Large Language Models as Priors for Randomized Experiments *

Alexis Perakis and Samad Sarferaz

KOF ETH Zurich

June 1, 2026

Abstract

We develop a framework for using large language models as priors in randomized experiments. The method augments experimental data with LLM-generated synthetic observations, treated not as substitutes for participants but as weighted dummy observations that encode experiment-specific prior information. The weight on the synthetic block is learned from the data subject to a finite-sample safe-region bound, yielding a disciplined bias–variance trade-off that preserves identification from random assignment. We characterize the estimator's bias and variance, derive conditions under which synthetic priors improve precision, and implement posterior inference through a Metropolis-within-Gibbs algorithm. We apply the method to two settings: a household inflation-expectations information experiment and a controlled forecasting experiment. In low-power designs, LLM-informed priors reduce posterior standard deviations by roughly one third, equivalent to more than doubling the effective sample size, while their gains decline as real-data evidence accumulates. The results show that LLMs can provide useful, transparent, and bounded prior information for experimental inference without replacing real observations.

*The authors thank Elliott Ash, Olivier Coibion, Maurizio Daniele, Andreas Dibiasi, Hans Gersbach, Domenico Giannone, Yuriy Gorodnichenko, Lukas Hack, Stephen Hansen, Daniel Kaufmann, Christian Matthes, Franck Portier, Frank Schorfheide, Adrien Wicht, Timo Wochner, Jan-Egbert Sturm and participants of the Unine-KOF-UniBasel PhD Macro Workshop 2026, the KOF Brown Bag Seminar, the 6th Dolomiti Macro Meetings, and the 12th Monash-Paris-Warwick-Zurich-CEPR Text-As-Data Workshop for helpful comments and suggestions. Authors' contact: perakis@kof.ethz.ch and sarferaz@kof.ethz.ch.

1 Introduction

Empirical researchers increasingly have access to a new form of structured background information: outputs from large language models (LLMs) that can be conditioned on experimental designs, participant characteristics, treatments, and outcome questions. This creates a natural opportunity for Bayesian empirical work, where prior information can be valuable but useful priors are often difficult to construct. LLMs offer a way to translate broad regularities about beliefs and behavior into experiment-specific prior information. This opportunity is especially relevant for randomized experiments, where identification is credible but statistical power is often limited by costly recruitment, small target populations, and multi-arm designs.

We operationalize this idea by treating LLM-generated outputs as weighted dummy observations in the experimental dataset. The synthetic observations are constructed to mirror the structure of the experiment, including its covariates, treatments, and outcomes, and are appended to the human sample. A scalar weight, ω, governs how much influence this synthetic block has on estimation. This weighted augmentation has a Bayesian interpretation: LLM-generated observations provide structured prior information, while the randomized experimental data remain the source of causal identification. The premise is therefore not that synthetic observations replace human responses, but that they provide a disciplined way to incorporate prior information when real experimental evidence is limited.

We formalize this idea in a framework for estimating treatment effects with a finite experimental sample and a synthetic auxiliary block. Incorporating the synthetic block induces a bias–variance trade-off: increasing its weight can improve precision, but may introduce bias if the synthetic and real conditional outcome processes differ. We characterize this trade-off theoretically and derive a conservative safe region for the dummy-observation weight ω, which prevents the synthetic block from dominating the real-data likelihood. Within this region, ω is learned from the data using the marginal likelihood of the augmented sample. The resulting procedure allows synthetic observations to contribute when they are informative, while preventing them from overwhelming the experimental evidence.

We evaluate the method in two complementary empirical settings. The first is the household inflation-expectations experiment of Coibion, Gorodnichenko and Weber (2022), a conventional information-provision experiment with multiple treatment arms and demographic controls. The second is the forecasting experiment of Afrouzi et al. (2023), where partici-

pants forecast controlled AR(1) processes and the estimand is forecast-implied persistence. Together, these applications allow us to assess the method in both a standard randomized information-provision design and a tightly controlled forecasting environment. In each case, we construct underpowered empirical designs from the original data, augment the reduced real-data sample with LLM-generated synthetic dummy observations, and compare real-only and mixed posterior inference. This strategy allows us to quantify whether synthetic augmentation can recover statistical power that would otherwise require additional real observations, how it affects posterior uncertainty and point estimates, and how the gains vary as the amount of real evidence increases.

The empirical results show that the method can deliver large precision gains in empirically relevant low-power settings. In the inflation-expectations experiment, our benchmark specification with 500 real observations reduces the posterior standard deviation of treatment-effect estimates by roughly 35 percent relative to the real-data-only analysis. Achieving the same reduction by increasing sample size alone would require approximately 135 percent more real observations. The stress-test results show that these gains decline smoothly as the real sample grows, remaining positive even when the sample is extended up to 20,000 observations, but becoming very small once the real-data likelihood is already highly informative. The forecasting experiment delivers a similar message in a different setting: LLM-generated dummy observations reduce uncertainty in estimates of forecast-implied persistence, especially when the real panel is shortened to only a small number of usable forecast rounds per participant. Across both applications, the results support the interpretation of the method as a disciplined bias–variance tradeoff: synthetic observations improve precision when real experimental evidence is scarce, while their influence is learned subject to an explicit safe-region bound that keeps inference anchored in real data.

Recent work has started to connect LLMs to experimental and expectations research.[1] Horton, Filippas and Manning (2023) argue that modern LLMs can be treated as "homo silicus"—implicit computational models that, when endowed with information and roles, reproduce qualitative patterns from classic experiments and can be used to pilot designs in silico. Morevoer, applied demonstrations show LLMs can emulate specific survey populations: Hansen et al. (2024) simulate the Survey of Professional Forecasters by building forecaster personas, feeding them real-time macro data, and find AI forecasts to often match or exceed human accuracy, while emphasizing LLMs as complements rather than substi-

[1]For overviews on LLMs in economics see Ash and Hansen (2023), Korinek (2023), Ash, Hansen and Muvdi (2024) and Korinek (2025).

tutes for human respondents. In inflation-expectations settings, Zarifhonarvar (2026) runs information-provision experiments on LLMs, documenting belief updating and treatment effects that qualitatively mirror human surveys, highlighting the roles of knowledge sources, model choice, and persona prompts. Work by Wu, Xi and Xie (2025) demonstrates that LLMs can replicate state-dependent treatment effects, especially in the context of variable inflation regimes.[2] Against this backdrop, Ludwig, Mullainathan and Rambachan (2026) provide an econometric "contract" for using LLM outputs in empirical work—warning that their use requires benchmark/validation data to model LLM error—implying that in-silico subjects should amplify, not replace, human data. Our approach closely aligns with that caution by leveraging dummy-observations and their Bayesian interpretation: rather than treating LLM outputs as real-world data, we use them only as structured prior information, aiming to alleviate power constraints in experiments while keeping identification anchored in human observations.

Our paper builds on the econometric literature on dummy-observation priors. Theil mixed estimators provide a classical way to incorporate prior information through artificial observations (Theil and Goldberger, 1961; Theil, 1963). This idea has been especially influential in Bayesian macroeconometrics, where dummy observations are used to encode shrinkage priors in VARs (e.g., Litterman, 1986; Doan, Litterman and Sims, 1984). Relatedly, the DSGE-VAR approach of Del Negro and Schorfheide (2004) uses artificial observations implied by a structural DSGE model to construct priors for VAR estimation. More recent work studies how the tightness of such priors can be calibrated or learned in high-dimensional VARs (e.g., Bańbura, Giannone and Reichlin, 2010; Giannone, Lenza and Primiceri, 2015). We extend this logic to randomized experiments, where the auxiliary observations are generated by LLMs and therefore require explicit safeguards against misspecification. Our contribution is to treat these observations as experiment-specific prior information, derive finite-sample bias–variance bounds for their use, and learn their weight subject to a safe region that keeps causal identification anchored in the randomized real-data sample.

We contribute to the literature that seeks to improve the precision of treatment-effect estimators in randomized experiments. We are closest to Iacovone, McKenzie and Meager (2025), who use informative priors elicited from experts and firms. They operationalize Bayesian approaches to randomized experiments discussed in Spiegelhalter, Freedman and Parmar (1994), Gelman et al. (1995) and Imbens and Rubin (2015). Crucially, our priors

[2]For other LLM applications see for example Bybee (2023), Manning, Zhu and Horton (2024), Kazinnik and Sinclair (2025), Brynjolfsson et al. (2025), Cook et al. (2025), Manning and Horton (2026).

are elicited from LLMs, which are instructed to emulate human behavior.[3] While Iacovone, McKenzie and Meager (2025) explicitly elicit full prior distributions from respondents, our framework does so implicitly. By using dummy observations, we simplify prior elicitation and provide substantial flexibility in prior representation. Studies motivated by external validity have used Bayesian hierarchical analysis (Meager, 2019, 2022; Vivalt, 2020; Bandiera et al., 2021; Crosta et al., 2024), while others, such as Coey and Cunningham (2019), have leveraged empirical Bayes to shrink treatment effect estimators.[4] This paper also connects to recent work studying the informativeness of ex-ante expert forecasts about experimental results (DellaVigna and Pope, 2018; DellaVigna, Otis and Vivalt, 2020). Under the assumption that LLMs contain a form of wisdom of crowds, we use analogous ex-ante beliefs as prior information for treatment-effect estimation.

We also contribute to the growing literature on information-provision experiments,[5] focusing in particular on settings where sample size constraints limit statistical power. These include firm samples, where recruitment is costly and response rates are low (e.g., Drechsel et al., 2015; Link et al., 2023; Mikosch et al., 2024; Dibiasi, Mikosch and Sarferaz, 2025; Weber et al., 2025; Abberger et al., 2026), experiments with many treatment arms (e.g. Coibion, Gorodnichenko and Weber, 2022; Ash et al., 2026) and expert-elicitation studies in macroeconomics, finance, or policy analysis (e.g., Andre et al., 2022, 2026) where access to participants is inherently limited and cell sizes are small. Traditional remedies for low power—such as pooling across waves or reducing treatment arms—can weaken design goals or hinder heterogeneity analysis. By introducing LLM-generated responses as dummy-observations, our method offers a flexible means of incorporating structured prior information without replacing actual participants, preserving design credibility while improving precision, power, and robustness in contexts where expanding the real-world sample is prohibitively difficult.

The remainder of the paper is organized as follows. Section 2 introduces the synthetic dummy observation priors and their theoretical properties. Section 3 presents the Bayesian estimation of the model parameters. Section 4 describes the synthetic data generation process with LLMs and Section 5 illustrates the approach. Section 6 concludes.

[3] Wu, Xi and Xie (2025) propose a perfectly balanced synthetic panel to reduce estimator variance, although their approach relies exclusively on LLM-generated observations. In a different spirit, Gui and Kim (2025) use LLMs to summarize a large number of covariates for stratification.

[4] For expositions of empirical Bayes, see Robbins (1964), Efron (2010), and Azevedo et al. (2019).

[5] For recent overviews see Haaland, Roth and Wohlfart (2023), Fuster and Zafar (2023), Stantcheva (2023) and Coibion and Gorodnichenko (2026).

2 Synthetic Data as Dummy–Observation Priors

This section introduces the dummy-observation prior used throughout the paper. In this and the next section, we take the LLM-generated synthetic block as given and append it to the real experimental sample as weighted dummy observations. We first define the augmented linear model and then characterize the bias–variance tradeoff induced by the weight ω, which controls the effective influence of the synthetic block.

Consider the standard Gaussian linear model

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

where $\boldsymbol{y}$ is an $n \times 1$ vector of observations, $\boldsymbol{X}$ is the $n \times k$ design matrix, $\boldsymbol{\beta}$ is the $k \times 1$ coefficient vector, and σ^2 is the error variance.

