The Safe-Tail Paradox:

Stress Testing AI Exposure of Banks' Borrowers*

Christophe Hurlin† Christophe Pérignon‡

May 27, 2026

Abstract

We document a Safe-Tail Paradox in banks' credit portfolios: retail borrowers classified as safest by scoring models are also the most exposed to artificial intelligence (AI)-related labor income risk. The paradox arises because AI exposure is positively correlated with borrower characteristics historically associated with low default risk (e.g., stable employment, high income), while AI exposure can weaken repayment capacity through displacement and wage compression. Credit risk therefore becomes concentrated in the safest segments of mortgage portfolios, precisely where regulatory capital buffers are thinnest. We design a borrower-level AI stress test and apply it to a synthetic portfolio calibrated to the French residential mortgage market. As AI adoption intensifies, capital requirements rise sixfold more in the safest class than in the riskiest, highlighting the need for AI-aware risk management.

Keywords: generative artificial intelligence, stress testing, retail credit risk, mortgage lending, prudential regulation.

JEL Classification: G21, G28, J24, O33.

*We gratefully acknowledge financial support from the Institut Universitaire de France (IUF) and from the Deloitte Chair on AI for Business Innovation.

†University of Orléans and Institut Universitaire de France, Rue de Blois, 45067 Orléans, France. E-mail: christophe.hurlin@univ-orleans.fr

‡HEC Paris, 1 Rue de la Libération, 78350 Jouy-en-Josas, France. E-mail: perignon@hec.fr

1 Introduction

Generative artificial intelligence (AI) is reshaping the landscape of cognitive work at a pace that has no historical precedent.[1] Unlike previous waves of automation, which displaced routine, manual tasks and spared high-skill workers, AI targets precisely the cognitive tasks that command wage premia: writing, coding, and reasoning. The emerging empirical evidence suggests that the labor market effects are already visible, concentrated among young, high-wage workers in exposed occupations, and largely invisible in aggregate statistics (Eloundou et al., 2024; Brynjolfsson et al., 2025; Humlum and Vestergaard, 2025; Massenkoff and McCrory, 2026). What distinguishes this wave from prior technological transitions is not its speed but its scope: for the first time, automation reaches into the core tasks of high-skill, well-paid occupations that have historically been sheltered from substitution.

Yet no study has examined the consequences of these potential labor market disruptions for the creditworthiness of bank borrowers. This paper asks a simple question: what happens to the credit risk of a mortgage portfolio when a technological shock hits borrowers along occupational lines? We focus on residential mortgages, which constitute the core of banks' retail exposure. The mechanism we identify is general, but we anchor the analysis in the French market, where the mortgage stock exceeded €1,284 billion in January 2026, representing nearly 48 percent of GDP (Fédération Bancaire Française, 2026). The sensitivity of mortgage default to income disruptions is well established: the *Haut Conseil de Stabilité Financière* (HCSF) estimates that a 3-percentage-point rise in unemployment among borrowers raises the long-run default rate by roughly 9 percent (HCSF, 2019).[2] The AI-driven disruptions now emerging in the data operate through the same income channel, but target a very different population of workers.

The key to understanding the mechanism lies in an empirical regularity: AI exposure is positively correlated with wages. Software developers, financial analysts, and lawyers are highly exposed by this measure, and precisely because they earn high incomes, are employed under stable contracts, and have historically exhibited low default rates, scoring models assign them to the safest rating classes.

[1] Throughout this paper, *generative AI* refers to large language models (LLMs) and related systems capable of producing text, code, and structured outputs from natural language instructions, exemplified by models such as GPT-4, Claude, and Gemini.

[2] The HCSF is the French macroprudential authority responsible for overseeing the financial system as a whole, with the aim of preserving its stability.

This gives rise to what we term the *Safe-Tail Paradox*. Banks assess mortgage default risk through scoring models that rely on income, employment status, and leverage, variables that reliably predicted default in a stable labor market. Occupational task content is absent from these models by construction: a borrower's AI exposure is neither recorded in credit files nor priced into default probabilities. Because income is positively correlated with AI exposure at origination and negatively correlated with past estimated default risk, this creates a systematic concentration of high-exposure borrowers in the safest rating classes, those with the lowest calibrated default probabilities and the thinnest regulatory capital buffers.

Standard stress-testing frameworks are not designed to detect this pattern. Existing supervisory exercises (such as the EBA stress tests or the Federal Reserve's DFAST) apply aggregate macroeconomic shocks and infer borrower-level impacts through satellite models: a coherent architecture for macro-financial risks, but one that suppresses the cross-sectional heterogeneity of AI exposure by construction. Banking supervisory authorities have begun to recognize this limitation: in April 2026, the Bank of England's Deputy Governor for Financial Stability called for stress testing and scenario analysis to better capture AI-related risks, echoing concerns raised by the Financial Stability Oversight Council in its 2025 annual report.

