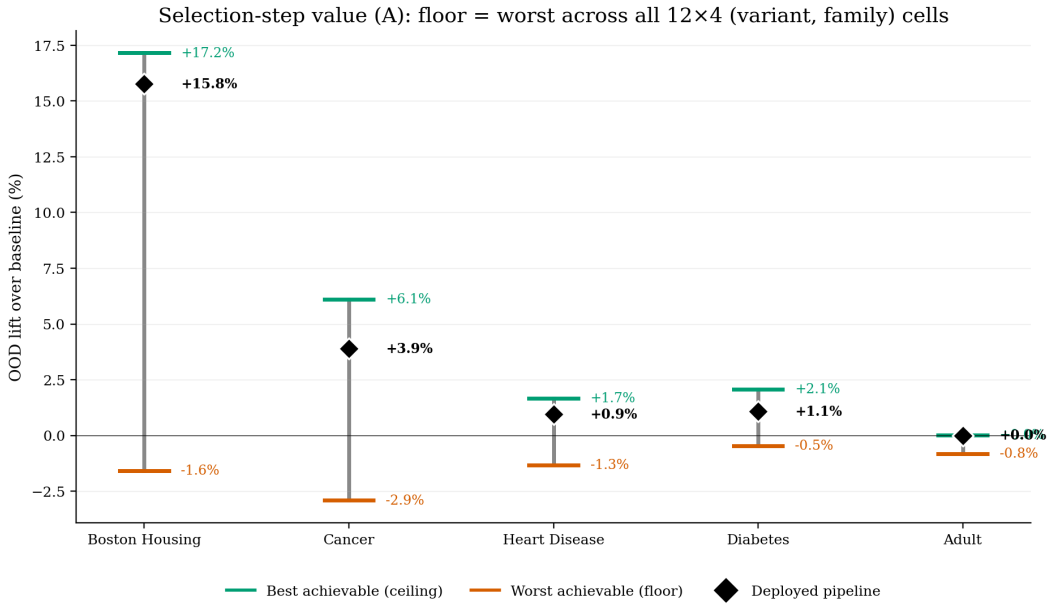


Figure 5.22: *Per-dataset OOD lift over baseline: worst achievable across all 48 (variant, family) deployments at their CV-tuned λ (lower tick of each whisker, vermillion), best achievable across the same pool (upper tick, bluish green), and the lift produced by the deployed pipeline (black diamond marker).*

A separate question is whether different LLMs even agree on what the prior should look like before any tuning: do they pick consistent signs and magnitudes for each feature, or do they disagree? Figures 5.23 and 5.24 answer that on the four datasets where the priors helped, plotting each variant's prior-mean coefficient (= mean across its 50 LLM samples) on shared axes, one panel per dataset. Heart Disease is restricted to its five numeric features so the agreement story isn't crowded out by 13 one-hot rows.

A handful of mid-magnitude features carry sign disagreements driven by one or two variants. On Cancer's *MedianAge* and *PctBlack*, the priors range from 0 on Opus 4.5 to 10 on several of the other variants. On Cancer's *PctHS25_Over*, Haiku and both Gemini Flash variants disagree with the rest on the sign. On Boston's *Distance to employment*, Opus 4.5 (R), GPT-5.1, and Haiku 4.5 fall on the opposite side of zero from the other nine variants. On Boston's *Highway accessibility*, both Gemini Flash variants cross zero but stay close to the rest and GPT-5.1 sits further on the negative scale. On Diabetes' *2h serum insulin*, Opus 4.5 alone is on the opposite side while the reasoning version is on zero. Three outliers fall off the panel edges and

are annotated at the boundary: Cancer *medIncome*, Boston *Residential zoning*, and Diabetes *Plasma glucose*.



Figure 5.23: *Cross-LLM coefficient agreement on the two regression datasets (Cancer and Boston Housing). Each panel shows the prior-mean coefficient for all twelve frontier LLM variants (three providers × two model sizes × {non-reasoning, reasoning}) per feature, in source-standardized feature space. Filled circles: non-reasoning. Open circles: reasoning. Gray vertical tick: across-variant mean. Off-panel points are annotated with their value at the boundary.*

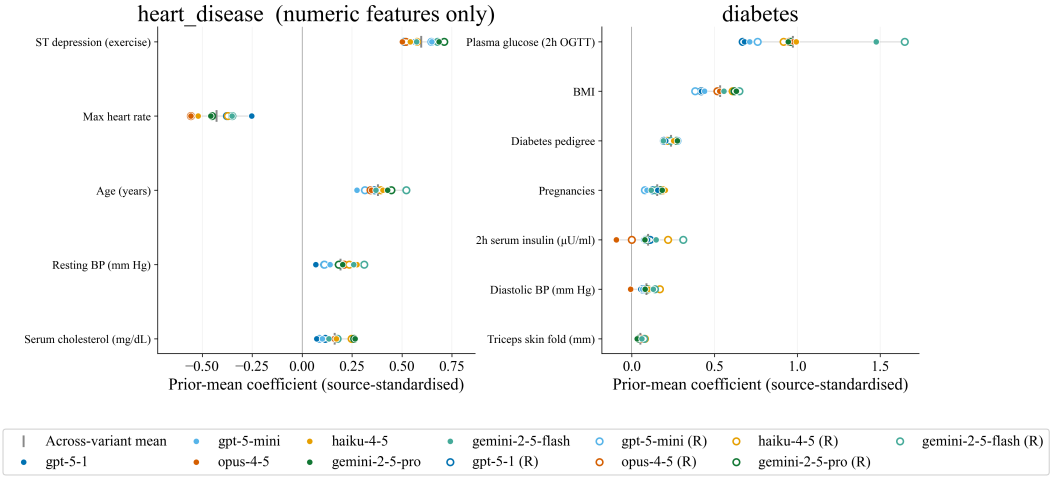


Figure 5.24: *Cross-LLM coefficient agreement on the two classification datasets (Heart Disease and Diabetes). Same conventions as Figure 5.23. Heart Disease shows numeric features only.*

All twelve variants agree on the sign of the dominant feature in each panel (*incidenceRate*, *Property tax rate*, *ST depression*, *Plasma glucose*); disagreement is mostly on weaker, ambiguous features. But the priors cluster around the same point which indicates that the LLMs base the priors on logical domain knowledge and not random guesses. A few single-cell outliers (most strikingly Gemini Flash on Boston *Property tax rate*) are clipped off-panel and annotated.