We augment the dataset with r LLM–generated dummy observations $(\boldsymbol{y}_d, \boldsymbol{X}_d)$, which encode prior beliefs about plausible coefficient values. To control their influence, each dummy row is multiplied by $\sqrt{\omega}$ before stacking with the real data:

$$\boldsymbol{X}_a = \begin{bmatrix} \boldsymbol{X} \\ \sqrt{\omega}\boldsymbol{X}_d \end{bmatrix}, \qquad \boldsymbol{y}_a = \begin{bmatrix} \boldsymbol{y} \\ \sqrt{\omega}\boldsymbol{y}_d \end{bmatrix}.$$

The augmented system can be written in the same form as the original model,

$$\boldsymbol{y}_a = \boldsymbol{X}_a\beta + \varepsilon_a, \tag{1}$$

where

$$\varepsilon_a = \begin{bmatrix} \varepsilon \\ \sqrt{\omega}\,\eta_d \end{bmatrix},$$

and where η_d collects the deviations of the LLM-generated dummy outcomes from the conditional mean structure implied by $\boldsymbol{X}_d\beta$. A formal decomposition of $\boldsymbol{y}_d$ and bounds on η_d are provided in Appendix A.1. We henceforth analyze $(\boldsymbol{y}_a, \boldsymbol{X}_a)$ as if all rows were actual observations: large values of ω make the dummy block exert strong prior influence, while $\omega = 0$ removes it entirely.

A central requirement for our mixed estimation framework is that the LLM-generated dummy observations capture regularities in the true data-generating process (DGP) in order

to serve as informative but imperfect prior information. In the next section, we lay out the assumptions that formalize this requirement and derive the finite-sample properties of the mixed estimator that follow from them.

2.1 Assumptions and Theoretical Properties

To formalize the conditions under which the mixed estimator is a valid and informative combination of real and synthetic data, we begin by stating two assumptions that characterize how the LLM-generated dummy observations may differ from the true data-generating process. These assumptions restrict the degree of misspecification in the synthetic block while allowing for flexible deviations in conditional means, variances, and support. Building on these conditions, we derive finite-sample bounds that characterize the bias–variance tradeoff induced by the dummy-observation prior and identify a conservative region for the prior weight ω within which the mixed estimator remains well behaved.

Let (Y, D, X) denote the outcome, treatment assignment, and covariates for the real observations, and let (Y_d, D_d, X_d) denote the corresponding quantities for the LLM-generated dummy observations. In the regression framework presented above (Equation 1), the treatment indicator D is included within the design matrix X (e.g. as treatment dummies), so the Gaussian linear model is only written in terms of (y, X).

Assumption 1 (Conditional moment closeness)**.** *For all (X, D) in the support of the real data,*

$$\big|\, E[Y_d \mid X, D] - E[Y \mid X, D] \,\big| \le \delta_1, \qquad \big|\, \mathrm{Var}(Y_d \mid X, D) - \mathrm{Var}(Y \mid X, D) \,\big| \le \delta_2.$$

Assumption 2 (Support overlap)**.** *The synthetic observations span the covariate space relevant for the real data:* $\mathrm{Supp}(X) \subseteq \mathrm{Supp}(X_d)$.

Assumption 1, together with the support overlap condition in Assumption 2, ensures that LLM-generated observations may be misspecified but nevertheless encode structured and relevant information about the conditional outcome distribution. Under these conditions, the mixed estimator can be interpreted as a bias–variance tradeoff between the real-data estimator and an LLM-informed dummy-observation prior. Because the treatment indicator D is included in the regression vector X (Equation 1), Assumption 1 also bounds potential misspecification in the treatment effect, which corresponds to the coefficient on D in β. The

following proposition formalizes these finite-sample properties.

Proposition 1 (Bias–variance bound for the mixed estimator)**.** *Consider the Gaussian linear model and augmented system described above. Let* $\boldsymbol{X}$ *be the* $n \times k$ *real-data design matrix and let* $\boldsymbol{X}_d$ *be the* $r \times k$ *dummy-data design matrix generated by the LLM. Let*

$$\hat{\beta}_{\text{real}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \qquad \textit{and} \qquad \hat{\beta}_{\text{mix}} = (\boldsymbol{X}'\boldsymbol{X} + \omega \boldsymbol{X}_d'\boldsymbol{X}_d)^{-1}\big(\boldsymbol{X}'\boldsymbol{y} + \omega \boldsymbol{X}_d'\boldsymbol{y}_d\big)$$

denote the real-data-only estimator and the mixed estimator, respectively. Under Assumptions 1 and 2 and bounded regressors, in the sense that there exists a constant $C_X < \infty$ *such that* $|X_{d,ij}| \le C_X$ *for all dummy observations* $i = 1, \ldots, r$ *and all regressors* $j = 1, \ldots, k$*. Let* $\lambda_{\min}(\boldsymbol{X}'\boldsymbol{X})$ *and* $\lambda_{\min}(\boldsymbol{X}_d'\boldsymbol{X}_d)$ *denote the smallest eigenvalues of the real and dummy design matrices, respectively.*

Then the bias of the mixed estimator satisfies

$$\big\|E[\hat{\beta}_{\text{mix}}] - \beta\big\| \;\le\; \frac{\omega\,\sqrt{k}\,C_X\,r\,\delta_1}{\lambda_{\min}(\boldsymbol{X}'\boldsymbol{X}) + \omega\,\lambda_{\min}(\boldsymbol{X}_d'\boldsymbol{X}_d)},$$

and its variance matrix satisfies

$$\big\|\mathrm{Var}(\hat{\beta}_{\text{mix}})\big\| \;\le\; \frac{\sigma^2}{\lambda_{\min}(\boldsymbol{X}'\boldsymbol{X}) + \omega\,\lambda_{\min}(\boldsymbol{X}_d'\boldsymbol{X}_d)}.$$

Consequently, the mean-squared error of $\hat{\beta}_{\text{mix}}$ *is bounded above by*

$$\mathrm{MSE}(\hat{\beta}_{\text{mix}}) \;\le\; \left(\frac{\omega\,\sqrt{k}\,C_X\,r\,\delta_1}{\lambda_{\min}(\boldsymbol{X}'\boldsymbol{X}) + \omega\,\lambda_{\min}(\boldsymbol{X}_d'\boldsymbol{X}_d)}\right)^2 + \frac{k\sigma^2}{\lambda_{\min}(\boldsymbol{X}'\boldsymbol{X}) + \omega\,\lambda_{\min}(\boldsymbol{X}_d'\boldsymbol{X}_d)}.$$

For the real-data-only estimator, $\mathrm{MSE}(\hat{\beta}_{\text{real}}) = \mathrm{trace}(\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}) \le k\sigma^2/\lambda_{\min}(\boldsymbol{X}'\boldsymbol{X})$*. Thus, for any value of* ω *satisfying*

$$\left(\frac{\omega\,\sqrt{k}\,C_X\,r\,\delta_1}{\lambda_{\min}(\boldsymbol{X}'\boldsymbol{X}) + \omega\,\lambda_{\min}(\boldsymbol{X}_d'\boldsymbol{X}_d)}\right)^2 + \frac{k\sigma^2}{\lambda_{\min}(\boldsymbol{X}'\boldsymbol{X}) + \omega\,\lambda_{\min}(\boldsymbol{X}_d'\boldsymbol{X}_d)} \;<\; \frac{k\sigma^2}{\lambda_{\min}(\boldsymbol{X}'\boldsymbol{X})},$$

the mixed estimator achieves a strictly smaller mean-squared error than the real-data-only estimator.

Proposition 1 provides a finite-sample characterization of how the dummy-observation

prior affects inference. The result shows that the influence of the synthetic block enters through two channels. (i) A bias term that depends on the relative weight of the misspecified dummy block. Its numerator increases with the dummy weight ω, the size of the dummy block, and the extent of LLM misspecification δ_1, while its denominator also increases with the information contained in the weighted dummy design. Hence, the bias bound rises approximately linearly for small values of ω, but eventually levels off as the mixed estimator approaches the LLM-only estimator. (ii) A variance term that declines with the combined information in the real and dummy design matrices. Together these bounds make transparent the bias–variance tradeoff implied by ω: for moderate values of ω, the variance reduction from the dummy observations dominates, improving precision, whereas for very large ω the estimator contracts towards the LLM-only estimator and the bias term becomes dominant.

Proposition 2 (A conservative safe region for the dummy-observation weight)**.** *Maintain the assumptions and notation of Proposition 1, and let*

$$A = \lambda_{\min}(\boldsymbol{X}'\boldsymbol{X})$$

denote the smallest eigenvalue of the real-data design matrix. Define

$$\bar{\omega} := \frac{\sigma\sqrt{A}}{C_X\, r\, \delta_1}.$$

Then for every weight $0 \le \omega \le \bar{\omega}$, the bias of the mixed estimator satisfies

$$\left\| E[\hat{\beta}_{\text{mix}}] - \beta \right\| \;\le\; \sqrt{k}\,\sigma,$$

i.e. the norm of the bias is no larger than the one-standard-error benchmark $\sqrt{k}\,\sigma$ implied by the real-data-only estimator. Moreover, for all such ω the variance of the mixed estimator is weakly smaller than that of the real-data-only estimator:

$$\left\| \mathrm{Var}(\hat{\beta}_{\text{mix}}) \right\| \;\le\; \left\| \mathrm{Var}(\hat{\beta}_{\text{real}}) \right\| = \left\| \sigma^2 (\boldsymbol{X}'\boldsymbol{X})^{-1} \right\|.$$

Proposition 2 provides an explicit upper bound on ω under which the bias remains smaller than a one-standard-error benchmark and the variance is weakly reduced, offering a conservative "safe region" for practical implementation.

3 Posterior Simulation

The previous section shows how LLM-generated observations enter the econometric model as a weighted dummy-observation prior and how the dummy weight ω governs the resulting bias–variance tradeoff. We now describe how we simulate from the posterior distribution induced by this augmented system.

The likelihood function for the augmented system is

$$p(\boldsymbol{y}_a \mid \boldsymbol{\beta}, \sigma^2, \omega) \propto (\sigma^2)^{-(n+r)/2} \exp\Big\{ -\tfrac{1}{2\sigma^2}(\boldsymbol{y}_a - \boldsymbol{X}_a\boldsymbol{\beta})'(\boldsymbol{y}_a - \boldsymbol{X}_a\boldsymbol{\beta}) \Big\}, \tag{2}$$

which is simply the Gaussian likelihood for the stacked data.

Conjugate conditionals (given ω). Given the scale–invariant, flat, prior $p(\boldsymbol{\beta}, \sigma^2) \propto \sigma^{-2}$, let $S(\boldsymbol{\beta}) = (\boldsymbol{y}_a - \boldsymbol{X}_a\boldsymbol{\beta})'(\boldsymbol{y}_a - \boldsymbol{X}_a\boldsymbol{\beta})$ be the residual sum of squares for the augmented data, and

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}_a'\boldsymbol{X}_a)^{-1}\boldsymbol{X}_a'\boldsymbol{y}_a$$

be the posterior mean of $\boldsymbol{\beta}$ under this prior, which can be rewritten as:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X} + \omega\boldsymbol{X}_d'\boldsymbol{X}_d)^{-1}(\boldsymbol{X}'\boldsymbol{y} + \omega\boldsymbol{X}_d'\boldsymbol{y}_d)$$

Conditional on σ^2, the posterior for $\boldsymbol{\beta}$ is multivariate normal:

$$\boldsymbol{\beta} \mid \sigma^2, \omega \sim \mathcal{N}\big(\hat{\boldsymbol{\beta}},\, \sigma^2(\boldsymbol{X}_a'\boldsymbol{X}_a)^{-1}\big). \tag{3}$$

Conditional on $\boldsymbol{\beta}$, the variance σ^2 follows an inverse–gamma distribution:

$$\sigma^2 \mid \boldsymbol{\beta}, \omega \sim \text{Inv-Gamma}\Big(\tfrac{n+r}{2},\, \tfrac{S(\boldsymbol{\beta})}{2}\Big), \tag{4}$$

with shape parameter $(n + r)/2$ and scale proportional to the augmented residual sum of squares.