We respond to this agenda by designing and implementing an AI stress test whose architecture differs from existing exercises in one essential respect: the shock is a vector of individual AI exposures e_i, one per borrower, calibrated to the task content of each occupation rather than to aggregate macro-financial variables. We define AI exposure at the occupational level as the share of tasks that can be performed by AI systems, based on task-level measures of technological capability and observed usage (Massenkoff and McCrory, 2026). This exposure can be set to zero to represent the pre-ChatGPT baseline implicit in current underwriting standards, to the level currently observed in practice (e_i^{obs}), or to the full automation potential of each occupation (e_i^{pot}). The scenario severity is parameterized by $\delta \in [0, 1]$, which interpolates between the observed and potential exposure levels: at $\delta = 0$, the portfolio is repositioned at the exposure level already observable in 2026, a non-trivial shock for any bank whose underwriting implicitly assumes $e_i = 0$; at $\delta = 1$, the full automation frontier is reached. This granular shock feeds into satellite equations governing employment displacement and wage compression, from which post-shock default probabilities and regulatory capital shortfalls are

derived for each rating class.

We implement this framework on a synthetic portfolio calibrated to reproduce the main features of the French residential mortgage market. A synthetic portfolio is the appropriate instrument here for two reasons. First, no public dataset simultaneously contains borrower-level income, contract type, and occupational task content. Second, and more fundamentally, identifying the Safe-Tail Paradox requires independent variation in income and AI exposure across occupations, a condition that real portfolios cannot satisfy because these variables are endogenously correlated through occupational sorting.

Our main findings confirm the paradox and quantify its regulatory implications. Across the rating scale, AI exposure and default risk move in opposite directions: the safest borrower classes carry the highest occupational AI exposure. When the shock materializes, regulatory capital requirements rise substantially more in the highest-rated classes than in the lowest-rated ones, by a factor of approximately six at the observed exposure level. The distinguishing feature of the paradox is not the aggregate magnitude of the shortfall, but its systematic sign pattern along the rating scale.

This paper makes two contributions to the literature on credit risk and banking regulation. First, we identify the Safe-Tail Paradox as a structural feature of credit scoring systems exposed to AI-driven labor market disruptions: because AI exposure is concentrated among high-income, cognitively intensive occupations, any scoring model that uses income as a proxy for creditworthiness will systematically misallocate risk toward its safest classes. This mechanism inverts the usual relationship between measured safety and true resilience, generating hidden risk precisely where capital buffers are thinnest. Second, we provide the first stress-testing framework designed for AI-related retail credit risk, extending the literature on thematic stress tests (Parlatore and Philippon, 2025) by replacing macro aggregates with a granular, occupation-level shock. In doing so, we connect the growing literature on AI's labor market effects (Autor et al., 2003; Acemoglu and Restrepo, 2022; Eloundou et al., 2024; Brynjolfsson et al., 2025) with the stress-testing literature in financial economics (Acharya et al., 2018; Farmer et al., 2022; Leitner and Williams, 2023; Shapiro and Zeng, 2024). More broadly, our results suggest that technological shocks may not increase credit risk uniformly, but may instead invert its location within portfolios.

The remainder of the paper is organized as follows. Section 2 documents the source of the credit-risk measurement gap. Section 3 presents the model and its main implications. Section 4 describes the stress testing exercise and Section 5 presents the results. Section 6 concludes.

2 Structural Source of the Credit Risk Mismeasurement

Internal rating systems for retail mortgage portfolios are typically calibrated on observable borrower characteristics such as gross income, employment contract type, loan-to-value ratio, and the debt service-to-income ratio. These variables are effective predictors of default risk in stable labor market conditions. Their limitation emerges not from how the model is built, but from how the environment is changing: a technological shock that operates along occupational lines is invisible to any scoring system calibrated on income and leverage alone.

The transmission from AI to repayment capacity operates through labor income along two related margins. On the *employment* margin, exposure to AI increases the risk of job loss or occupational downgrading in jobs where a significant share of tasks can be automated. Brynjolfsson et al. (2025) document a 16 percent relative decline in employment in the United States among young workers in highly exposed occupations. On the *earnings* margin, the effects are heterogeneous. Workers in high-exposure occupations face downward earnings pressure as AI substitutes for their core tasks, while the effects on workers in less exposed occupations remain empirically uncertain. This pattern is consistent with the task-based framework of Acemoglu and Restrepo (2022). Using Danish administrative data, Humlum and Vestergaard (2025) find near-zero average earnings effects coexisting with heterogeneous adjustments across workers. Using longitudinal worker-level data from Germany, Engberg et al. (2025) document that this heterogeneity operates along occupational lines: professionals in knowledge-intensive occupations experience wage declines as AI exposure rises, while other workers exhibit wage gains.