5.4 Summary

Across the five datasets in this study, four produced statistically meaningful prior-informed improvements over the regularized baseline on at least one OOD evaluation metric. Boston Housing and Cancer show direct improvements on the primary metric (OOD MSE): on Boston Housing the largest single lift is Independent-Prior with Opus 4.5 at -15.8% OOD MSE; on Cancer, Prior-Mean with Opus 4.5 at -3.9%, with every prior family's paired-Δ CI excluding zero. The classification AUC lifts on Heart Disease and Diabetes are not statistically significant; the significant gains for those datasets show up on F1 (Heart Disease Prior-Mean F1 $+0.048$ $[+0.018, +0.077]$; Diabetes Prior-Mean F1 $+0.089$ $[+0.039, +0.141]$) . Brier and LogLoss both lower for the prior model than for baseline. Adult is the one dataset where no prior family beat baseline: at $n_{\text{train}} \approx 27{,}000$ the L2-Logistic baseline already saturates AUC on the OOD, CV picks tiny λ on the prior families, and they collapse to baseline predictions within ± 0.001 AUC.

Table 5.9 compiles the headline numbers per dataset.

Two regression datasets (Boston Housing, Cancer) carry the headline OOD lifts. Heart Disease and Diabetes show non-significant AUC lifts but significant gains on F1, sensitivity, and probabilistic accuracy, with a small specificity tradeoff. Adult shows no useful lift.

Table 5.9: *Headline OOD performance per dataset: the best prior-informed model and its percentage improvement over baseline on the primary metric (OOD MSE for regression, OOD AUC for classification). Positive percentages indicate the prior-informed model beats baseline (lower MSE / higher AUC).*

Dataset	Metric	Winning model	% vs baseline
Boston Housing	MSE	Independent (Opus 4.5)	+15.8%
Cancer	MSE	Prior-Mean (Opus 4.5)	+3.9%
Heart Disease	AUC	Correlated (Opus 4.5)	+0.9%
Diabetes	AUC	Independent (Opus 4.5)	+1.1%
Adult	AUC	Prior-Mean (GPT-5-mini, tied)	$\approx 0\%$

Chapter 6

Discussion

6.1 Predictive Performance

Across the five datasets, the prior-informed models improved out-of-distribution performance in 4 out of 5 of them. The clearest gains occurred in the two regression tasks. On Boston Housing and Cancer, all LLM-informed prior families reduced OOD MSE relative to the Ridge baseline significantly. These results show that replacing generic shrinkage toward zero with shrinkage toward a domain-informed coefficient target can improve prediction when predicting on OOD data.

The classification datasets showed a more mixed pattern. On Heart Disease and Diabetes, the main AUC differences were small and generally within bootstrap uncertainty, meaning that the prior did not substantially change the rank ordering of observations. However, the threshold-based metrics, especially F1 and sensitivity, improved for some prior-informed models. This suggests that the priors may have helped move the fitted decision boundary or the scale of predicted probabilities, even when they did not strongly improve discrimination. In this sense, the classification results are weaker than the regression results on the primary metric, but they still suggest that LLM-informed priors can affect clinically relevant operating-point behavior.

Adult was the exception. None of the prior-informed families produced a meaningful improvement over the regularized baseline. This might be because of the characteristics of the dataset and split rather than the models. Adult has by far the largest source-domain training set and the OOD shift is comparatively mild, so the L2-Logistic baseline already has enough data to estimate stable coefficients. In that setting, there is little room for external prior information to improve the model. Cross-validation selected very weak prior contributions, and the deployed prior-informed models collapsed to baseline predictions. Adult therefore illustrates a safe-failure mode: when the LLM prior adds little beyond the data, the tuning procedure can effectively turn it off.

The winning prior structure also varied across datasets. Independent-Prior was best on two of the four datasets where the prior meaningfully helped, while Prior-

Mean and Correlated-Prior each won once. This matters because it shows that the additional uncertainty structure is not uniformly beneficial or useless. In some cases, the simple prior mean was enough; in others, feature-specific variance or full covariance structure gave the better deployment model. The three prior families should therefore be interpreted as complementary ways of using the same elicited information, rather than as a strict hierarchy where the most complex prior is always expected to win.

Taken together, the performance results support a qualified conclusion. LLM-elicited priors can improve OOD prediction, especially in moderate-sized regression problems with meaningful covariate shift, but the benefit is dataset-dependent. The method is most useful when the training data alone are not sufficient to identify stable coefficients, and least useful when the baseline already has enough data and the deployment shift is mild.

This pattern of dataset-dependent benefit is consistent with Rezaeimanesh and Ghassemi (2026), who report that LLM-derived priors help most when feature-target relationships reflect generalizable domain knowledge and fail when those relationships are dataset-specific or behavioral. The Adult result adds a complementary observation: even when the elicitation produces a sensible prior, cross-validation correctly drives the prior strength toward zero on a dataset where the OOD baseline already saturates, and the prior-informed estimators collapse to baseline rather than degrade it. That two quite different elicitation routes (token log-probabilities in LoID, sampled coefficient values here) produce broadly compatible benefit profiles suggests the gain reflects the LLM's underlying domain knowledge rather than a quirk of any one extraction mechanism.

6.2 The Role of Model Selection

A central result from this thesis is that the prior-informed models need a selection pipeline around them. The LLM priors are not automatically safe just because they are domain-informed. Across the full set of provider and prior-family combinations, some candidates improve OOD performance, while others perform substantially worse than the regularized baseline. This means that a study using only one provider, one prompt setup, and one prior model could easily draw a misleading conclusion from the same general method. Depending on which LLM and prior family happened to be chosen, the method could appear clearly beneficial, neutral, or harmful.

Figure 5.22 makes this point directly. The worst candidate model is below baseline on every dataset, showing that LLM-informed priors can damage prediction when the elicited prior or the chosen prior structure is poorly matched to the task. The prior has to be treated as a candidate source of information, not as an authority. Cross-validation over prior strength, provider variant, and prior family is what prevents the weakest candidates from reaching deployment.

At the same time, the deployed model is not usually the absolute best model on the OOD test set. This is expected, because all tuning and selection are performed using

only source-domain data. The OOD set is held out, so by design the hyper-parameter tuning and selection of provider is done to maximise performance on the ID data; this is a central struggle in all OOD prediction tasks. But still, when looking at the performance range of the models, the selection pipeline used in this thesis consistently chooses providers that are at the top end of the spectrum. For example, on the Boston Housing dataset the range runs from -1.6% to 17.2% and the selected model ends up with 15.8%, which is very close to the absolute best performance.