Posterior simulation with unknown ω. To allow the dummy-observation weight ω to be learned from the data while preventing the synthetic block from exerting implausibly strong influence, we embed the model in a *Metropolis–within–Gibbs* algorithm with an explicit

guardrail on the admissible range of ω. Proposition 2 implies a conservative safe region

$$0 \le \omega \le \bar{\omega}, \qquad \bar{\omega} = \frac{\sigma\sqrt{A}}{C_X\, r\, \delta_1}, \qquad A = \lambda_{\min}(\boldsymbol{X}'\boldsymbol{X}).$$

In implementation, we estimate the discrepancy term in the theoretical bound by comparing the conditional mean functions implied by the dummy-only and real-data-only posterior means. Specifically, let $\bar{\beta}_d$ and $\bar{\beta}_r$ denote the posterior mean coefficient vectors from separate flat-prior Gaussian estimations using, respectively, only the synthetic dummy block and only the real-data subsample. We set

$$\hat{\delta}_1 = \max_{i \in \mathcal{X}_{\text{eval}}} \left| x_i'\bar{\beta}_d - x_i'\bar{\beta}_r \right|,$$

where $\mathcal{X}_{\text{eval}}$ contains the covariate–treatment profiles observed in the real and synthetic design matrices. This quantity estimates the largest discrepancy between the synthetic and real conditional mean functions over the empirical support used for estimation. Let $\hat{\sigma}$ denote the posterior estimate of the residual standard deviation from the real-data-only model, we then set

$$\omega_{\max} = \frac{\hat{\sigma}\sqrt{A}}{C_X\, r\, \hat{\delta}_1},$$

and restrict prior support to $(0, \omega_{\max}]$. Thus, the cap remains anchored in the safe-region logic of Proposition 2: it rules out values of ω for which the synthetic block could dominate the real data and make inference unduly sensitive to misspecification in the LLM-generated observations, while still allowing the data to determine the effective prior weight within the admissible region.

Metropolis–within–Gibbs update for ω. We target the marginal posterior

$$p(\omega \mid \mathcal{D}) \;\propto\; p_0(\omega)\, \mathbb{1}\{0 < \omega \le \omega_{\max}\}\, p(\boldsymbol{y}, \boldsymbol{y}_d \mid \boldsymbol{X}, \boldsymbol{X}_d, \omega),$$

where $\mathcal{D} = \{(\boldsymbol{y}, \boldsymbol{X}), (\boldsymbol{y}_d, \boldsymbol{X}_d)\}$ and $p_0(\omega)$ is a proper baseline prior on $(0, \omega_{\max}]$. Under the formulation,

$$\boldsymbol{X}_a(\omega) = \begin{bmatrix} \boldsymbol{X} \\ \sqrt{\omega}\, \boldsymbol{X}_d \end{bmatrix}, \qquad \boldsymbol{y}_a(\omega) = \begin{bmatrix} \boldsymbol{y} \\ \sqrt{\omega}\, \boldsymbol{y}_d \end{bmatrix},$$

define

$$\boldsymbol{A}(\omega) = \boldsymbol{X}'\boldsymbol{X} + \omega\, \boldsymbol{X}_d'\boldsymbol{X}_d, \qquad \boldsymbol{b}(\omega) = \boldsymbol{X}'\boldsymbol{y} + \omega\, \boldsymbol{X}_d'\boldsymbol{y}_d, \qquad c(\omega) = \boldsymbol{y}'\boldsymbol{y} + \omega\, \boldsymbol{y}_d'\boldsymbol{y}_d.$$

Let

$$\widehat{\boldsymbol{\beta}}(\omega) = \boldsymbol{A}(\omega)^{-1}\boldsymbol{b}(\omega) \qquad \text{and} \qquad \mathrm{SSE}(\omega) = c(\omega) - \boldsymbol{b}(\omega)'\boldsymbol{A}(\omega)^{-1}\boldsymbol{b}(\omega),$$

so that $\mathrm{SSE}(\omega)$ is the augmented residual sum of squares evaluated at the least-squares solution.

In the empirical implementation, we use a Gamma baseline prior,

$$p_0(\omega) \propto \omega^{a_\omega - 1} \exp(-b_\omega \omega),$$

truncated to the admissible interval $(0, \omega_{\max}]$. Our baseline sets $a_\omega = 2$ and $b_\omega = 0.1$.

Normalization for learning ω. As emphasized by Sims (2005), the dummy-observation construction yields the correct posterior kernel for $(\boldsymbol{\beta}, \sigma^2)$ conditional on a fixed ω, but posterior learning about ω itself requires retaining all ω-dependent normalizing terms. A convenient interpretation is that the dummy block supplies observations with precision proportional to ω:

$$\boldsymbol{y}_d \mid \boldsymbol{\beta}, \sigma^2, \omega \sim \mathcal{N}(\boldsymbol{X}_d\boldsymbol{\beta},\ \sigma^2\omega^{-1}\boldsymbol{I}_r).$$

Hence the dummy likelihood contributes the factor $\omega^{r/2}$. With the scale-invariant prior $p(\boldsymbol{\beta}, \sigma^2) \propto \sigma^{-2}$, the marginal likelihood obtained by integrating out $(\boldsymbol{\beta}, \sigma^2)$ is, up to ω-independent constants,

$$p(\boldsymbol{y}, \boldsymbol{y}_d \mid \boldsymbol{X}, \boldsymbol{X}_d, \omega) \ \propto\ \omega^{r/2}\, |\boldsymbol{A}(\omega)|^{-1/2} \big(\mathrm{SSE}(\omega)\big)^{-(n+r-k)/2}. \tag{5}$$

Accordingly, the log marginal likelihood is

$$\ell(\omega) = \frac{r}{2}\log\omega - \frac{1}{2}\log|\boldsymbol{A}(\omega)| - \frac{n+r-k}{2}\log\big(\mathrm{SSE}(\omega)\big) \quad + \text{const.}$$

Proposal and acceptance step. We update ω on the log scale, $\eta = \log\omega$, to preserve positivity:

$$\eta^\star = \eta + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, s_\eta^2), \qquad \omega^\star = \exp(\eta^\star).$$

Because the posterior support is truncated at $\omega_{\max}$, any proposal with $\omega^\star > \omega_{\max}$ is rejected automatically. For admissible proposals, since the target density is written with respect to $d\omega$ but the random-walk proposal is symmetric in $\eta = \log \omega$, the change of variables contributes the Jacobian term $\log |d\omega/d\eta| = \eta$. Hence the log acceptance ratio includes $\eta^\star - \eta$. The acceptance probability is therefore

$$\alpha = \begin{cases} \min\Big\{1,\ \exp\Big(\ell(\omega^\star) - \ell(\omega) + \log p_0(\omega^\star) - \log p_0(\omega) + \eta^\star - \eta\Big)\Big\}, & \text{if } \omega^\star \le \omega_{\max}, \\ 0, & \text{if } \omega^\star > \omega_{\max}. \end{cases} \tag{6}$$

Equivalently, one may view the cap as part of the prior support, in which case $\log p(\omega^\star) = -\infty$ outside $(0, \omega_{\max}]$.

Interpretation of the guardrail. This restriction preserves the desirable feature that ω is learned from the data, but only within a region where the finite-sample theory guarantees that the synthetic block remains a disciplined source of prior information rather than a substitute for the real sample. In that sense, the Metropolis step selects the effective prior weight endogenously, while the cap ensures that posterior learning remains anchored to the credibility of the human data.

Metropolis–within–Gibbs algorithm. Each iteration $m = 1, \ldots, M$ proceeds as follows:

1. **Gibbs step:** draw $\boldsymbol{\beta}^{(m)} \mid \sigma^{2(m-1)}, \omega^{(m-1)}, \mathcal{D}$ from the Gaussian conditional in (3).
2. **Gibbs step:** draw $\sigma^{2(m)} \mid \boldsymbol{\beta}^{(m)}, \omega^{(m-1)}, \mathcal{D}$ from the inverse-gamma conditional in (4).
3. **Metropolis step:** propose $\omega^\star$ via the random walk above and set $\omega^{(m)} = \omega^\star$ with probability α in (6); otherwise set $\omega^{(m)} = \omega^{(m-1)}$.

The resulting draws $\{\boldsymbol{\beta}^{(m)}, \sigma^{2(m)}, \omega^{(m)}\}$ jointly characterize the posterior distribution while ensuring that the learned weight remains inside the theoretically admissible region.

In the empirical implementation, we initialize the chain at a positive value truncated to the admissible interval and compute posterior summaries after discarding an initial burn-in period. We assess convergence using trace and recursive mean plots for ω and the treatment-effect coefficients, together with the Metropolis acceptance rate.

The present implementation is based on a homoskedastic Gaussian working likelihood. This choice is convenient but not essential. The synthetic block enters through the prior information matrix and can be combined with alternative Bayesian likelihoods or posterior adjustments that allow for heteroskedasticity, clustered or serial dependence, or outliers. Existing Bayesian interpretations of heteroskedasticity-consistent and cluster-robust covariance estimators (Poirier, 2011; Szpiro, Rice and Lumley, 2010; Startz, 2012; Lewis, 2022), as well as flexible heteroskedastic or heavy-tailed likelihoods (Norets, 2015; West, 1984; Lange, Little and Taylor, 1989; Geweke, 1993), provide natural starting points for such extensions.

4 Generating Synthetic Observations

The posterior simulation algorithm developed above treats the synthetic dummy block $(\boldsymbol{y}_d, \boldsymbol{X}_d)$ as given. This section describes a general protocol for constructing that block in a way that is portable across experimental settings. The goal is to generate synthetic observations that are aligned with the real experiment: they contain the same treatment indicators, covariates, and outcomes as the real data. The resulting block is an elicited prior object, expressed in the same data format as the real experiment and subsequently disciplined by the dummy-observation weight ω.

The protocol is as follows. The researcher first defines the relevant support of the experimental design. She then translates this support into textual profiles and exposes each profile (i.e. persona) to the experimental protocol. Finally, she records the resulting responses and appends the resulting observations as the dummy block $(\boldsymbol{y}_d, \boldsymbol{X}_d)$. While the implementation is especially direct in information-provision experiments, it can be adapted to other experimental settings.

Define the support of the design. We start by identifying the study-relevant characteristics for the economic agents subject to the experiment. For households they typically include age, gender, education, and household income, among others. To ensure tractability, we aggregate to a reasonable number of categories for each characteristic (for example, age: <25, 25–64, and 65+). We then construct the target population cells, which correspond to the unique combinations of the chosen characteristics. The resulting grid determines the support points at which the prior is elicited by constructing the personas for which the LLM simulates responses.

Translate cells into profiles and expose them to the experimental protocol. For each population cell we create a brief persona in the spirit of Zarifhonarvar (2026) and Hansen et al. (2024), which is operationalized as a textual description of the specific economic agent's characteristics. This description is embedded into a prompt and incorporated into the experiment's sequence. In the setting of randomized information-provision experiments, the sequence corresponds to the participant's experimental journey: pre-treatment belief elicitation, randomized information provision, and post-treatment questions. For each persona–treatment combination we generate at least one synthetic observation in the information-provision experiment.

Preserve the experimental protocol. A central feature of the construction is that the elicitation inherits the design of the real-world experiment. The wording of the questions, the treatment scripts, the order of elicitation and the outcome scales are fixed by the experimental protocol. Researcher discretion enters primarily through the definition of the population cells and the translation of their values into concise persona descriptions. In our framework, these descriptions simply assign values to the covariates used to define the cells, rather than adding narrative detail. The LLM is therefore not asked a different version of the experimental question; it is exposed to the same protocol under explicitly specified covariate values.

Example To fix ideas, we present an illustration of the LLM elicitation strategy below. The words in blue correspond to the persona characteristics that are programmatically varied to capture all population cells needed for the experiment. The set of characteristics can be easily adapted and extended to meet the needs of the experimental design.

1. Persona definition and pre-treatment elicitation:

> "You are a female American respondent, between 40 and 64 years old, living in the Southern United States. You have high household income ($100,000 or more). You have your own views about the economy based on your background and experiences. Today is June 10th, 2018.
>
> By how much do you expect prices in the economy (as measured by the Consumer Price Index) to change over the next 12 months?
>
> Please provide an answer as a percentage change from current prices. Please use a

percent between -100 and 100, if no change please enter a 0. Answer as this specific person would, taking into account their demographic background, likely economic knowledge, and personal circumstances."

After recording the pre-treatment elicitation step, the LLM is prompted again with the assigned treatment (here in red) and the post-treatment question(s). The synthetic data generation pipeline varies the treatment to expose every population cell to every treatment in the experiment.