Professional occupational status is typically negatively correlated with default risk and positively correlated with AI exposure. Borrowers with high income and stable employment are typically assigned to the safest rating classes, yet these characteristics are associated with occupations that are more exposed to task automation. The consequence is a systematic negative relationship between AI exposure and rating class assignment, formalized as $\mathrm{Cov}(e_i, k_i) < 0$,

where e_i denotes the individual exposure to AI and k_i the rating class assigned at origination. This implies that changes in default risk induced by AI are concentrated in the safest rating classes, reversing the usual ordering of risk across rating classes. Because these effects are not captured by aggregate indicators, standard stress testing frameworks fail to detect this concentration, leading to a misallocation of capital across the rating ladder.

3 A Model of Rating Incoherence

3.1 Setup

A bank holds a portfolio of n mortgage loans indexed by $i \in \{1, \dots, n\}$. At origination, the bank observes a vector of borrower characteristics $\mathbf{x}_i = (w_i, \mathrm{DTI}_i, \mathbf{z}_i)$, where w_i denotes gross annual income, $\mathrm{DTI}_i = L_i/w_i$ the debt service to income ratio, and $\mathbf{z}_i$ additional characteristics. Probabilities of default (PD) are assessed through a logistic scoring model,

$$\mathrm{PD}_i^{(0)} = \Lambda\big(\alpha + \beta_w w_i + \beta_{\mathrm{DTI}}\,\mathrm{DTI}_i + \beta_z^\top \mathbf{z}_i\big), \tag{1}$$

where $\alpha \in \mathbb{R}$, $\beta_w < 0$, $\beta_{\mathrm{DTI}} > 0$, and $\beta_z \in \mathbb{R}^q$ are parameters estimated from historical data, and $\Lambda(\cdot)$ denotes the logistic cumulative distribution function (CDF).

Borrowers are assigned to K rating classes based on their estimated PD using thresholds $0 < \theta_1 < \cdots < \theta_{K-1} < 1$, with class $C_k = \{i : \theta_{k-1} \le \mathrm{PD}_i^{(0)} < \theta_k\}$, so that lower-index classes correspond to safer borrowers. The class-level calibrated PD is

$$\bar{\pi}_k = \frac{1}{|C_k|} \sum_{i \in C_k} \mathrm{PD}_i^{(0)}, \tag{2}$$

which implies $\bar{\pi}_1 < \cdots < \bar{\pi}_K$ by construction. The partition $\{C_k\}$ is fixed at origination and underlies both internal risk management and regulatory capital allocation.

AI exposure blind spot. Two features of the environment are central to the analysis. First, each borrower is employed in an occupation characterized by an AI exposure $e_i \in [0, 1]$. This variable is determined by sector, task content, and qualification level, which are dimensions absent from standard credit files and therefore excluded from the scoring model. As a result, two borrowers with identical characteristics $\mathbf{x}_i$ may face substantially different AI related income risk, a source of heterogeneity that is invisible to the bank by construction.

Second, and more consequentially, this exposure is not randomly distributed across the portfolio. AI driven automation exerts its sharpest substitution pressure on high wage, cognitively intensive occupations, through both wage compression (Azar et al., 2025) and elevated displacement risk (Kochhar, 2023; Brynjolfsson et al., 2025). Borrowers with high income, those assigned by the scoring model to low risk rating classes, are therefore precisely those with the greatest unobserved exposure. The scoring model's blind spot is thus systematically concentrated where it is least expected: among borrowers rated as safe.

These two features together imply a structural misalignment between measured and actual risk. Formalizing them requires two assumptions.

Assumption 1 (Information Constraint)**.** AI exposure satisfies $e_i = g(\mathbf{x}_i) + \nu_i$, where ν_i is mean-zero with $\text{Var}(\nu_i) > 0$, and ν_i is not recoverable from standard credit file information.

Assumption 2 (AI Exposure-Income Relationship)**.** The conditional mean AI exposure $\mathbb{E}[e_i \mid w_i = w]$ is strictly increasing in income.

Assumption 1 formalizes the first feature of the environment described above. The decomposition of e_i makes precise the sense in which two observationally identical borrowers can differ in AI exposure: $g(\mathbf{x}_i)$ is the share of AI exposure that the bank could in principle infer from income and repayment variables, while ν_i is the occupation-specific remainder that escapes the credit file entirely. The condition $\text{Var}(\nu_i) > 0$ ensures that this remainder is economically meaningful. Assumption 2 captures the direction of the blind spot: unobserved risk is not merely present but is systematically higher among borrowers that the model rates as safe. Together, the two assumptions imply that the scoring model is not just incomplete, it is incomplete in a directionally biased way.