This tradeoff works well on the datasets tested in this thesis: the pipeline gives up the chance of always selecting the absolute OOD-best model in exchange for more stable performance. Selecting the best OOD model directly would only be possible by chance, if we open the possibility of chance to play a part we risk getting a model that hurts the performance. The selection process across providers and prior families is therefore a crucial step toward making the method deployable in real-world settings.

6.3 What the LLMs Actually Know

The cross-LLM agreement panels (Figures 5.23 and 5.24) provide a direct view of how much the elicited priors agree when the same problem is shown to twelve different frontier-LLM variants. The dominant feature in each panel (*incidenceRate* on Cancer, *property tax rate* on Boston Housing, ST depression on Heart Disease, and plasma glucose on Diabetes) receives the same sign from every variant, and the across-variant means cluster around a shared central value. Sign disagreement, where it appears, is concentrated on lower-magnitude coefficients: mid-magnitude features on Cancer and weak features near zero on Boston Housing and Diabetes. This indicates that the LLMs contain useful domain information for the features that carry the strongest signal, while being noisier where the association is weaker or less obvious. Twelve independently prompted variants agreeing on the sign of the dominant features for cancer mortality or Boston housing prices is unlikely to be random.

But as discussed in the last section, the performance can differ quite drastically even with the general agreement on the priors. This suggests that smaller magnitude differences, occasional sign disagreements, covariance or variance structure, and individual outlying samples can still matter once the priors are implemented into the fitted model. The agreement plots therefore support the idea that the LLMs are drawing on shared domain structure, while the performance spread shows why this shared structure still needs to be filtered through validation and model selection.

A small number of extreme single-cell outliers do remain even after the $M = 50$ averaging procedure. The most striking case is Gemini Flash on Boston *Property tax rate*: 49 of its 50 elicited coefficients sit close to the cluster (raw median ≈ -0.01, standardized ≈ -1.4), but a single draw returned roughly -50 in raw units, the equivalent of claiming that a one-unit increase in the property tax rate lowers the median home value by \$50,000, and that lone draw drags the $M = 50$ mean to roughly -148 in standardized space. The estimator design partially compensates for

this through CV-tuned λ, which lets the data down-weight an unreliable prior at deployment time. Future work would implement filtering during the prompt elicitation process so that priors like this do not come through. In this case it caused no harm to the final performance because the selection pipeline worked as it was designed to and the Gemini Flash family was not deployed on Boston Housing.

The reasoning-vs-non-reasoning comparison (Figure 5.21) shows a clean vendor-based pattern. All four non-Anthropic providers (OpenAI's GPT-5.1 and GPT-5-mini, Google's Gemini Pro and Gemini Flash) had higher average OOD lift in reasoning mode than in non-reasoning mode. Both Anthropic providers (Opus 4.5 and Haiku 4.5) went the other way.

There is one possibility for why this happens, although it is not confirmed: Anthropic's extended-thinking mode is incompatible with forced tool-use (§3.3.1), so the JSON schema for the elicited coefficients is enforced by prompt only on the Anthropic R variants, whereas every other R variant enforces the schema in the API call. Although it was not found specifically what went wrong here, this is the biggest contrast between the Anthropic models and the others.

The relative performance drop from not using reasoning are not very substantial, so a budget-constrained user could realistically prompt the non-reasoning variants for pure JSON output and get comparable performance. Also, given that several of the best-performing models are non-reasoning and the best performer overall is Opus 4.5 (NR), one could safely try to get a higher quantity of providers and samples rather than spending money on tokens for reasoning.

6.4 Limitations and Future Work

A few limitations are worth discussing.

The elicitation procedure used a single prompt variant per dataset (the `feature_context` variant). During development, alternative prompt forms were briefly explored at varying lengths and information content, but the project deliberately did not pursue prompt engineering further. The reasoning was that any improvement obtained by tuning prompts to the specific datasets and OOD splits used here would not be guaranteed to transfer to other domains, and would risk producing a methodology that performs well on the chosen test bed but does not generalise. The cost of that decision is that the results do not characterize sensitivity to prompt design.

The empirical variance of the M sampled coefficient vectors is treated throughout as a proxy for the LLM's underlying uncertainty about the coefficient, although this was not directly tested. The estimator was designed under the hypothesis that the LLM's underlying distribution over coefficient values is a more honest signal than its semantic self-reports of uncertainty would be, but whether the temperature-induced sampling distribution actually tracks the LLM's belief structure in a perfect way is not confirmed. The token-log-probability approach of LoID, mentioned in §1.2, is one alternative that grounds the per-coefficient variance more directly in the model's

internal probabilities and would be a natural cross-check for future work.

Another possible limitation is the choice of datasets. The methodology was designed under the assumption that the LLM has to actually reason or know about the data domain rather than recall it directly, but several of the datasets used here are well known and likely overlap heavily with public LLM training data (Sainz et al., 2023). The cross-LLM agreement plots show several features on which the elicited priors had very low variance, which can be read either as the LLMs being genuinely confident about the effect or as the LLMs being able to deduce from the prompt which dataset they are being asked about and recalling its known coefficient structure. The two cases cannot be distinguished with the available data. An attempt to address this was made by including a dataset supplied directly by an NMBU researcher that had not been published publicly, but after multiple rounds of testing it was decided that the dataset was too small for an OOD split and prediction task (λ collapsed to zero on every model because of the low n), and it was dropped from the study. Rezaeimanesh and Ghassemi (2026) successfully tested this in their paper and reported a positive result, which can indicate that the LLMs actually do produce original priors with a grounding in actual domain knowledge.

The OOD-oracle coefficient vectors used in the per-feature comparison plots are unregularized model fits on the OOD partition (OLS for regression, unregularized logistic for classification), not the true underlying generative coefficients of the OOD distribution. They are useful as a visual reference for comparison against the LLM prior, but they should not be read as ground truth. Sampling variability in the OOD set means the oracle is itself an estimate, and a prior that disagrees with it on a feature may not actually be disagreeing with the true effect. A natural alternative would be synthetic data, where the coefficients are known exactly and the LLM prior could be evaluated against true ground truth rather than an oracle estimate. This was deliberately not done in this thesis because the whole concept relies on the LLM bringing in domain knowledge on real data, and synthetic data has no domain for the LLM to draw on.