2. Treatment and post-treatment elicitation

"Before you continue, we would like you to know the following: The U.S. Federal Open Market Committee (which sets short-term interest rates) forecasts 1.9% inflation rate in 2018.

By how much do you expect prices in the economy (as measured by the Consumer Price Index) to change over the next 12 months?

Please provide an answer as a percentage change from current prices. Please use a percent between -100 and 100, if no change please enter a 0. Answer as this specific person would, taking into account their demographic background, likely economic knowledge, and personal circumstances."

Graphical and other non-textual treatments. The same protocol applies when the treatment, or more generally the object presented to the subject, is not purely textual. Experimental designs may present subjects with graphs, images, videos, or other types of information; they may also involve multi-step or interactive tasks. In such cases, the synthetic elicitation should preserve the object faced by the human subject rather than replace it with a verbal summary. This keeps the synthetic observation tied to the original experimental design and avoids introducing an additional layer of researcher interpretation.

Figure 1 illustrates this point for the forecasting experiment of Afrouzi et al. (2023). In that experiment, subjects observe a graphical history of the stochastic process when reporting their forecasts. We therefore embed the same graphical representation in the LLM prompt. After each synthetic forecast is recorded, the history is updated according to the experimental protocol, so that the next prompt contains the updated graph.

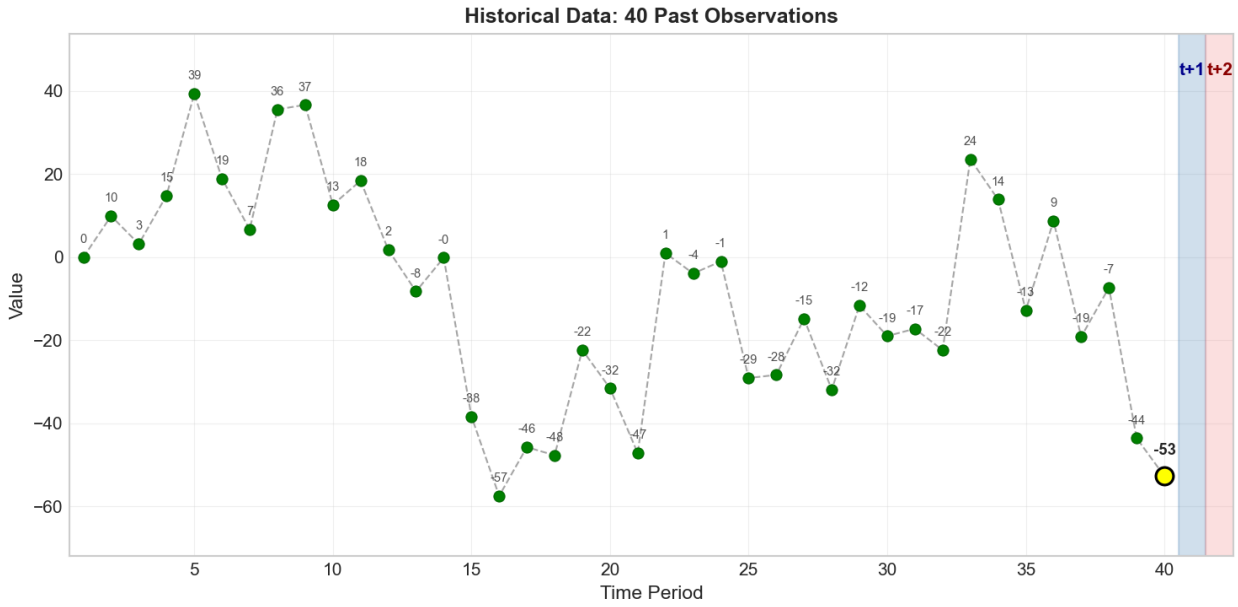


Figure 1: Visual Treatment

Notes: The figure illustrates how the treatment from Afrouzi et al. (2023) is incorporated into the LLM prompt used to generate synthetic observations. At each iteration, the LLM makes one- and two-step-ahead predictions. The synthetic-observation pipeline records these predictions and then updates the figure by adding the newly observed data point, thereby expanding the graphical observation window over forecasting rounds.

5 Empirical Results

We evaluate the dummy-observation prior in two complementary applications. The first is the household inflation-expectations experiment of Coibion, Gorodnichenko and Weber (2022), a randomized information-provision experiment in which the target parameters are treatment effects on revisions in 12-month inflation expectations. The second is the forecasting experiment of Afrouzi et al. (2023), a controlled environment in which participants forecast stochastic AR(1) processes and the target parameter is the forecast-implied persistence of expectations. In both applications, we construct deliberately low-power real-data samples, augment them with a fixed block of LLM-generated synthetic dummy observations, and compare posterior uncertainty from the real-data-only estimator with posterior uncertainty from the mixed estimator. The dummy-observation weight ω is re-estimated for each real-data subsample under the safe-region cap described above, so the reported gains reflect repeated implementation of the full procedure rather than a single favorable subsample draw.

5.1 Inflation-Expectations Experiment

We begin with the household inflation-expectations experiment of Coibion, Gorodnichenko and Weber (2022). The analysis focuses on the specification with demographic and socioeconomic controls, corresponding to the controls used in the original empirical specification.

Table 1 reports the main precision results. The table is based on 100 repeated random real-data subsamples of size $n = 500$. Across treatment effects, the real-data-only posterior standard deviation is about 2.63 on average, whereas the mixed posterior standard deviation falls to about 1.72. This corresponds to an average posterior-SD reduction of 34.69 percent. Expressed in effective-sample-size units, achieving the same reduction by increasing the real sample alone would require approximately 134.69 percent more real observations. Put differently, in this design the mixed estimator with 500 real observations attains a level of precision comparable to a real-data-only analysis with more than twice as many observations. Appendix B examines the sensitivity of these results to the effective information content assigned to the LLM-generated block. The gains decline sharply as the synthetic block is discounted, indicating that the baseline precision improvements depend on the informativeness of the LLM prior rather than on synthetic rows mechanically tightening the posterior.

The precision gains are highly stable across treatment arms. Posterior-SD reductions range from 34.51 percent for the FOMC-statement treatment to 34.89 percent for the USA Today treatment, and effective-sample-size gains range from 133.43 percent to 136.13 percent. The standard deviations across repeated subsamples are small relative to the mean gains. Thus, the improvement is not driven by one particular treatment arm or by one favorable subsample draw; it is a systematic consequence of adding the LLM-generated dummy-observation prior.

Figure 2 examines how these gains vary with the number of real observations. As expected from the bias–variance logic, the value of the dummy-observation prior is largest when the real-data sample is smallest. With 250 real observations, the mean posterior-SD reduction is close to 50 percent, corresponding to an effective-sample-size gain of almost 300 percent. At the benchmark sample size of 500, the reduction is about 35 percent and the effective-sample-size gain is about 135 percent, consistent with Table 1. The gains decline monotonically as the real sample grows: they remain sizeable at 1,000 observations, become smaller at 2,500 observations, and are modest at 5,000 observations. Extending the stress test to 10,000 and 20,000 real observations reinforces the same pattern. Even at 20,000 observations, the

Table 1: Precision Gains: Inflation-Expectations Experiment

Treatment	Real SD	Mixed SD	SD reduction (%)	Eff. n gain (%)
Population growth	2.61	1.71	34.56	133.74
	(0.16)	(0.10)	(1.13)	(8.10)
Unemployment rate	2.65	1.73	34.86	135.88
	(0.21)	(0.13)	(1.16)	(8.44)
Gasoline prices	2.63	1.72	34.61	134.11
	(0.18)	(0.12)	(1.17)	(8.43)
Past inflation	2.64	1.72	34.83	135.64
	(0.19)	(0.12)	(1.10)	(8.04)
Inflation target	2.62	1.71	34.69	134.67
	(0.19)	(0.13)	(1.21)	(8.49)
Fed inflation forecast	2.63	1.72	34.58	133.93
	(0.18)	(0.11)	(1.29)	(9.28)
FOMC statement	2.61	1.71	34.51	133.43
	(0.18)	(0.12)	(1.34)	(9.63)
USA Today	2.63	1.71	34.89	136.13
	(0.17)	(0.12)	(1.23)	(8.93)
Mean across treatments	2.63	1.72	34.69	134.69
	(0.18)	(0.12)	(1.20)	(8.67)

Notes: The table is based on the specification of Coibion, Gorodnichenko and Weber (2022) with demographic and socioeconomic controls. Entries report means of the corresponding statistics across 100 repeated real-data subsamples of size 500. Standard deviations across repeated subsamples are reported in parantheses below the means. SD reduction is computed within each repeated subsample as $100 \times (1 - SD_{\text{mixed}}/SD_{\text{real}})$; effective-sample-size gain is computed within each repeated subsample as $100 \times ((SD_{\text{real}}/SD_{\text{mixed}})^2 - 1)$.

estimated precision gain remains positive, but it is close to zero, indicating that the synthetic dummy block adds little once the real-data likelihood is already highly informative. This is precisely what the framework predicts: synthetic prior information is most valuable when real experimental evidence is scarce, while its marginal contribution naturally diminishes as the real-data likelihood becomes more informative.

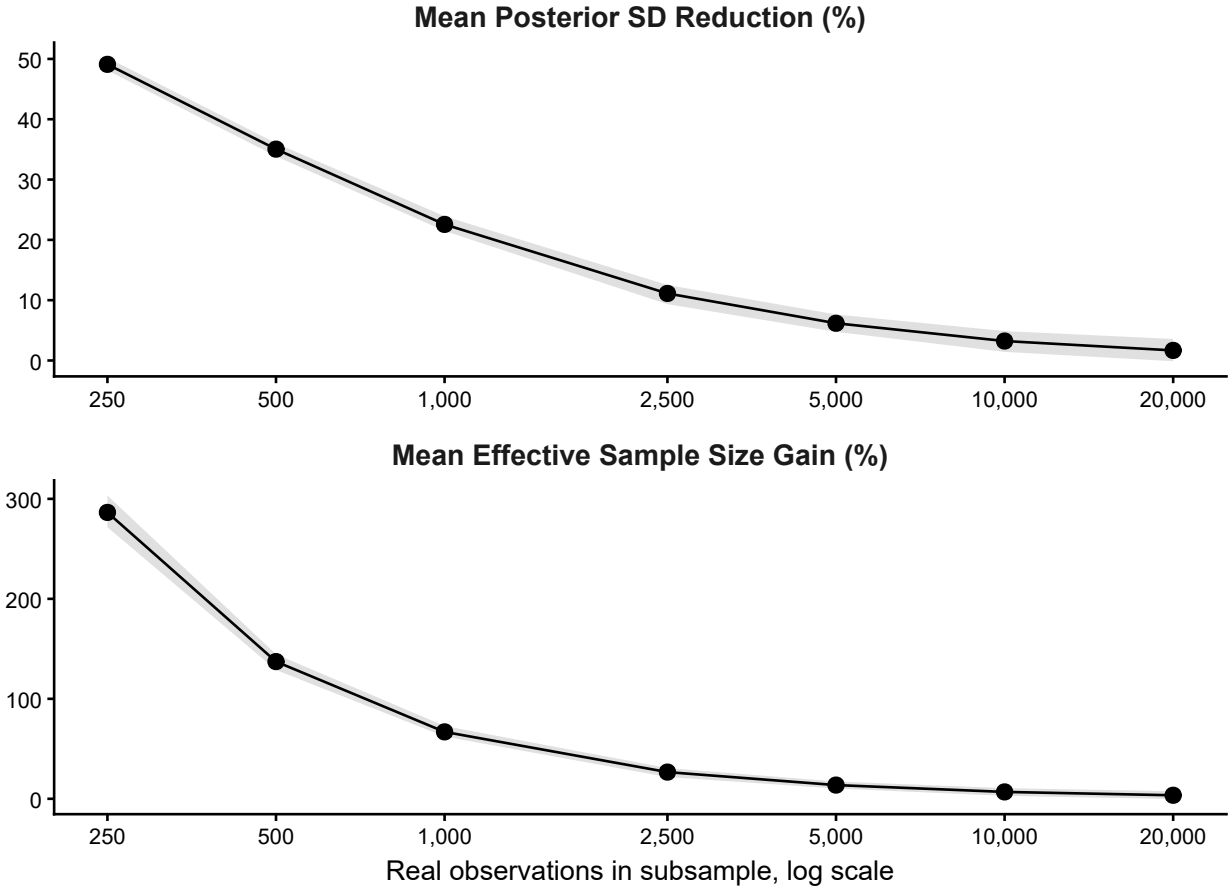


Figure 2: Stress Test: Precision Gains in the Inflation-Expectations Experiment

Notes: The figure is based on the specification of Coibion, Gorodnichenko and Weber (2022) with demographic and socioeconomic controls. The upper panel reports the mean reduction in posterior standard deviations from adding LLM-generated dummy observations, relative to the real-data-only benchmark. The lower panel reports the corresponding effective-sample-size gain. Lines denote means across repeated random subsamples; shaded areas denote the central 90 percent range across subsamples, defined by the 5th and 95th percentiles. The analysis holds the synthetic dummy block fixed and re-estimates the dummy-observation weight ω for each subsample.