Assumption 2 is validated in Figure 1. The left panel uses the *observed exposure* index of Massenkoff and McCrory (2026), which captures the share of occupational tasks currently being delegated to AI in practice. The right panel uses the *potential exposure* index of Eloundou et al. (2024), which measures the share of tasks that LLMs could plausibly accelerate regardless of current deployment. Both indices are measured at the six-digit SOC occupation level and merged with annual mean wages from the BLS Occupational Employment and Wage Statistics survey; full data details are provided in Appendix A.

A Nadaraya–Watson estimator of $\hat{\mathbb{E}}[e_i \mid w_i]$ is strictly increasing across the wage support under both measures. The nonparametric monotonicity test of Ghosal et al. (2000) rejects the null of a non-increasing relationship in each case (bootstrap $p < 0.01$). Mean exposure rises from approximately 6% in the lowest wage decile to 27% in the top decile for the observed measure, and from 11% to 46% for the potential benchmark. This positive association between wages and AI exposure is consistent with Eloundou et al. (2024) and Felten et al. (2023), who document a significant and monotone relationship between occupational wages and LLM exposure across US occupations using different methods.

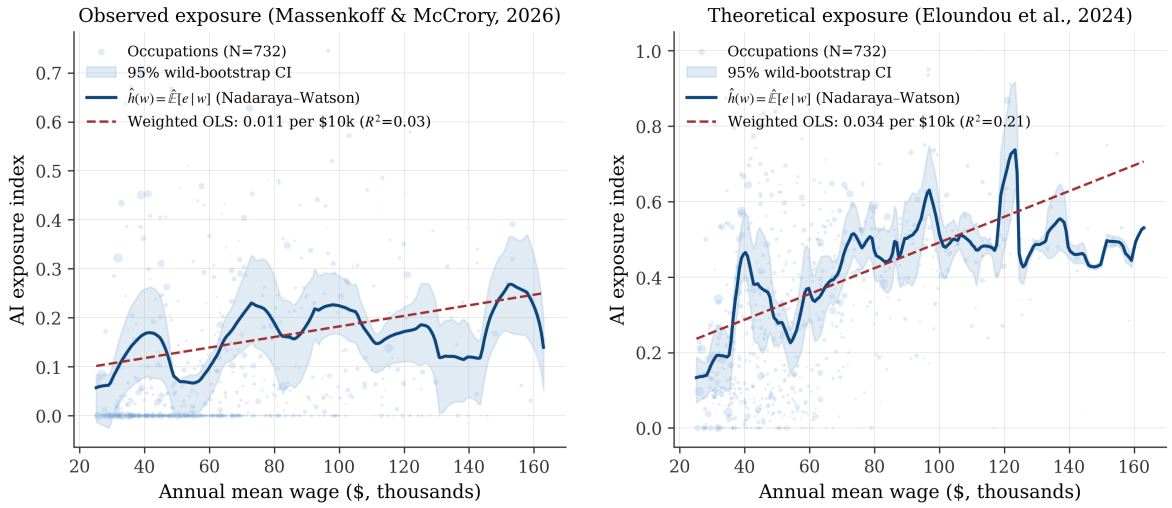

Figure 1. AI exposure and wage across occupations. This figure presents Nadaraya–Watson estimates of $\hat{h}(w) = \hat{\mathbb{E}}[e_i \mid w_i = w]$ across $m = 729$ six-digit SOC occupations. Horizontal axis: annual mean wage ($, thousands); vertical axis: AI exposure index in $[0, 1]$. Left panel: observed-exposure index of Massenkoff and McCrory (2026); right panel: theoretical β-index of Eloundou et al. (2024). Scatter points are weighted by employment. Shaded band: 95% wild-bootstrap confidence interval. Red dashed line: employment-weighted OLS fit.

3.2 Formal Characterization of the Safe-Tail Paradox

The safest borrowers in the portfolio are the most exposed to AI-driven income disruption, a property we term the *Safe-Tail Paradox*. This is not a sampling artefact: it is a structural consequence of any credit scoring model that ignores AI exposure and uses income as a proxy for creditworthiness. To formalize this claim, let the class-average AI exposure be defined as

$$\bar{e}_k \;=\; \frac{1}{|C_k|} \sum_{i \in C_k} e_i, \qquad k = 1, \ldots, K. \tag{3}$$

Conditional on the information available at origination, $\mathrm{PD}_i^{(0)}$ is correctly calibrated: it ranks

borrowers by their historical default frequency, and the partition $\{C_k\}$ reflects this ordering faithfully. The paradox arises because the scoring rule is strictly decreasing in w_i: higher income lowers measured default risk and pushes borrowers toward safer rating classes. Assumption 2 establishes that AI exposure moves in the opposite direction. Neither relationship is anomalous in isolation. Their coincidence on the same variable is what renders the misallocation directional and systematic.