The pipeline works well in practice but is slow. Each (dataset, provider, mode) cell requires $M = 50$ valid LLM responses, and reasoning-mode responses are substantially more expensive than non-reasoning responses in both wall-clock time and API spend. The deployment model is also chosen by fitting every (variant, family) combination and scoring each on cross-validation, which scales poorly with the number of variants and prior families. Two natural directions for future work are reducing the elicitation cost itself, for example through fewer samples or shorter prompts, and finding a faster way to select across providers and prior families without having to fit and evaluate every combination.

The set of frontier LLMs evaluated here is a moving target: newer GPT-5, Claude, and Gemini variants have been released since the elicitation runs were closed, and whether the same selection pipeline produces comparable gains on the current frontier is left to future work.

6.5 Future Work

There are many directions this work could evolve into.

The most direct extension from this thesis is to test the methodology on a larger and more carefully chosen set of datasets, ideally datasets that are unlikely to appear in LLM training corpora and that allow multiple distinct OOD splits per dataset. The current results conflate "LLM prior helps OOD prediction" with "LLM prior helps the specific OOD split chosen," and separating that would strengthen the conclusions substantially.

A second direction is to use the prompt-rephrasing mechanism of Capstick et al. (2025) as the elicitation step in this pipeline, either in place of temperature sampling or alongside it. The elicitation step that produces the M samples can be swapped with this method, so AutoElicit's rephrasing approach could be used directly while keeping the downstream estimators, λ tuning, and provider-family selection unchanged. Rephrasing captures cross-prompt variability that temperature sampling alone does not reach, so using both should produce a richer distribution over the LLM's belief structure than either technique alone.

Another way to evolve the methods in this thesis would be to use the underlying log probabilities of the LLM's outputs to estimate uncertainty. This is close to Rezaeimanesh and Ghassemi (2026), which reads the logits over semantically opposed tokens like "positive" and "negative". A natural variant would be to keep the coefficient-elicitation prompt used here and read the log-probabilities of the actual numeric tokens the LLM emits as the coefficient value, which would preserve the magnitude information that the semantic-token approach discards and could in principle yield an uncertainty estimate from far fewer samples. The challenge is that a value like 0.35 usually fragments across several tokens ("0", ".", "3", "5"), and there is no obvious way to combine their probabilities into a single variance estimate. Whether this direction has any merit remains to be seen.

The first natural next step is a shared benchmark for LLM-aided tabular prediction under distribution shift that compares this thesis's method, AutoElicit, LoID, and LLM Processes as baselines on a common set of OOD splits which all future models could compare with. This is the most important next step for gaining a complete understanding about which methods systematically outperforms others, and which direction to move in for the future.

Chapter 7

Conclusion

This thesis investigated whether prior information elicited from an LLM can improve predictive performance under domain shift when used as a coefficient prior in regularized linear and logistic regression. Across five tabular datasets, the prior-informed estimators produced statistically meaningful improvements on four of the five tasks, with the clearest gains on the regression datasets and a 15.8% reduction in OOD MSE on Boston Housing. On Heart Disease and Diabetes, OOD AUC barely moved but F1, sensitivity, and probabilistic accuracy all improved: the priors did not change how predictions were ranked, but they shifted the fitted decision boundary toward better positive-class recall and produced better-calibrated probabilities. Adult illustrated the opposite case, where the baseline already saturated AUC and cross-validation drove the prior strength to zero, leaving the prior-informed estimators tied with baseline within practical noise. This is the safe-failure mode the methodology was designed to allow, showing that the benefit is real but dataset-dependent rather than universal.

Three contributions follow from this work. First, the global-λ treatment of the LLM-elicited covariance: the shape of the prior is read from the LLM while its scale is fit to the source data, which sidesteps the miscalibration of LLM-stated uncertainty documented in earlier work without discarding the structural information the LLM provides. Second, the cross-provider, cross-family selection pipeline: the deployed lift consistently lands close to the per-dataset OOD-best ceiling and well above the floor that any single (provider, family) commitment could risk, which gives the method the potential to be deployed in practice. Third, the thesis adds a new route for extracting prior information from LLMs (sampled coefficient values with variance and covariance estimation) alongside AutoElicit's direct standard-deviation elicitation and LoID's token log-probabilities.

This area of research is moving quickly, with several closely related papers appearing during the writing of this thesis. The results reported here, alongside that broader line of work, suggest that LLM-elicited priors are worth pursuing further as one of the few approaches to improving OOD prediction that scales without retraining or human domain expert elicitation.

Bibliography

Amin, M. T., Ahmmod, T., Ferdus, Z., Naem, T. N. H., Ferdous, E., Bhattacharjee, A., Solaiman, I. A., & Noor, N. B. (2026). Integrating machine learning ensembles and large language models for heart disease prediction using voting fusion. https://arxiv.org/abs/2602.22280

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., … Liang, P. (2022). On the opportunities and risks of foundation models. https://arxiv.org/abs/2108.07258

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. https://arxiv.org/abs/2005.14165

Capstick, A., Krishnan, R. G., & Barnaghi, P. (2025). Autoelicit: Using large language models for expert prior elicitation in predictive modelling. https://arxiv.org/abs/2411.17284

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

Detrano, R., Jánosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J.-J., Sandhu, S., Guppy, K. H., Lee, S., & Froelicher, V. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. *The American Journal of Cardiology*, *64*(5), 304–310. https://doi.org/10.1016/0002-9149(89)90524-9

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Chapman & Hall.

Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., Yu, T., Zhang, R., & Ahmed, N. K. (2024). Bias and fairness in large language models: A survey. https://arxiv.org/abs/2309.00770

Goh, E., Gallo, R., Hom, J., et al. (2024). Large language model influence on diagnostic reasoning: A randomized clinical trial. *JAMA Network Open*, *7*(10), e2440969. https://doi.org/10.1001/jamanetworkopen.2024.40969

Harrison, D., & Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, *5*(1), 81–102. https://doi.org/10.1016/0095-0696(78)90006-2

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd). Springer.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67. https://doi.org/10.1080/00401706.1970.10488634

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., et al. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, *43*(2), Article 73. https://doi.org/10.1145/3703155

Kohavi, R. (1996). Scaling up the accuracy of naive-Bayes classifiers: A decision-tree hybrid. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD)*, 202–207.

Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. https://arxiv.org/abs/1612.01474

Liu, J., Shen, Z., He, Y., Zhang, X., Xu, R., Yu, H., & Cui, P. (2023). Towards out-of-distribution generalization: A survey. https://arxiv.org/abs/2108.13624

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, *55*(9), Article 195. https://doi.org/10.1145/3560815

Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., & Gao, J. (2024). Large language models: A survey.

O'Hagan, A. (2019). Expert knowledge elicitation: Subjective but scientific. *The American Statistician*, *73*(sup1), 69–81. https://doi.org/10.1080/00031305.2018.1518265

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J. V., Lakshminarayanan, B., & Snoek, J. (2019). Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. https://arxiv.org/abs/1906.02530

Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A., & Lawrence, N. D. (Eds.). (2009). *Dataset shift in machine learning*. MIT Press.

Requeima, J., Bronskill, J., Choi, D., Turner, R. E., & Duvenaud, D. (2024). Llm processes: Numerical predictive distributions conditioned on natural language. https://arxiv.org/abs/2405.12856

Rezaeimanesh, S., & Ghassemi, M. M. (2026). What language models know but don't say: Non-generative prior extraction for generalization. https://arxiv.org/abs/2601.17609

Riegler, M. A., Hellton, K. H., Thambawita, V., & Hammer, H. L. (2025). Using large language models to suggest informative prior distributions in bayesian statistics. https://arxiv.org/abs/2506.21964

Rippner, N. (2017). Cancer trials [Compiled from US Census American Community Survey, ClinicalTrials.gov, and Cancer.gov; accessed via Kaggle]. https://data.world/nrippner/cancer-trials

Sainz, O., Campos, J. A., García-Ferrero, I., Etxaniz, J., de Lacalle, O. L., & Agirre, E. (2023). Nlp evaluation in trouble: On the need to measure llm data contamination for each benchmark. https://arxiv.org/abs/2310.18018

Smith, J. W., Everhart, J. E., Dickson, W. C., Knowler, W. C., & Johannes, R. S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. *Proceedings of the Annual Symposium on Computer Application in Medical Care*, 261–265.

Sokolov, A., Carlin, D. E., Paull, E. O., Baertsch, R., & Stuart, J. M. (2016). Pathway-based genomics prediction using generalized elastic net. *PLOS Computational Biology*, *12*(3), 1–23. https://doi.org/10.1371/journal.pcbi.1004790

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the royal statistical society: Series B (Methodological)*, *36*(2), 111–133.

van de Schoot, R., Depaoli, S., King, R., Kramer, B., Märtens, K., Tadesse, M., Vannucci, M., Gelman, A., Veen, D., Willemsen, J., & Yau, C. (2021). Bayesian statistics and modelling. *Nature Reviews Methods Primers*, *1*(1). https://doi.org/10.1038/s43586-020-00001-2

Wang, S., Kreuzer, M., Braun, U., & Schwarm, A. (2017). Effect of unconventional oilseeds (safflower, poppy, hemp, camelina) on in vitro ruminal methane production and fermentation. *Journal of the Science of Food and Agriculture*, *97*(11), 3864–3870. https://doi.org/10.1002/jsfa.8260

Zeng, Y., Liu, J., Lam, H., & Namkoong, H. (2024). Llm embeddings improve test-time adaptation to tabular $Y|X$-shifts. https://arxiv.org/abs/2410.07395

Zhang, E., Goto, R., Sagan, N., Mutter, J., Phillips, N., Alizadeh, A., Lee, K., Blanchet, J., Pilanci, M., & Tibshirani, R. (2025). Llm-lasso: A robust framework for domain-informed feature selection and regularization. https://arxiv.org/abs/2502.10648

Appendix A

Prompt and Reasoning-Mode Response Example

This appendix reproduces one full prompt-response cycle for the Cancer dataset under the OpenAI GPT-5.1 reasoning-mode configuration (sample 0 of 50). The system and user prompts are reproduced verbatim. The response is the contents of the `reasoning` and `coefficients` fields of the model's structured JSON output; this is the model's self-narrated reasoning written into the response schema, not the hidden chain-of-thought tokens that some reasoning APIs expose separately.

A.1 System prompt

> You are helping construct informative priors for a regression model. Adopt the perspective of an epidemiologist specializing in population cancer outcomes, health disparities, and interpretable statistical modeling. Use the feature definitions and feature-scale summaries to reason about BOTH sign and magnitude. All coefficients must be given in the original raw-feature units used by the model matrix. Do not standardize, rescale, or reinterpret the coefficients as standardized effects. Do not include an intercept. Before committing to numbers, reason step-by-step: for each predictor, weigh its likely sign, its magnitude given the feature-scale summary, and the mechanism that links it to the outcome. Also consider how predictors overlap, so that correlated variables share signal rather than each getting a large coefficient. Record that reasoning in the 'reasoning' field of the response JSON, then write the numeric coefficients in the 'coefficients' field. Return a single JSON object with exactly the keys 'reasoning' and 'coefficients'. Do not wrap it in markdown or code fences. Return one plausible coefficient vector draw from your belief distribution. Do not collapse uncertainty into a single average if multiple plausible settings exist. Keep the draw realistic and domain-grounded rather than intentionally noisy. Prefer interpretable, conservative, domain-plausible coefficients. Avoid extreme magnitudes unless they are clearly justified by the variable scale and domain mechanism.