Figure 3 provides complementary evidence on how the synthetic prior interacts with the real data. The LLM-generated dummy-observation prior is relatively concentrated, reflecting

the structured information encoded in the synthetic block. After combining the prior with a randomly drawn real-data subsample of size $n = 500$, the mixed posterior distributions are substantially more dispersed and, in several cases, shifted away from the prior. This shows that the procedure does not simply replace real observations with LLM-generated observations. Instead, the synthetic block acts as an informative prior anchor, while the real-data likelihood remains central in determining posterior location and uncertainty.

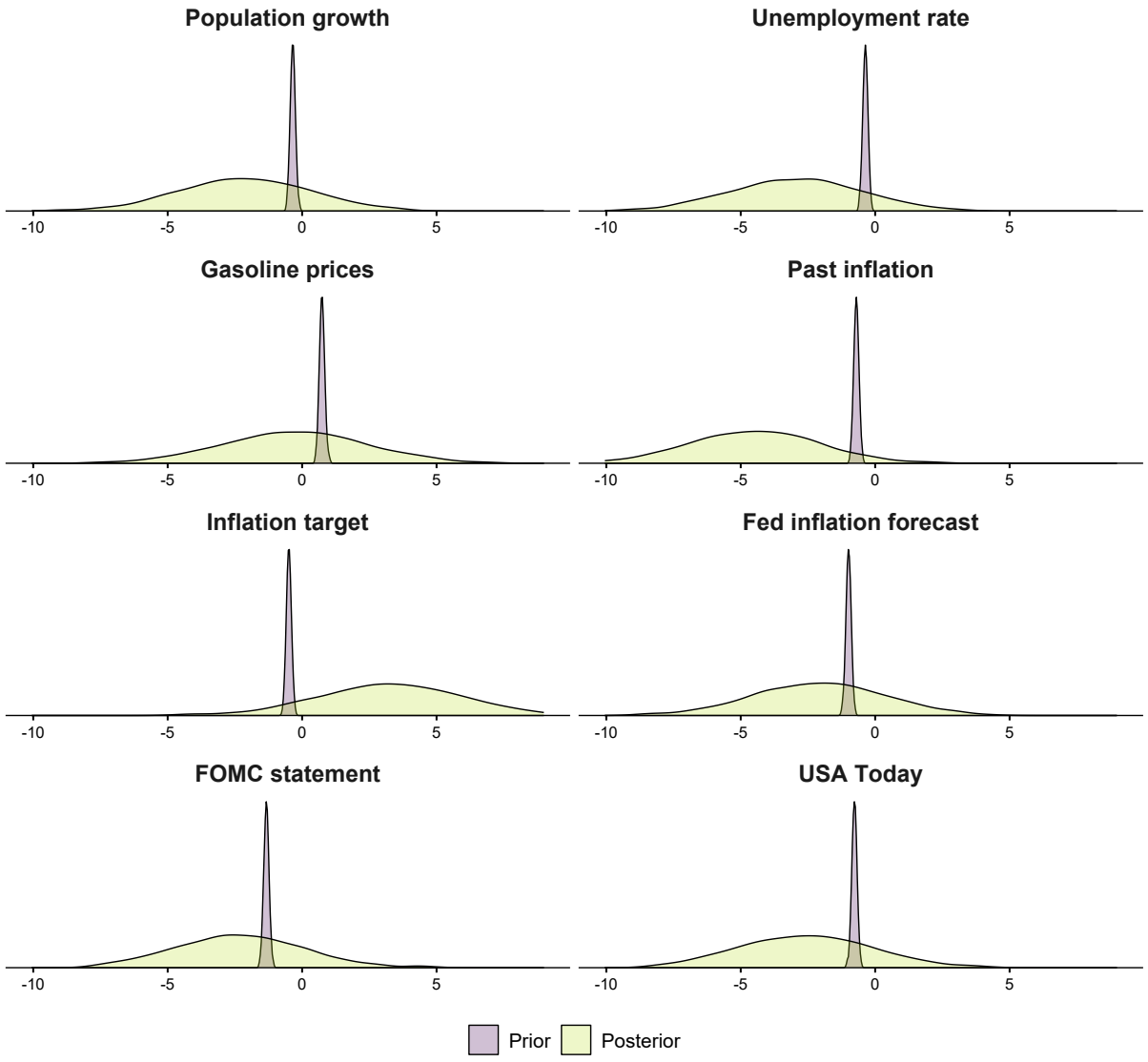


Figure 3: Prior and Posterior Distributions: Inflation-Expectations Experiment

Notes: The figure is based on the specification of Coibion, Gorodnichenko and Weber (2022) with demographic and socioeconomic controls. Each panel shows the distribution of one treatment-effect coefficient under the LLM-generated synthetic dummy-observation prior and under the posterior obtained after combining this prior with one randomly drawn real-data subsample of size $n = 500$. The synthetic dummy block is held fixed, and the dummy-observation weight ω is re-estimated for the displayed subsample. Lines are kernel density estimates based on retained MCMC draws. The vertical axis uses a square-root scale to make the relatively concentrated prior distributions and the more dispersed posterior distributions visually comparable.

To summarize, the results show that LLM-generated dummy observations can deliver large and stable precision gains in a conventional information-provision experiment. At the

same time, the posterior distributions remain responsive to the real data, and the stress-test results show that the gains decline smoothly as the amount of real evidence increases. The findings therefore align with the theoretical interpretation of the method as a disciplined bias–variance tradeoff: the synthetic block improves precision, but its influence is learned subject to an explicit safe-region bound, keeping inference anchored in real experimental observations.

5.2 Forecasting Experiment

The second application uses the forecasting experiment of Afrouzi et al. (2023). This setting differs from the previous experiment in two useful ways. First, the data-generating process is known: participants forecast realizations of AR(1) processes with true persistence ρ. Second, the empirical object is not a treatment effect, but the forecast-implied persistence of subjective expectations. For each true persistence cell, we estimate the slope from regressing the reported forecast on the currently observed state and report the posterior standard deviation of the resulting per-period forecast-implied persistence. This makes the experiment a useful robustness check for the dummy-observation prior: the method is applied to a continuous forecasting object in a tightly controlled environment rather than to treatment effects in a survey information experiment.

The low-power design retains all real participants but restricts each participant to 10 usable forecast rounds. This short-panel design mimics a setting in which the researcher has access to the full cross-section of forecasters but limited repeated observations per participant. For each repeated round subsample, the mixed estimator combines the real observations with the LLM-generated synthetic forecast panels, and the dummy-observation weight ω is learned subject to the same safe-region logic used in the previous application.

Table 2 summarizes the precision gains by true AR(1) persistence. The results show substantial improvements across the entire persistence grid. Averaging over ρ cells, the posterior standard deviation of the forecast-implied persistence falls from 0.04 in the real-data-only specification to 0.03 in the mixed specification. This corresponds to an average posterior-SD reduction of 33.66 percent and an effective-sample-size gain of 128.84 percent. The gains are present for all persistence values: posterior-SD reductions range from 29.47 percent for the random-walk process ($\rho = 1$) to 37.62 percent for the highly persistent but stationary process ($\rho = 0.8$). Thus, the LLM dummy-observation prior improves precision not only in

the most transitory cells, where forecasts are especially noisy, but also in high-persistence environments where forecast-implied persistence is already estimated more tightly.

Table 2: Precision Gains: Forecasting Experiment

True ρ	Real SD	Mixed SD	SD reduction (%)	Eff. n gain (%)
$\rho = 0$	0.06	0.04	36.60	148.93
	(0.00)	(0.00)	(0.79)	(6.20)
$\rho = 0.2$	0.06	0.04	35.82	142.87
	(0.00)	(0.00)	(0.71)	(5.35)
$\rho = 0.4$	0.05	0.04	31.49	113.17
	(0.00)	(0.00)	(0.90)	(5.65)
$\rho = 0.6$	0.05	0.03	30.94	109.82
	(0.00)	(0.00)	(0.93)	(5.73)
$\rho = 0.8$	0.03	0.02	37.62	157.12
	(0.00)	(0.00)	(0.77)	(6.33)
$\rho = 1$	0.01	0.01	29.47	101.11
	(0.00)	(0.00)	(0.88)	(5.05)
Mean across ρ cells	0.04	0.03	33.66	128.84
	(0.00)	(0.00)	(0.83)	(5.72)

Notes: The table reports precision gains for the forecasting experiment of Afrouzi et al. (2023). The low-power design keeps all real participants but samples a limited number of usable forecast rounds per participant. Entries report means across 100 repeated real-data round subsamples with 10 usable forecast rounds per participant. Standard deviations across repeated subsamples are in parentheses. Real SD and Mixed SD denote posterior standard deviations of the per-period forecast-implied persistence. SD reduction is $100 \times (1 - SD_{\text{mixed}}/SD_{\text{real}})$. Effective-sample-size gain is $100 \times ((SD_{\text{real}}/SD_{\text{mixed}})^2 - 1)$.

Figure 4 studies how these gains vary with the number of usable forecast rounds per participant. The figure mirrors the stress-test logic in the inflation-expectations application, but the relevant dimension of real-data information is now the length of the participant-level forecasting panel rather than the number of cross-sectional survey observations. The pattern is again consistent with the bias–variance interpretation of the method. When only a few rounds per participant are available, the real-data likelihood is weak and the synthetic dummy block provides substantial precision gains. As the number of usable rounds increases, the real-data-only estimator becomes more informative and the marginal contribution of the synthetic prior declines. The fact that the gains decrease with panel length is important: it indicates that the dummy-observation prior is most valuable precisely in the low-information designs for which it is intended, rather than mechanically reducing posterior uncertainty by

a fixed amount.