Proposition 1 (Safe-Tail Paradox)**.** *Under Assumptions 1–2, class-average AI exposure is strictly decreasing in the rating class. Thus, the rating system simultaneously satisfies*

$$\bar{e}_1 > \bar{e}_2 > \cdots > \bar{e}_K$$
$$\bar{\pi}_1 < \bar{\pi}_2 < \cdots < \bar{\pi}_K$$

with lower-index classes corresponding to safer borrowers: safer rating classes concentrate borrowers with higher exposure to AI-related income risk.

Proof. Fix $k < k'$. Write $e_i = h(w_i) + \tilde{\nu}_i$ where $h(w) \equiv \mathbb{E}[e_i \mid w_i = w]$ and $\tilde{\nu}_i \equiv e_i - h(w_i)$ satisfies $\mathbb{E}[\tilde{\nu}_i \mid w_i] = 0$ by construction. Since $\beta_w < 0$ and $\beta_{\text{DTI}} > 0$, the scoring rule is therefore strictly decreasing in w_i conditional on $(L_i, \mathbf{z}_i)$, so safer classes contain borrowers with stochastically higher incomes. Integrating over the distribution of $(L_i, \mathbf{z}_i)$, the marginal distribution of $w_i \mid i \in C_k$ first-order stochastically dominates that of $w_i \mid i \in C_{k'}$:

$$F_{w|C_k}(t) \leq F_{w|C_{k'}}(t) \quad \text{for all } t \in \mathbb{R}, \tag{4}$$

where $F_{w|C_k}(t)$ is the cumulative distribution function of the income conditional distribution. Since h is strictly increasing by Assumption 2 and the FSD ordering (4) holds strictly,

$$\mathbb{E}[e_i \mid i \in C_k] = \mathbb{E}\big[h(w_i) \mid i \in C_k\big] + \mathbb{E}\big[\tilde{\nu}_i \mid i \in C_k\big] > \mathbb{E}\big[h(w_i) \mid i \in C_{k'}\big] = \mathbb{E}[e_i \mid i \in C_{k'}], \tag{5}$$

where $\mathbb{E}[\tilde{\nu}_i \mid i \in C_k] = 0$ follows from $\mathbb{E}[\tilde{\nu}_i \mid w_i] = 0$ and the law of iterated expectations. Since $|C_k| \to \infty$ as $n \to \infty$, the law of large numbers gives $\bar{e}_k \xrightarrow{p} \mathbb{E}[e_i \mid i \in C_k]$, so that $\bar{e}_k > \bar{e}_{k'}$ for all $k < k'$, which establishes the proposition. □

The paradox is structural. The same variable that lowers measured default risk loads positively on an omitted factor that increases vulnerability to a common shock. The rating system there-

fore ranks borrowers correctly with respect to historical default frequencies, yet systematically misorders them with respect to their exposure to AI-driven income disruption.

The misalignment has three direct implications for credit risk management. First, the concentration of high-exposure borrowers in the safest rating classes creates a hidden systematic risk: borrowers who appear well diversified along observable dimensions share a common latent vulnerability to AI-driven income disruption, generating within-class default correlation that is absent from both the scoring model and the calibration of asset correlation parameters in standard retail credit frameworks (Vasicek, 2002; Gordy, 2003).

Second, the initial rating order itself becomes unreliable once AI risk materializes. Because safer classes concentrate borrowers with higher AI exposure, a sufficiently large labor market shock will disproportionately raise default risk precisely within those classes, potentially reversing the ranking that the scoring model imposes. A borrower assigned to a safe class may carry higher true default risk than one assigned to a riskier class, not because the model was misestimated, but because the omitted factor loads inversely on the observable used for classification.

Third, and most directly relevant for prudential regulation, regulatory capital under the internal ratings-based (IRB) approach formula is calibrated on $\bar{\pi}_k$ and therefore increasing in the rating class. The Safe-Tail Paradox implies that the classes with the lowest capital buffers are precisely those with the highest latent exposure. An AI-driven shock would thus generate a capital shortfall that is concentrated in the safest segment of the portfolio and increasing in the severity of the shock, the opposite of what the regulatory architecture is designed to handle.

4 AI Stress Test: Design, Shock, and Transmission Channels

4.1 Borrower-Level Stress Test

The Safe-Tail Paradox implies that an adverse AI shock would concentrate losses in the safest rating classes, precisely where capital buffers are lowest. Quantifying this misalignment requires holding the scoring model fixed and evaluating losses under a counterfactual exposure scenario. Stress testing, understood here as a diagnostic of the existing rating system rather than a forecast of portfolio-level outcomes, provides such a framework. The diagnostic targets

the internal consistency of the rating system: whether borrowers assigned to the same rating class remain correctly ordered in terms of true risk once exposure heterogeneity is taken into account. Concretely, the exercise consists in holding the bank's rating model fixed, introducing a borrower-specific AI exposure shock, and recomputing default risk and losses under this counterfactual exposure profile.