A.2 User prompt

```
Task context
- County-level linear regression for TARGET_deathRate: y = intercept + sum_j beta_j
↪  * x_j

Domain context:
- Source domain is U.S. South-region counties (AL, AR, DE, DC, FL, GA, KY, LA, MD,
↪  MS, NC, OK, SC, TN, TX, VA, WV).
- Priors should capture broad, transferable structural effects rather than local
↪  artifacts.

Dataset notes:
- This is an ecological county-level dataset, so effects are population-level
↪  associations rather than individual-level clinical effects.
- Race-composition variables should be handled cautiously as structural correlates,
↪  not causal or essentialist claims.

Guidance:
- Use plausible signs grounded in epidemiology.
- Keep magnitudes realistic and conservative for county-level regressors.
- Favor stable, interpretable structural effects over local artifacts.

Feature meanings:
- incidenceRate: cancer incidence rate
- medIncome: median household income
- povertyPercent: percent of residents below the poverty line
- MedianAge: median age
- PctHS25_Over: percent ages 25+ with high school as highest education
- PctBachDeg25_Over: percent ages 25+ with a bachelor's degree
- PctEmployed16_Over: employment rate among ages 16+
- PctUnemployed16_Over: unemployment rate among ages 16+
- PctPrivateCoverage: percent with private health coverage
- PctPublicCoverage: percent with public health coverage
- PctBlack: percent Black
- PctOtherRace: percent other race
- PctMarriedHouseholds: percent of households that are married-couple households

Feature scale summaries from the actual design matrix:
- incidenceRate: continuous/raw numeric; median=454, IQR=[420, 481],
↪  central95%~=[355, 525]
- medIncome: continuous/raw numeric; median=4.52e+04, IQR=[3.89e+04, 5.25e+04],
↪  central95%~=[3.18e+04, 7e+04]
- povertyPercent: continuous/raw numeric; median=15.9, IQR=[12.1, 20.4],
↪  central95%~=[8.3, 28.7]
- MedianAge: continuous/raw numeric; median=41, IQR=[37.7, 44], central95%~=[31.9,
↪  50.2]
- PctHS25_Over: continuous/raw numeric; median=35.3, IQR=[30.4, 39.7],
↪  central95%~=[22.2, 45.4]
- PctBachDeg25_Over: continuous/raw numeric; median=12.3, IQR=[9.4, 16.1],
↪  central95%~=[6.4, 23.5]
- PctEmployed16_Over: continuous/raw numeric; median=54.5, IQR=[48.9, 60],
↪  central95%~=[40.2, 66.3]
```

```
- PctUnemployed16_Over: continuous/raw numeric; median=7.6, IQR=[5.5, 9.7],
↪  central95%~=[2.8, 13.8]
- PctPrivateCoverage: continuous/raw numeric; median=65.1, IQR=[57.2, 72.1],
↪  central95%~=[46.4, 80.6]
- PctPublicCoverage: continuous/raw numeric; median=36.3, IQR=[30.9, 41.5],
↪  central95%~=[23, 49]
- PctBlack: continuous/raw numeric; median=2.25, IQR=[0.621, 10.5],
↪  central95%~=[0.101, 42.2]
- PctOtherRace: continuous/raw numeric; median=0.826, IQR=[0.295, 2.18],
↪  central95%~=[0.0141, 7.86]
- PctMarriedHouseholds: continuous/raw numeric; median=51.7, IQR=[47.8, 55.4],
↪  central95%~=[39.7, 61]

Important modeling reminders:
- Binary indicators use 0/1 coding, so their coefficients represent the change
↪  associated with turning the indicator on.
- When multiple variables capture overlapping structure, distribute signal
↪  conservatively rather than giving each one a large effect.
- The output schema already fixes the keys. Choose only the numeric values.
```

Note: the "∼=" marker above is rendered as ≈ (approximately equal to) in the prompt actually sent to the model.

A.3 Response: reasoning

I treat TARGET_deathRate as an age-adjusted cancer mortality rate per 100,000, similar in scale to incidenceRate. I'm constructing coefficients in raw units, thinking about (1) plausible direction of association, (2) how large a 1-unit change is given the observed ranges, and (3) overlap among SES, coverage, and race variables so no single predictor carries an unrealistically large effect.

1) incidenceRate - Mechanism: More incident cancer cases per 100k should broadly translate into more cancer deaths, modulated by stage at diagnosis and treatment quality. At the county level, even with some variation in survival, mortality will move substantially with incidence. - Scale: central 95% ≈ 355–525; a range of ∼170. If each additional incident case per 100k produced ∼0.3 additional deaths per 100k, that would yield about 0.3*170 ≈ 51 deaths difference across that range, which is plausible given that county all-cancer mortality can vary by ∼100+ per 100k. - Collinearity: incidence overlaps somewhat with age and access, but is still a primary driver. I give it a moderately strong positive slope, not so large that it dominates everything. → Chosen: incidenceRate = +0.30 deaths per 100k per 1-unit increase in incidence.

2) medIncome - Mechanism: Higher median household income is generally protective via better access to care, earlier diagnosis, healthier environments, and lower comorbidity burden. At the ecological level this usually shows as lower mortality. - Scale: 95% ≈ 31.8k–70k (range ∼38k). A slope of -0.0003 implies a 10,000 USD increase in income lowers mortality by about 3 deaths per 100k; across the full 38k range, about 11–12 deaths difference. That's meaningful but not huge. - Overlap: Income is strongly correlated with education, employment,

insurance, and marital structure, so I keep this effect moderate so it shares signal with those variables rather than overwhelming them. → Chosen: medIncome = -0.0003 deaths per 100k per $1.

3) povertyPercent - Mechanism: Higher poverty is associated with worse baseline health, later stage at diagnosis, poorer treatment, and generally higher cancer mortality. - Scale: 95% ≈ 8.3–28.7% (range ∼20 percentage points). A slope of +0.35 gives about 7 deaths per 100k difference over that span, which is modest but important. - Overlap: Poverty is inversely related to income, education, employment, and private coverage, so I keep it moderate to avoid redundancy. → Chosen: povertyPercent = +0.35 deaths per 100k per 1 percentage-point increase.

4) MedianAge - Mechanism: Older populations have both higher cancer incidence and higher cancer mortality, even after age-adjustment artifacts and some correlation with public coverage. - Scale: 95% ≈ 31.9–50.2 (≈18 years). A slope of +0.8 implies roughly 14–15 deaths per 100k extra mortality from the youngest to oldest typical counties, which is plausible for structural age differences. - Overlap: Age is correlated with public coverage and incidence, but is conceptually distinct; giving it a moderate coefficient spreads risk across these. → Chosen: MedianAge = +0.8 deaths per 100k per year.

5) PctHS25_Over - Mechanism: Higher share of adults whose highest education is high school (with relatively fewer BA/advanced degrees) reflects lower educational attainment overall, associated with lower health literacy, more risk factors, and worse access. - Scale: 95% ≈ 22.2–45.4 (range ∼23 points). With a coefficient around +0.18, going from the lowest to highest share adds about 4 deaths per 100k. That leaves room for the BA variable to capture additional educational protection. - Overlap: Since this is partly the complement of higher education, I keep it modest. → Chosen: PctHS25_Over = +0.18 deaths per 100k per 1 percentage point.