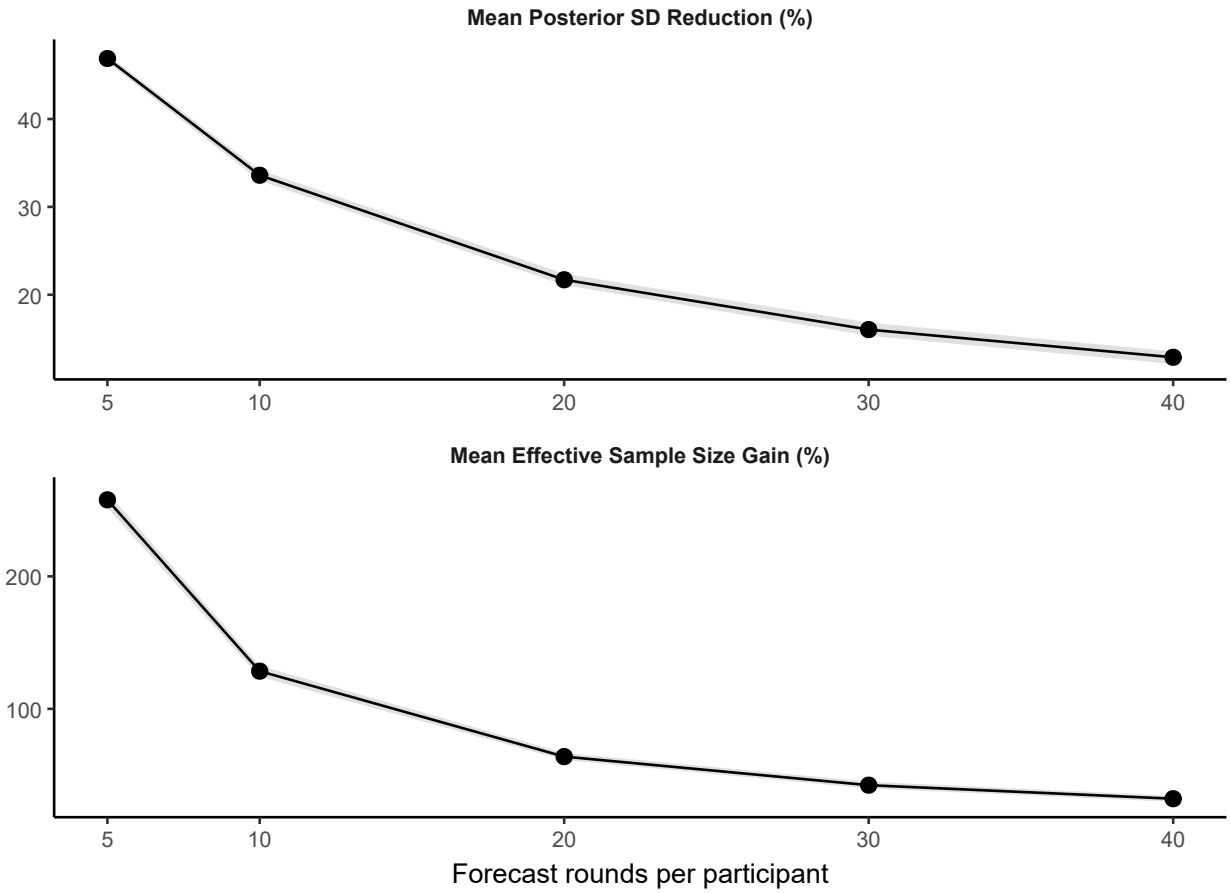


Figure 4: Stress Test: Precision Gains in the Forecasting Experiment

Notes: The figure is based on the forecasting experiment of Afrouzi et al. (2023). The stress test varies the number of usable forecast rounds per participant while retaining all real participants. The upper panel reports the mean reduction in posterior standard deviations from adding LLM-generated dummy observations, relative to the real-data-only benchmark. The lower panel reports the corresponding effective-sample-size gain. Precision gains are computed for the per-period forecast-implied persistence and averaged across true AR(1) persistence values ρ. Lines denote means across repeated random round subsamples; shaded areas denote the central 90 percent range across subsamples, defined by the 5th and 95th percentiles. The analysis holds the synthetic dummy block fixed and re-estimates the dummy-observation weight ω for each subsample.

Figure 5 shows the prior and posterior distributions of forecast-implied persistence by true persistence cell. The figure illustrates that the LLM-generated synthetic block provides a structured prior over the persistence response, but the posterior after combining it with real forecast rounds is not identical to the prior. The mixed posterior remains broader and can shift relative to the LLM-induced distribution, reflecting information from the real

participant forecasts. This is particularly useful in the forecasting setting because the target object has a clear behavioral interpretation: posterior mass above the 45-degree rational benchmark corresponds to overreaction in expectations. The figure therefore shows not only that the synthetic block improves precision, but also that the augmented posterior preserves the empirical content of the real forecasting data.

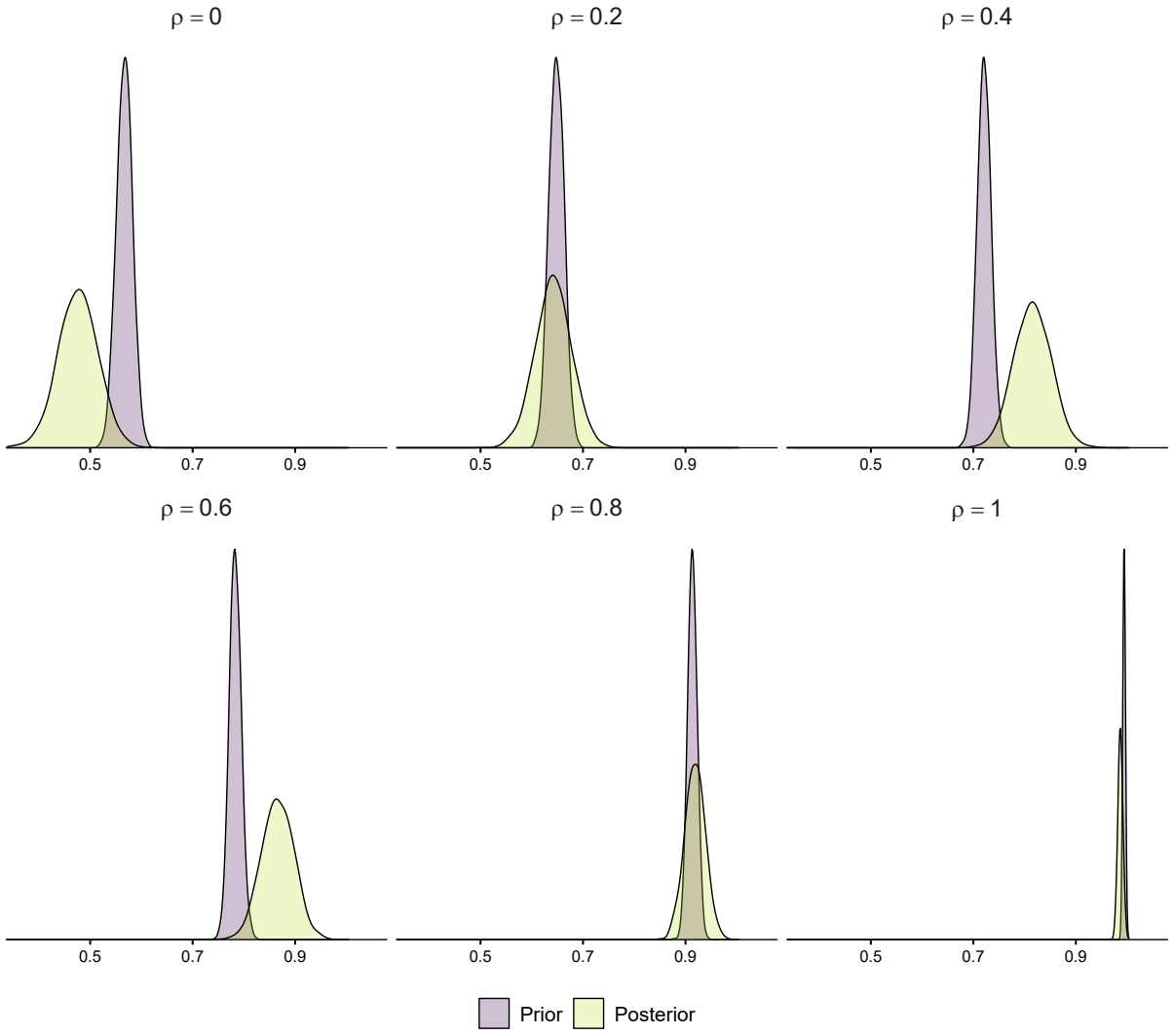


Figure 5: Prior and Posterior Distributions: Forecasting Experiment

Notes: The figure is based on the forecasting experiment of Afrouzi et al. (2023). Each panel shows the distribution of the per-period forecast-implied persistence for a given true AR(1) persistence ρ. The prior distribution is induced by the LLM-generated synthetic dummy-observation block, while the posterior distribution is obtained after combining this prior with one randomly drawn low-power real-data round subsample. The low-power design keeps all real participants but uses only 10 usable forecast rounds per participant. Lines are kernel density estimates based on retained MCMC draws.

The forecasting experiment therefore reinforces the main empirical message from the inflation-expectations application. LLM-generated dummy observations improve precision in an underpowered design, the gains are stable across economically meaningful cells, and

the benefits decline as more real information becomes available. The fact that the method performs similarly for treatment-effect estimates in the inflation-expectations experiment and for forecast-implied persistence in the Afrouzi et al. experiment suggests that the dummy-observation prior is not tied to a particular estimand. Instead, it provides a general way to encode experiment-specific prior information while preserving the central role of real observations in posterior inference.

6 Conclusion

This paper develops a framework for using LLM-generated synthetic observations as prior information in randomized experiments. The central distinction is between using LLMs as substitutes for participants and using them as structured, weighted priors for experimentally identified estimands. Our approach takes the second route. By representing synthetic observations as dummy observations, the framework makes the content of the prior explicit, allows its influence to be learned from the data, and restricts that influence through a finite-sample safe region. The resulting estimator formalizes a disciplined bias–variance tradeoff: synthetic observations can improve precision when real experimental evidence is scarce, while their influence naturally declines as the real-data likelihood becomes more informative.

Across two applications, this logic delivers sizeable precision gains in low-power designs while keeping identification anchored in real observations. The inflation-expectations application shows that LLM-informed priors can improve treatment-effect estimation in a standard information-provision experiment with multiple treatment arms. The forecasting application shows that the same framework can improve inference for a different estimand, forecast-implied persistence, in a controlled experimental environment. These results suggest that the method is not tied to a particular outcome, estimand, or experimental design. More broadly, LLMs need not be viewed either as invalid synthetic subjects or as replacements for costly experiments. They can instead serve as a flexible source of bounded, inspectable, and empirically disciplined prior information.

References

Abberger, Klaus, Arbërim Bibaj, Hans Gersbach, Alexis Perakis, Alexander Rathke, Samad Sarferaz, and Kieran James Walsh. 2026. "Tariffs and Firm Expectations." CESifo Working Paper 12510.

Afrouzi, Hassan, Spencer Y Kwon, Augustin Landier, Yueran Ma, and David Thesmar. 2023. "Overreaction in Expectations: Evidence and Theory*." *The Quarterly Journal of Economics*, 138(3): 1713–1764.

Andre, Peter, Carlo Pizzinelli, Christopher Roth, and Johannes Wohlfart. 2022. "Subjective models of the macroeconomy: Evidence from experts and representative samples." *Review of Economic Studies*, 89(6): 2958–2991.

Andre, Peter, Ingar Haaland, Christopher Roth, Mirko Wiederholt, and Johannes Wohlfart. 2026. "Narratives about the Macroeconomy." *The Review of Economic Studies*.

Ash, Elliott, and Stephen Hansen. 2023. "Text Algorithms in Economics." *Annual Review of Economics*, 15: 659–688.

Ash, Elliott, Heiner Mikosch, Alexis Perakis, and Samad Sarferaz. 2026. "Seeing and hearing is believing: The role of audiovisual communication in shaping inflation expectations." *European Economic Review*, 184: 105253.

Ash, Elliott, Stephen Hansen, and Yabra Muvdi. 2024. "Large Language Models in Economics." CEPR Discussion Paper No. 19479.

Azevedo, Eduardo M., Alex Deng, José L. Montiel Olea, and E. Glen Weyl. 2019. "Empirical Bayes Estimation of Treatment Effects with Many A/B Tests: An Overview." *AEA Papers and Proceedings*, 109: 43–47.

Bańbura, Marta, Domenico Giannone, and Lucrezia Reichlin. 2010. "Large Bayesian Vector Auto Regressions." *Journal of Applied Econometrics*, 25(1): 71–92.

Bandiera, Oriana, Greg Fischer, Andrea Prat, and Erina Ytsma. 2021. "Do Women Respond Less to Performance Pay? Building Evidence from Multiple Experiments." *American Economic Review: Insights*, 3(4): 435–54.

Brynjolfsson, Erik, José Ramón Enríquez, Sophia Kazinnik, and David Nguyen. 2025. "Augmenting Survey Data with Generative AI: An Application to Economic Research." *SSRN Working Paper*, 6343598.

Bybee, J Leland. 2023. "The ghost in the machine: Generating beliefs with large language models." *arXiv preprint arXiv:2305.02823*, 349–389.