Existing stress-testing frameworks, whether supervisory exercises such as the EBA stress tests (BCBS, 2018) or DFAST (Federal Reserve, 2019), or internal stress tests conducted at the initiative of individual institutions, share a common architecture. A macro-financial scenario is specified over aggregate variables such as GDP, unemployment, or asset prices, and satellite models, the econometric mappings that translate macroeconomic variables into default probabilities or losses, distribute this scenario across borrowers (Budnik et al., 2024). This structure holds across thematic variants, including climate or geopolitical stress tests, where the shock remains defined at the aggregate or sector level. In such settings, borrower heterogeneity is entirely induced by the mapping and cannot overturn the ranking implied by the initial scoring model. By construction, these frameworks cannot detect misranking within rating classes.

The present framework departs from this architecture by specifying the shock directly at the borrower level as a vector of occupation-specific exposures, rather than as an aggregate disturbance transmitted through a satellite model. Heterogeneity is not transmitted from a macro scenario; it is the scenario. This shift is essential to the Safe-Tail Paradox: the misalignment between ratings and true risk arises from variation in AI exposure *within* rating classes, a pattern that any aggregate or sectoral specification suppresses by construction. Sectoral aggregates appear only as a convenient reading grid in the presentation of results, not as a modelling assumption that shapes the shock itself.

The most relevant theoretical antecedent is Parlatore and Philippon (2025), who show that the value of stress testing lies primarily in learning about unobserved risk exposures, particularly when regulators can engage in targeted interventions. The present framework inverts this premise: borrower-level AI exposure is observable, albeit imperfectly, from occupational task indices. The question is therefore not whether the rating system uncovers hidden exposures, but whether it correctly ranks this observable dimension of risk within rating classes. The object of interest is whether the loss distribution is aligned with the capital allocation implied

by ratings, formalized in the next section through the exposure shock and its mapping into borrower-level risk.

4.2 Structure of the Shock

Two features distinguish the AI shock from those used in conventional stress tests. First, the shock is a vector rather than a scalar: it assigns a borrower-specific intensity calibrated to the task content of each occupation, rather than a common multiplier of an aggregate risk factor. Second, the shock is defined relative to the bank's information set: the relevant baseline is not a counterfactual about the economy, but the level of AI exposure that current credit models implicitly price, namely zero. The shock therefore measures the gap between the lender's implicit pricing assumption and any chosen point on the realized-to-potential adoption frontier.

Let $e = (e_1, \ldots, e_n)^\top \in [0,1]^n$ denote the AI-exposure vector of the portfolio of n borrowers, where e_i is the share of borrower i's occupational tasks that can be performed or augmented by AI systems. The *implicit* exposure $e_i^{(0)} = 0$ is the level priced by standard credit scoring models, which contain no variable capturing occupational AI exposure (Section 2) and therefore assign all borrowers a zero exposure by construction. The *observed* exposure e_i^{obs} (Massenkoff and McCrory, 2026) is the share of tasks for which AI usage is currently recorded as of March 2026, and the *potential* exposure e_i^{pot} (Eloundou et al., 2024) is the share for which AI could halve task completion time. By construction, $e_i^{\text{pot}} \geq e_i^{\text{obs}} \geq e_i^{(0)} = 0$ for every borrower; both indices are described in Appendix A.

The stress scenario is the transition from the vector $e^{(0)} = 0_n$ to a post-shock vector $e^{(1)}(\delta)$ defined by

$$e_i^{(1)}(\delta) \;=\; e_i^{\text{obs}} + \delta\left(e_i^{\text{pot}} - e_i^{\text{obs}}\right), \qquad \delta \in [0,1], \tag{6}$$

where δ is a single severity index: $\delta = 0$ recovers the level of AI penetration already observed in the economy, $\delta = 1$ the full automation frontier, and $\delta = 0.5$ the mid-transition scenario used in Section 5.

The specification rests on two choices that warrant comment. Anchoring the baseline at $e_i^{(0)} = 0$ is a property of the bank's information set, not a counterfactual claim that AI exposure was nil before ChatGPT. Some exposure existed earlier in some occupations; what

matters here is that no such variable enters internal scoring models, so the lender's effective pricing baseline is zero regardless of the underlying state of the world. The linear interpolation between $\mathbf{e}^{\text{obs}}$ and $\mathbf{e}^{\text{pot}}$ is parsimonious by design: it summarizes the realized-to-potential gap through a single severity parameter, preserves the cross-sectional rank order of exposure at every value of δ, and remains agnostic about the temporal trajectory of adoption. Concave or S-shaped paths can be accommodated by replacing δ with a monotone transformation $\phi(\delta) \in [0, 1]$ without altering the cross-sectional pattern that drives the Safe-Tail Paradox; Appendix B reports robustness to occupation-specific adoption speeds.