6) PctBachDeg25_Over - Mechanism: Higher BA attainment is strongly protective via income, health behaviors, navigation of care, and community resources. After adjusting for income and coverage, we still expect some residual protective effect. - Scale: 95% ≈ 6.4–23.5 (range ∼17). A slope of -0.25 yields about -4 to -5 deaths per 100k across that range, which is reasonable after controlling for related SES factors. - Overlap: Strongly correlated with income, private coverage, and married households, so I choose a moderate negative rather than a very large one. → Chosen: PctBachDeg25_Over = -0.25 deaths per 100k per 1 percentage point.

7) PctEmployed16_Over - Mechanism: Higher employment rates reflect stronger local economies, employer-based insurance, and better material conditions, so we expect a protective association. - Scale: 95% ≈ 40.2–66.3 (range ∼26). With -0.12, the full range yields about -3 deaths per 100k, small but directionally consistent. - Overlap: This overlaps with income and poverty, so I deliberately keep the effect small. → Chosen: PctEmployed16_Over = -0.12 deaths per 100k per 1 percentage point.

8) PctUnemployed16_Over - Mechanism: Higher unemployment indicates economic distress and instability, often linked to loss of insurance, delayed care,

and poorer health. - Scale: 95% ≈ 2.8–13.8 (range ∼11). With +0.20, that implies roughly +2 deaths per 100k across the typical range, which is modest. - Overlap: Since employment, unemployment, poverty, and income are interrelated, I keep both employment and unemployment coefficients conservative to avoid double-counting. → Chosen: PctUnemployed16_Over = +0.20 deaths per 100k per 1 percentage point.

9) PctPrivateCoverage - Mechanism: Higher private coverage generally signals better access to timely diagnosis and treatment, more comprehensive benefits, and higher SES, so it should be protective. - Scale: 95% ≈ 46.4–80.6 (range ∼35). A coefficient of -0.20 yields about -7 deaths per 100k difference across that span, consistent with an important but not dominant protective factor once income, education, and employment are accounted for. → Chosen: PctPrivateCoverage = -0.20 deaths per 100k per 1 percentage point.

10) PctPublicCoverage - Mechanism: Higher public coverage (Medicare, Medicaid) at the county level typically indicates an older and/or more medically burdened poor population. For cancer, more Medicare-eligible older adults generally implies higher mortality; higher Medicaid share often correlates with poverty and comorbidity. - Scale: 95% ≈ 23–49 (range ∼26). A coefficient of +0.30 gives about +7.8 deaths per 100k through that range, which is plausible after accounting for age and SES. - Overlap: Because this overlaps with age and poverty, I avoid a very large positive coefficient. → Chosen: PctPublicCoverage = +0.30 deaths per 100k per 1 percentage point.

11) PctBlack - Mechanism: In the U.S. South, counties with higher Black population share generally experience higher cancer mortality due to structural racism, residential segregation, limited access to high-quality care, and cumulative disadvantage. Many of these mechanisms are captured by SES and access variables, so the residual association should be positive but tempered. - Scale: 95% ≈ 0.1–42.2 (range ∼42). With +0.18, moving from very low to high Black share is associated with about +7–8 deaths per 100k, consistent with a meaningful but not extreme residual disparity after adjustment. → Chosen: PctBlack = +0.18 deaths per 100k per 1 percentage point.

12) PctOtherRace - Mechanism: This category aggregates diverse groups (e.g., Asian, Native American, multiracial). In many Southern counties, higher "other race" share is often driven by groups that, on average, do not have systematically higher all-site cancer mortality than White populations, and some (e.g., Asian) often have lower mortality. Ecologically this might show as neutral or slightly protective after SES adjustment. - Scale: 95% ≈ 0.014–7.86 (range ∼7.8). With a coefficient of -0.12, the total effect across the entire range is on the order of -1 death per 100k, i.e., very small. → Chosen: PctOtherRace = -0.12 deaths per 100k per 1 percentage point.

13) PctMarriedHouseholds - Mechanism: Higher proportion of married-couple households often proxies more social support, higher incomes, and more stable living situations, generally associated with better health outcomes and earlier cancer diagnosis. - Scale: 95% ≈ 39.7–61 (range ∼21). A coefficient of -0.15 implies about -3 deaths per 100k over this range, which is modest but directionally appropriate. - Overlap: This overlaps with income, education, and employment;

hence the modest magnitude.

Overall, these coefficients together imply that moving from a low-incidence, younger, socioeconomically advantaged county to a high-incidence, older, disadvantaged county could produce differences on the order of tens of deaths per 100k, which is consistent with observed county-level variation in cancer mortality. At the same time, correlated SES and coverage variables each have moderate rather than extreme coefficients, so they share explanatory weight rather than any single variable dominating the structural prior.

A.4 Response: coefficients


```
{
  "incidenceRate": 0.3,
  "medIncome": -0.0003,
  "povertyPercent": 0.35,
  "MedianAge": 0.8,
  "PctHS25_Over": 0.18,
  "PctBachDeg25_Over": -0.25,
  "PctEmployed16_Over": -0.12,
  "PctUnemployed16_Over": 0.2,
  "PctPrivateCoverage": -0.2,
  "PctPublicCoverage": 0.3,
  "PctBlack": 0.18,
  "PctOtherRace": -0.12,
  "PctMarriedHouseholds": -0.15
}
```

Appendix B

Adult Prior Diagnostics

The per-feature LLM prior versus the OOD-oracle coefficient and the per-family coefficient-shrinkage diagram for Adult, deferred from Section 5.2. They support the finding noted there: the LLM prior on Adult is essentially a zero vector with tight intervals and systematic under-shooting of the oracle coefficients, and the tiny tuned λ leaves Independent and Correlated anchored near the unregularized reference rather than near the prior, because the data signal dominates.