Coey, Dominic, and Tom Cunningham. 2019. "Improving Treatment Effect Estimators Through Experiment Splitting." *WWW '19*, 285–295. New York, NY, USA:Association for Computing Machinery.

Coibion, Olivier, and Yuriy Gorodnichenko. 2026. *Expectations matter: The new causal macroeconomics of surveys and experiments.* Princeton University Press.

Coibion, Olivier, Yuriy Gorodnichenko, and Michael Weber. 2022. "Monetary policy communications and their effects on household inflation expectations." *Journal of Political Economy*, 130(6): 1537–1584.

Cook, Thomas R., Zach Modig, Sophia Kazinnik, and Nathan Palmer. 2025. "What Do LLMs Want?" *SSRN Working Paper*, 5812143.

Crosta, Tommaso, Dean Karlan, Finley Ong, Julius Rüschenpöhler, and Christopher R. Udry. 2024. "Unconditional Cash Transfers: A Bayesian Meta-Analysis of Randomized Evaluations in Low and Middle Income Countries." *NBER Working Paper*, 32779. National Bureau of Economic Research.

DellaVigna, Stefano, and Devin Pope. 2018. "Predicting Experimental Results: Who Knows What?" *Journal of Political Economy*, 126(6): 2410–2456.

DellaVigna, Stefano, Nicholas Otis, and Eva Vivalt. 2020. "Forecasting the Results of Experiments: Piloting an Elicitation Strategy." *AEA Papers and Proceedings*, 110: 75–79.

Del Negro, Marco, and Frank Schorfheide. 2004. "Priors from General Equilibrium Models for VARs." *International Economic Review*, 45(2): 643–673.

Dibiasi, Andreas, Heiner Mikosch, and Samad Sarferaz. 2025. "Uncertainty Shocks, Adjustment Costs, and Firm Beliefs: Evidence from a Representative Survey." *American Economic Journal: Macroeconomics*, 17(3): 36–73.

Doan, Thomas, Robert B. Litterman, and Christopher A. Sims. 1984. "Forecasting and Conditional Projection Using Realistic Prior Distributions." *Econometric Reviews*, 3(1): 1–100.

Drechsel, Dirk, Heiner Mikosch, Samad Sarferaz, and Matthias Bannert. 2015. "How are firms affected by exchange rate shocks? Evidence from survey based impulse responses." *KOF Working Papers*, 371. KOF Swiss Economic Institute, ETH Zurich.

Efron, Bradley. 2010. *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction. Institute of Mathematical Statistics Monographs*, Cambridge:Cambridge University Press.

Fuster, Andreas, and Basit Zafar. 2023. "Survey experiments on economic expectations." In *Handbook of Economic Expectations*. 107–130. Academic Press.

Gelman, Andrew, John B. Carlin, Hal S. Stern, and Donald B. Rubin. 1995. "Bayesian Data Analysis." Chapman and Hall/CRC.

Geweke, John. 1993. "Bayesian Treatment of the Independent Student-t Linear Model." *Journal of Applied Econometrics*, 8(S1): S19–S40.

Giannone, Domenico, Michele Lenza, and Giorgio E. Primiceri. 2015. "Prior Selection for Vector Autoregressions." *The Review of Economics and Statistics*, 97(2): 436–451.

Gui, George, and Seungwoo Kim. 2025. "Leveraging LLMs to Improve Experimental Design: A Generative Stratification Approach." *SSRN Working Paper*, 5549285.

Haaland, Ingar, Christopher Roth, and Johannes Wohlfart. 2023. "Designing Information Provision Experiments." *Journal of Economic Literature*, 61(1): 3–40.

Hansen, Anne Lundgaard, John J. Horton, Sophia Kazinnik, Daniela Puzzello, and Ali Zarifhonarvar. 2024. "Simulating the Survey of Professional Forecasters." *SSRN Working Paper*, 5066286.

Horton, John J., Apostolos Filippas, and Benjamin S. Manning. 2023. "Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?" *NBER Working Paper*, 31122. National Bureau of Economic Research.

Iacovone, Leonardo, David McKenzie, and Rachael Meager. 2025. "Bayesian Impact Evaluation With Informative Priors: An Application to a Colombian Management and Export Improvement Program." *Econometrica*, 93(5): 1915–1935.

Imbens, Guido W., and Donald B. Rubin. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction.* Cambridge:Cambridge University Press.

Kazinnik, Sophia, and Tara M. Sinclair. 2025. "FOMC *In Silico*: A Multi-Agent System for Monetary Policy Decision Modeling." *SSRN Working Paper*, 5424097.

Korinek, Anton. 2023. "Generative AI for Economic Research: Use Cases and Implications for Economists." *Journal of Economic Literature*, 61(4): 1281–1317.

Korinek, Anton. 2025. "AI Agents for Economic Research." *NBER Working Paper*, 34202. National Bureau of Economic Research.

Lange, Kenneth L., Roderick J. A. Little, and Jeremy M. G. Taylor. 1989. "Robust Statistical Modeling Using the t Distribution." *Journal of the American Statistical Association*, 84(408): 881–896.

Lewis, Gabriel. 2022. "Heteroskedasticity and Clustered Covariances from a Bayesian Perspective." University Library of Munich, Germany MPRA Paper 116662.

Link, Sebastian, Andreas Peichl, Christopher Roth, and Johannes Wohlfart. 2023. "Information frictions among firms and households." *Journal of Monetary Economics*, 135: 99–115.

Litterman, Robert B. 1986. "Forecasting with Bayesian Vector Autoregressions — Five Years of Experience." *Journal of Business & Economic Statistics*, 4(1): 25–38.

Ludwig, Jens, Sendhil Mullainathan, and Ashesh Rambachan. 2026. "Large Language Models: An Applied Econometric Framework." *Annual Review Economics*, 18.

Manning, Benjamin S., and John J. Horton. 2026. "General Social Agents." *NBER Working Paper*, 34937. National Bureau of Economic Research.

Manning, Benjamin S., Kehang Zhu, and John J. Horton. 2024. "Automated Social Science: Language Models as Scientist and Subjects." *NBER Working Paper*, 32381. National Bureau of Economic Research.

Meager, Rachael. 2019. "Understanding the Average Impact of Microcredit Expansions: A Bayesian Hierarchical Analysis of Seven Randomized Experiments." *American Economic Journal: Applied Economics*, 11(1): 57–91.

Meager, Rachael. 2022. "Aggregating Distributional Treatment Effects: A Bayesian Hierarchical Analysis of the Microcredit Literature." *American Economic Review*, 112(6): 1818–1847.

Mikosch, Heiner, Christopher Roth, Samad Sarferaz, and Johannes Wohlfart. 2024. "Uncertainty and information acquisition: Evidence from firms and households." *American Economic Journal: Macroeconomics*, 16(2): 375–405.

Norets, Andriy. 2015. "Bayesian Regression with Nonparametric Heteroskedasticity." *Journal of Econometrics*, 185(2): 409–419.

Poirier, Dale J. 2011. "Bayesian Interpretations of Heteroskedastic Consistent Covariance Estimators Using the Informed Bayesian Bootstrap." *Econometric Reviews*, 30(4): 457–468.

Robbins, Herbert. 1964. "The Empirical Bayes Approach to Statistical Decision Problems." *The Annals of Mathematical Statistics*, 35(1): 1–20.

Sims, Christopher A. 2005. "Dummy Observation Priors Revisited." Princeton University Manuscript. Available at Christopher A. Sims's webpage: `https://sims.princeton.edu/yftp/DummyObs/DumObsPrior.pdf`; accessed May 6, 2026.

Spiegelhalter, David J., Laurence S. Freedman, and Mahesh K. B. Parmar. 1994. "Bayesian Approaches to Randomized Trials." *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 157(3): 357–416.

Stantcheva, Stefanie. 2023. "How to Run Surveys: A Guide to Creating Your Own Identifying Variation and Revealing the Invisible." *Annual Review of Economics*, 15: 205–234.

Startz, Richard. 2012. "Bayesian Heteroskedasticity-Robust Standard Errors." Department of Economics, University of California, Santa Barbara Economics Working Paper Series qt69c4x8m9.

Szpiro, Adam A., Kenneth M. Rice, and Thomas Lumley. 2010. "Model-Robust Regression and a Bayesian "Sandwich" Estimator." *The Annals of Applied Statistics*, 4(4): 2099–2113.

Theil, Henri. 1963. "On the Use of Incomplete Prior Information in Regression Analysis." *Journal of the American Statistical Association*, 58(302): 401–414.

Theil, Henri, and Arthur S. Goldberger. 1961. "On Pure and Mixed Statistical Estimation in Economics." *International Economic Review*, 2(1): 65–78.

Vivalt, Eva. 2020. "How Much Can We Generalize From Impact Evaluations?" *Journal of the European Economic Association*, 18(6): 3045–3089.

Weber, Michael, Bernardo Candia, Hassan Afrouzi, Tiziano Ropele, Rodrigo Lluberas, Serafin Frache, Brent Meyer, Saten Kumar, Yuriy Gorodnichenko, Dimitris Georgarakos, Olivier Coibion, Geoff Kenny, and Jorge Ponce. 2025. "Tell Me Something I Don't Already Know: Learning in Low- and High-Inflation Settings." *Econometrica*, 93(1): 229–264.

West, Mike. 1984. "Outlier Models and Prior Distributions in Bayesian Linear Regression." *Journal of the Royal Statistical Society: Series B (Methodological)*, 46(3): 431–439.

Wu, Jing Cynthia, Jin Xi, and Shihan Xie. 2025. "LLM Survey Framework: Coverage, Reasoning, Dynamics, Identification." *NBER Working Paper*, 34308. National Bureau of Economic Research.

Zarifhonarvar, Ali. 2026. "Generating inflation expectations with large language models." *Journal of Monetary Economics*, 157: 103859.

A Proofs for Section 2

A.1 Proof of Proposition 1

Proof. We work in the Gaussian linear model described in Section 2,

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

and augment it with ω-weighted dummy observations $(\boldsymbol{y}_d, \boldsymbol{X}_d)$ to obtain the stacked system

$$\boldsymbol{X}_a = \begin{bmatrix} \boldsymbol{X} \\ \sqrt{\omega}\, \boldsymbol{X}_d \end{bmatrix}, \qquad \boldsymbol{y}_a = \begin{bmatrix} \boldsymbol{y} \\ \sqrt{\omega}\, \boldsymbol{y}_d \end{bmatrix}.$$

The mixed estimator is the OLS estimator on the augmented system,

$$\hat{\beta}_{\text{mix}} = (\boldsymbol{X}_a' \boldsymbol{X}_a)^{-1} \boldsymbol{X}_a' \boldsymbol{y}_a.$$

Bias. Write the real outcomes as $\boldsymbol{y} = \boldsymbol{X}\beta + \varepsilon$. For dummy observations we write

$$\boldsymbol{y}_d = \boldsymbol{X}_d \beta + \eta_d,$$

where η_d captures the systematic deviation of the LLM-generated conditional means from the truth.

Assumption 1 bounds the difference between the LLM-generated and true conditional means at any covariate–treatment profile (x, d) in the support of the real data. Because dummy observations are generated exactly at such profiles $(X_{d,i}, D_{d,i})$, we have

$$\left| E[Y_{d,i} \mid X_{d,i}, D_{d,i}] - E[Y \mid X_{d,i}, D_{d,i}] \right| \leq \delta_1.$$

Using the decomposition $Y_{d,i} = X_{d,i}'\beta + \eta_{d,i}$ and the fact that $E[Y \mid X_{d,i}, D_{d,i}] = X_{d,i}'\beta$ under the true DGP, this implies

$$|E[\eta_{d,i} \mid X_{d,i}, D_{d,i}]| \leq \delta_1 \qquad \text{for all dummy rows } i.$$

Assumption 1 together with bounded regressors further ensures $|X_{d,ij}| \leq C_X$ for some constant C_X.