Two implications follow. First, even the mildest scenario, $\delta = 0$, constitutes a non-trivial shock for any lender whose provisioning implicitly assumes $\mathbf{e}^{(0)} = \mathbf{0}$: the result obtained at $\delta = 0$ measures the gap between the rating system and an environment that has already materialized. Second, the cross-sectional dispersion of $\mathbf{e}^{(1)}(\delta)$ is preserved at every value of δ: severity scales the average shock without collapsing the heterogeneity that drives the paradox. Alternative specifications that aggregate the shock into a uniform PD multiplier or a sectoral perturbation suppress this heterogeneity by construction, and therefore cannot detect the misordering of risk that the framework is designed to identify.

4.3 Transmission Channels

The shocked exposure $e_i^{(1)}(\delta)$ transmits to borrower income through two labor market channels. The *employment channel* determines whether the borrower retains their job: displacement occurs with probability

$$p_i^{\text{emp}}(\delta) \;=\; \Lambda\big(\gamma_0 + \gamma_1\, e_i^{(1)}(\delta) + \gamma_2\, e_i^{(1)}(\delta)^2\big), \qquad \gamma_1, \gamma_2 > 0, \tag{7}$$

where the quadratic term captures the convexity of displacement risk at high exposure levels, and $\Lambda(\gamma_0)$ is the baseline displacement probability under $e_i^{(0)} = 0$. Cross-sectional heterogeneity in displacement risk arises solely through variation in $e_i^{(1)}(\delta)$.

The *wage channel* governs the income of employed borrowers and operates asymmetrically across the exposure distribution. Its theoretical basis lies in the task-based framework of Autor et al. (2003) and Acemoglu and Restrepo (2022), in which automation substitutes for labor in some tasks while complementing others. Generative AI partly overturns this logic: the cognitive, non-routine tasks that previously commanded a wage premium are precisely those

that large language models can now replicate (Eloundou et al., 2024). As a result, workers in highly exposed occupations face earnings compression, while the sign and magnitude of complementarity effects for low-exposure workers remain empirically uncertain (Handa et al., 2025; Brynjolfsson et al., 2025).

Aggregate evidence does not resolve this ambiguity. While firm-level studies for France show that AI adoption can coincide with employment and sales growth (Aghion et al., 2025), these average effects mask substantial dispersion in individual outcomes (Aghion and Bouverot, 2024). Stress testing targets this lower tail: the relevant question is not whether AI raises average productivity, but whether a non-negligible mass of borrowers experiences income losses sufficient to affect default risk. We therefore retain the substitution margin as the primary transmission channel and absorb complementarity gains into a stochastic residual.

For borrowers above a threshold $\bar{e}$, high exposure generates downward pressure on wages as tasks become replicable. Below the threshold, exposure is sufficiently limited that income is left unchanged. The two channels combine conditionally: a borrower first faces displacement risk with probability $p_i^{\text{emp}}(\delta)$ and, conditional on remaining employed, experiences the wage effect specified in equation (8):

$$w_i^{(1)}(\delta) \;=\; \begin{cases} \lambda^{\text{UI}}\, w_i^{(0)} & \text{if displaced,} \\ w_i^{(0)}\big(1 - \mu\, e_i^{(1)}(\delta)\big) + \varepsilon_i & \text{if employed and } e_i^{(1)}(\delta) \geq \bar{e}, \\ w_i^{(0)} & \text{if employed and } e_i^{(1)}(\delta) < \bar{e}, \end{cases} \tag{8}$$

where $\lambda^{\text{UI}} \in (0,1)$ is the unemployment replacement rate, $\mu > 0$ governs wage compression, and $\varepsilon_i \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ captures unobserved heterogeneity in adaptation, firm-level task reorganization, wage-setting conditions, and residual complementarity effects. The threshold $\bar{e}$ is calibrated on the cognitive–routine boundary documented by Autor et al. (2003); calibration values are reported in Appendix B.

This transmission structure is the dynamic counterpart of Assumption 2. The positive relationship between AI exposure and income at origination reflects a cross-sectional equilibrium in which cognitively intensive tasks command a wage premium. The wage channel captures instead the intertemporal erosion of these rents under automation. The shock preserves the initial ranking but weakens its economic foundation: the borrowers with the highest initial

incomes are also those facing the largest post-shock income declines. This alignment is the source of the systematic misallocation of risk documented in the Safe-Tail Paradox.

4.4 Dimensions of Analysis

The stress test examines three main quantities across the severity spectrum $\delta \in [0, 1]$: (i) post-shock default probabilities and the associated calibration bias, (ii) mismeasurement matrices, and (iii) regulatory capital shortfalls.