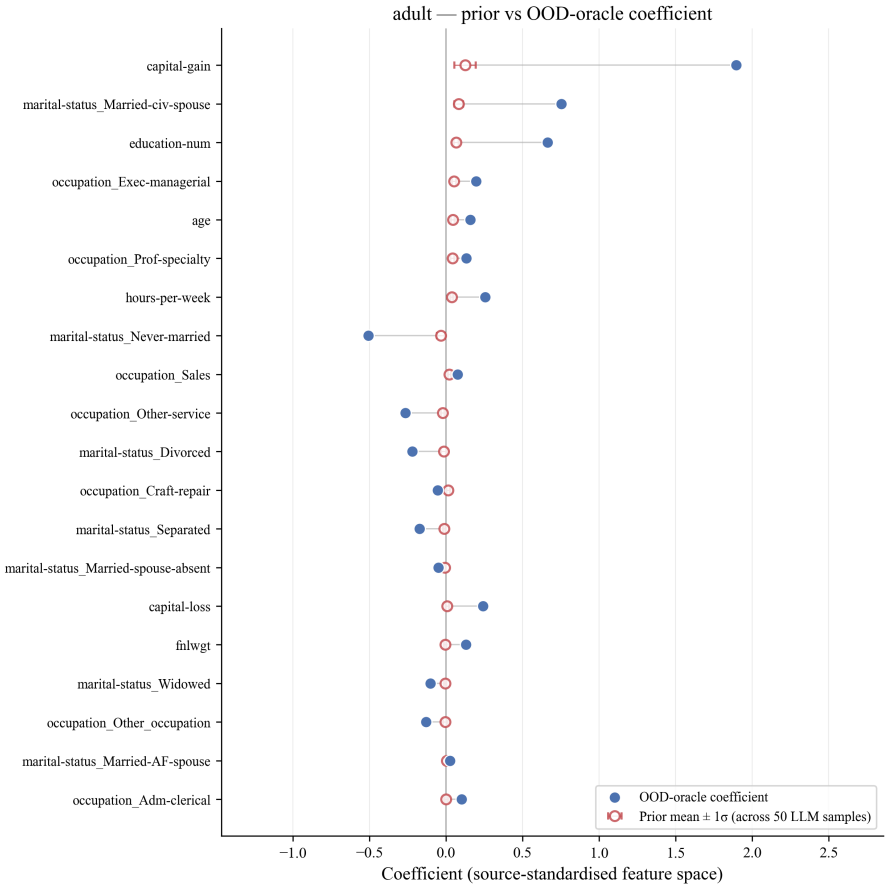


Figure B.0.1: *Per-feature comparison of the LLM prior and the OOD-oracle coefficient on Adult. Blue circles: OOD-oracle coefficient. Open red circles with whiskers: prior mean $\pm 1\sigma$ across 50 LLM samples from the deployment-winning LLM. Gray segments connect the two markers per feature as a visual shrinkage axis. Both axes in source-standardized feature space.*

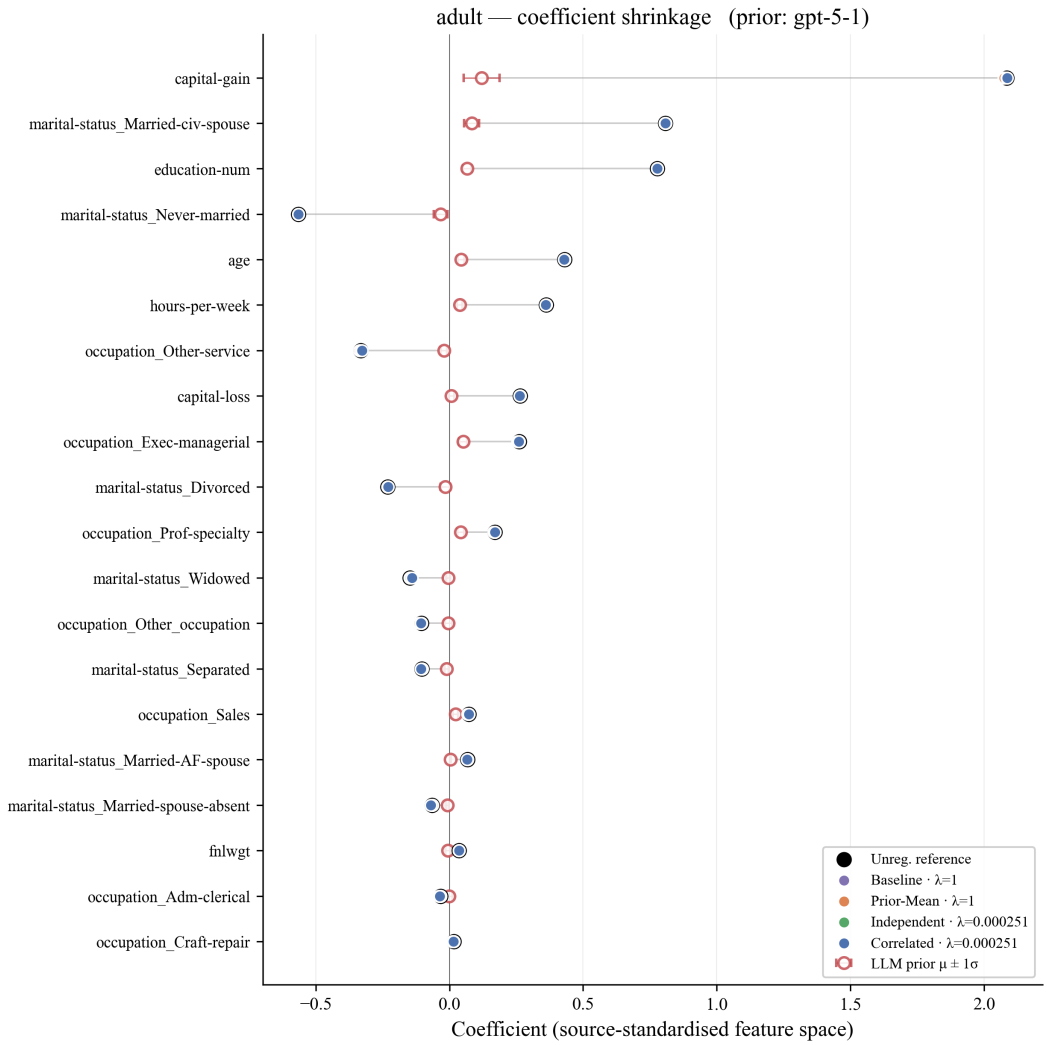


Figure B.0.2: *Coefficient shrinkage on Adult. Black filled circle: unregularized reference fit on the source split. Open red circle with whiskers: LLM prior mean $\pm 1\sigma$. Colored filled circles: each model family's deployed coefficient at its tuned λ. Gray segment per feature shows the reference→prior shrinkage axis. All in source-standardized feature space.*



Norges miljø- og biovitenskapelige universitet
Noregs miljø- og biovitskaplege universitet
Norwegian University of Life Sciences

Postboks 5003
NO-1432 Ås
Norway