Substituting $\boldsymbol{y} = \boldsymbol{X}\beta + \varepsilon$ and $\boldsymbol{y}_d = \boldsymbol{X}_d\beta + \eta_d$ into $\hat{\beta}_{\text{mix}}$ and taking expectations gives

$$E[\hat{\beta}_{\text{mix}}] = (\boldsymbol{X}'\boldsymbol{X}+\omega\boldsymbol{X}_d'\boldsymbol{X}_d)^{-1}\big(\boldsymbol{X}'\boldsymbol{X}\beta+\omega\boldsymbol{X}_d'\boldsymbol{X}_d\beta+\omega\boldsymbol{X}_d'E[\eta_d]\big) = \beta+(\boldsymbol{X}'\boldsymbol{X}+\omega\boldsymbol{X}_d'\boldsymbol{X}_d)^{-1}\big(\omega\boldsymbol{X}_d'E[\eta_d]\big).$$

Thus,

$$E[\hat{\beta}_{\text{mix}}] - \beta = (\boldsymbol{X}'\boldsymbol{X} + \omega\boldsymbol{X}_d'\boldsymbol{X}_d)^{-1}\big(\omega\boldsymbol{X}_d'E[\eta_d]\big).$$

We now bound the vector $\boldsymbol{X}_d'E[\eta_d]$. Its jth element is

$$(\boldsymbol{X}_d'E[\eta_d])_j = \sum_{i=1}^{r} X_{d,ij}\, E[\eta_{d,i}].$$

Each term satisfies

$$|X_{d,ij}E[\eta_{d,i}]| \le C_X\delta_1.$$

Summing over $i = 1, \ldots, r$ gives

$$|(\boldsymbol{X}_d'E[\eta_d])_j| \le r\, C_X\delta_1.$$

Thus every coordinate of the $k \times 1$ vector $\boldsymbol{X}_d'E[\eta_d]$ is bounded in magnitude by $rC_X\delta_1$, so the Euclidean norm satisfies

$$\|\boldsymbol{X}_d'E[\eta_d]\| \le \sqrt{k}\, C_X\, r\delta_1.$$

Next, because $(\boldsymbol{X}'\boldsymbol{X})$ and $(\boldsymbol{X}_d'\boldsymbol{X}_d)$ are positive definite with smallest eigenvalues at least $\lambda_{\min}(\boldsymbol{X}'\boldsymbol{X})$ and $\lambda_{\min}(\boldsymbol{X}_d'\boldsymbol{X}_d)$, their sum satisfies

$$\lambda_{\min}(\boldsymbol{X}'\boldsymbol{X} + \omega\boldsymbol{X}_d'\boldsymbol{X}_d) \ge \lambda_{\min}(\boldsymbol{X}'\boldsymbol{X}) + \omega\, \lambda_{\min}(\boldsymbol{X}_d'\boldsymbol{X}_d).$$

Therefore,

$$\big\|(\boldsymbol{X}'\boldsymbol{X} + \omega\boldsymbol{X}_d'\boldsymbol{X}_d)^{-1}\big\| \le \frac{1}{\lambda_{\min}(\boldsymbol{X}'\boldsymbol{X}) + \omega\, \lambda_{\min}(\boldsymbol{X}_d'\boldsymbol{X}_d)}.$$

Combining with the bound above yields

$$\|E[\hat{\beta}_{\text{mix}}] - \beta\| \le \frac{\omega\, \sqrt{k}\, C_X\, r\delta_1}{\lambda_{\min}(\boldsymbol{X}'\boldsymbol{X}) + \omega\, \lambda_{\min}(\boldsymbol{X}_d'\boldsymbol{X}_d)}.$$

This shows explicitly how the bias depends on ω, r, and δ_1.

Variance. Write the augmented regression as

$$\hat{\beta}_{\text{mix}} = (\boldsymbol{X}_a'\boldsymbol{X}_a)^{-1}\boldsymbol{X}_a'\varepsilon_a, \qquad \varepsilon_a = \begin{bmatrix} \varepsilon \\ \sqrt{\omega}\,\varepsilon_d \end{bmatrix},$$

where $\varepsilon_d = \boldsymbol{y}_d - E[\boldsymbol{y}_d \mid \boldsymbol{X}_d]$. Assumption 1 implies that the conditional variances of the real and dummy blocks are uniformly bounded. For notational simplicity, we therefore upper-bound the conditional variance of the augmented error vector by a common constant and write, conditioning on the design matrices,

$$\text{Var}(\varepsilon_a \mid \boldsymbol{X}, \boldsymbol{X}_d) \;\preceq\; \sigma^2 \boldsymbol{I}_{n+r}.$$

Allowing explicitly for the variance discrepancy in Assumption 1 would replace σ^2 by $\sigma^2 + \delta_2$ (or, more generally, by a uniform upper bound $\bar{\sigma}^2$) throughout the variance and MSE bounds without affecting their dependence on ω, n, and r.

Hence

$$\text{Var}(\hat{\beta}_{\text{mix}}) = (\boldsymbol{X}_a'\boldsymbol{X}_a)^{-1}\boldsymbol{X}_a'\text{Var}(\varepsilon_a)\,\boldsymbol{X}_a(\boldsymbol{X}_a'\boldsymbol{X}_a)^{-1} = \sigma^2(\boldsymbol{X}_a'\boldsymbol{X}_a)^{-1}.$$

As above, $\lambda_{\min}(\boldsymbol{X}_a'\boldsymbol{X}_a) = \lambda_{\min}(\boldsymbol{X}'\boldsymbol{X}) + \omega\lambda_{\min}(\boldsymbol{X}_d'\boldsymbol{X}_d)$, so the largest eigenvalue of the variance matrix satisfies

$$\|\text{Var}(\hat{\beta}_{\text{mix}})\| \le \frac{\sigma^2}{\lambda_{\min}(\boldsymbol{X}'\boldsymbol{X}) + \omega\lambda_{\min}(\boldsymbol{X}_d'\boldsymbol{X}_d)}.$$

Thus the posterior variance shrinks at the rate dictated by the combined information from n real and ωr synthetic observations.

Mean-squared error. The mean-squared error is

$$\text{MSE}(\hat{\beta}_{\text{mix}}) = \|E[\hat{\beta}_{\text{mix}}] - \beta\|^2 + \text{trace}\big(\text{Var}(\hat{\beta}_{\text{mix}})\big).$$

Using the two bounds derived above, the first term is at most

$$\left(\frac{\omega\,\sqrt{k}\,C_X\,r\,\delta_1}{\lambda_{\min}(\boldsymbol{X}'\boldsymbol{X}) + \omega\,\lambda_{\min}(\boldsymbol{X}_d'\boldsymbol{X}_d)}\right)^2,$$

and the second term is at most

$$\frac{k\sigma^2}{\lambda_{\min}(\boldsymbol{X}'\boldsymbol{X}) + \omega\,\lambda_{\min}(\boldsymbol{X}_d'\boldsymbol{X}_d)}.$$

For the real-data estimator $\hat{\beta}_{\text{real}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$, the bias is zero and

$$\text{MSE}(\hat{\beta}_{\text{real}}) = \text{trace}\big(\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\big) \le \frac{k\sigma^2}{\lambda_{\min}(\boldsymbol{X}'\boldsymbol{X})}.$$

Comparing the two expressions shows that the mixed estimator has smaller MSE than the real-data estimator whenever

$$\frac{\omega^2\,(C_X r\delta_1)^2}{\big(\lambda_{\min}(\boldsymbol{X}'\boldsymbol{X}) + \omega\lambda_{\min}(\boldsymbol{X}_d'\boldsymbol{X}_d)\big)^2} + \frac{k\sigma^2}{\lambda_{\min}(\boldsymbol{X}'\boldsymbol{X}) + \omega\lambda_{\min}(\boldsymbol{X}_d'\boldsymbol{X}_d)} \;<\; \frac{k\sigma^2}{\lambda_{\min}(\boldsymbol{X}'\boldsymbol{X})}.$$

Rearranging yields the comparison condition stated in the text: for moderate values of ω, the variance reduction from the additional dummy information outweighs the bias introduced by misspecification of order δ_1, whereas for very large ω the estimator approaches the dummy-only estimator and the bias dominates.

This completes the proof. □

A.2 Proof of Proposition 2

Proof. From Proposition 1, the bias norm is bounded by

$$\big\|E[\hat{\beta}_{\text{mix}}] - \beta\big\| \le \frac{\omega\,\sqrt{k}\,C_X\,r\,\delta_1}{\lambda_{\min}(\boldsymbol{X}'\boldsymbol{X}) + \omega\,\lambda_{\min}(\boldsymbol{X}_d'\boldsymbol{X}_d)}.$$

Since $\lambda_{\min}(\boldsymbol{X}'\boldsymbol{X}) + \omega\,\lambda_{\min}(\boldsymbol{X}_d'\boldsymbol{X}_d) \ge A$, we have

$$\big\|E[\hat{\beta}_{\text{mix}}] - \beta\big\| \le \frac{\omega\,\sqrt{k}\,C_X\,r\,\delta_1}{A}.$$

For any $\omega \le \bar{\omega} = \sigma\sqrt{A}/(C_X r\delta_1)$,

$$\frac{\omega\,C_X\,r\,\delta_1}{A} \;\le\; \frac{\sigma\sqrt{A}}{C_X r\delta_1}\cdot\frac{C_X r\delta_1}{A} = \frac{\sigma}{\sqrt{A}}.$$

Hence

$$\left\|E[\hat{\beta}_{\text{mix}}] - \beta\right\| \leq \sqrt{k}\,\frac{\sigma}{\sqrt{A}} = \sqrt{k}\,\sigma\,\left\|(\boldsymbol{X}'\boldsymbol{X})^{-1/2}\right\| \leq \sqrt{k}\,\sigma,$$

where the last inequality uses that the operator norm of $(\boldsymbol{X}'\boldsymbol{X})^{-1/2}$ is at most $1/\sqrt{A}$.

For the variance, Proposition 1 gives

$$\left\|\text{Var}(\hat{\beta}_{\text{mix}})\right\| \leq \frac{\sigma^2}{\lambda_{\min}(\boldsymbol{X}'\boldsymbol{X}) + \omega\,\lambda_{\min}(\boldsymbol{X}_d'\boldsymbol{X}_d)} \leq \frac{\sigma^2}{A} = \left\|\text{Var}(\hat{\beta}_{\text{real}})\right\|,$$

since $\lambda_{\min}(\boldsymbol{X}'\boldsymbol{X}) + \omega\,\lambda_{\min}(\boldsymbol{X}_d'\boldsymbol{X}_d) \geq A$. This proves the result. □

B Sensitivity to the Information Content of LLM-Generated Observations

We examine how the precision gains change when the LLM-generated dummy block is treated as containing less reliable information than in the baseline specification. Let $\lambda \geq 1$ denote the relative imprecision of the synthetic block and define $\phi = 1/\lambda$. The baseline corresponds to $\lambda = 1$, while larger values of λ discount the effective information content of the LLM-generated observations. The synthetic rows themselves are unchanged; instead, their contribution to posterior updating is scaled by ϕ. In the Gaussian updating equations, this scales the dummy-block cross-products and residual sum of squares by ϕ, and the variance update uses $n + \phi r$ effective information units. This exercise therefore asks how much precision remains when the LLM-generated block is treated as lower-quality prior information.

When the LLM-generated block is discounted substantially, the mixed posterior approaches the real-data-only benchmark. This confirms that the method's precision gains are driven by the effective information content of the synthetic prior rather than by the mere presence of synthetic rows.

Table 3: Sensitivity to LLM Information Content

λ	Real SD	Mixed SD	SD reduction (%)	Eff. n gain (%)
1.00	2.63	1.72	34.69	134.69
2.00	2.63	2.03	22.64	67.24
4.00	2.63	2.28	13.40	33.47
9.00	2.63	2.46	6.52	14.55
16.00	2.63	2.53	3.88	8.34

Entries summarize means across 100 repeated real-data subsamples of size 500. λ is the relative imprecision parameter. The corresponding information-content discount is $\phi = 1/\lambda$. The synthetic rows themselves are unchanged. Larger values of λ reduce the effective information content of the synthetic block. The dummy-block cross-products and residual sum of squares enter with factor ϕ, and the variance update uses $n + \phi r$ effective observations. SD reduction is computed as $100 \times (1 - SD_{mixed}/SD_{real})$. Effective-sample-size gain is computed as $100 \times ((SD_{real}/SD_{mixed})^2 - 1)$.