Post-shock default probabilities. Substituting post-shock income $w_i^{(1)}(\delta)$ into the scoring equation, with all coefficients fixed at their origination values, yields individual post-shock PDs

$$\mathrm{PD}_i^{(1)}(\delta) \;=\; \Lambda\big(\alpha + \beta_w\, w_i^{(1)}(\delta) + \beta_{\mathrm{DTI}}\, \tfrac{L_i}{w_i^{(1)}(\delta)} + \beta_z^\top \mathbf{z}_i\big), \tag{9}$$

and their average within each pre-shock rating class C_k

$$\bar{\pi}_k^{(1)}(\delta) \;\equiv\; \frac{1}{|C_k|} \sum_{i \in C_k} \mathrm{PD}_i^{(1)}(\delta). \tag{10}$$

For each rating class k, the calibration bias is defined as the relative gap between the post-shock class PD and the pre-shock anchor on which regulatory capital is based:

$$\mathcal{B}_k(\delta) \;\equiv\; \frac{\bar{\pi}_k^{(1)}(\delta) - \bar{\pi}_k^{(0)}}{\bar{\pi}_k^{(0)}}, \tag{11}$$

where $\bar{\pi}_k^{(0)}$ denotes the pre-shock calibrated PD for class k. A positive bias indicates that banks underestimate required capital by neglecting AI exposure; a negative value indicates overestimation.

Mismeasurement matrices. The cross-sectional manifestation of the Safe-Tail Paradox is captured by a *mismeasurement matrix* $\mathbf{M}(\delta) \in [0, 1]^{K \times K}$, structurally identical to a credit migration matrix but whose displacement is static rather than temporal: it records the gap between the rating assigned by an AI-blind scoring model and the rating that would obtain under a counterfactual scoring that prices AI exposure. Element $M_{k\ell}(\delta)$ is the share of borrowers originating in class k whose stressed PD falls in class ℓ:

$$M_{k\ell}(\delta) \;=\; \frac{1}{|\mathcal{C}_k|} \sum_{i \in \mathcal{C}_k} \mathbf{1}\Big\{\mathrm{PD}_i^{(1)}(\delta) \in [\theta_{\ell-1}, \theta_\ell)\Big\}. \tag{12}$$

Each row sums to one. Diagonal entries measure class stability under the shock; upper-triangular mass measures the share of class-k borrowers whose true risk warrants a downgrade once AI exposure is priced, and lower-triangular mass measures the corresponding upgrades.

Regulatory capital shortfall. Let

$$\kappa_k(\pi) \;\equiv\; \mathrm{LGD}\cdot\left[\Phi\left(\frac{\Phi^{-1}(\pi)+\sqrt{\rho_k}\,\Phi^{-1}(0.999)}{\sqrt{1-\rho_k}}\right)-\pi\right] \tag{13}$$

denote the Vasicek–Gordy IRB capital charge per unit of Exposure at Default (EAD) for class k, where ρ_k is the asset correlation of class k, held fixed throughout the stress test. The class-level shortfall measures the gap between the capital required under the stressed PD and the capital held on the basis of the pre-shock rating, expressed as a percentage of the pre-shock regulatory capital base:

$$\Delta K_k(\delta) \;\equiv\; \frac{\kappa_k\big(\bar{\pi}_k^{(1)}(\delta)\big)-\kappa_k\big(\bar{\pi}_k^{(0)}\big)}{\kappa_k\big(\bar{\pi}_k^{(0)}\big)}. \tag{14}$$

The aggregate portfolio shortfall is the EAD-weighted average across classes:

$$\Delta K(\delta) \;\equiv\; \frac{\sum_{k=1}^{K}\Big(\sum_{i\in C_k}\mathrm{EAD}_i\Big)\cdot\Delta K_k(\delta)}{\sum_{k=1}^{K}\sum_{i\in C_k}\mathrm{EAD}_i}, \tag{15}$$

which measures the average percentage undercapitalization of the portfolio relative to its pre-shock regulatory capital base.

5 AI Stress Test: Results

5.1 The Safe-Tail Paradox in a Calibrated Portfolio

A credible test of the Safe-Tail Paradox requires a portfolio that preserves the joint distribution of income, contract type, leverage, and occupational task content at the borrower level, and that reproduces the rating behavior of a retail book under the IRB approach. We construct a synthetic portfolio calibrated to reproduce the aggregate characteristics of the French residential mortgage market. The choice of France is pragmatic rather than structural: publicly available HCSF, ACPR, and INSEE statistics provide the regulatory constraints, origination parameters, and sectoral wage distribution required to discipline each borrower-level variable, but the mechanism itself is country-agnostic.[3]

[3] ACPR (*Autorité de Contrôle Prudentiel et de Résolution*) is the French banking authority; INSEE (*Institut National de la Statistique et des Études Économiques*) is the French national statistics office.

Portfolio Construction and Calibration. A synthetic portfolio is the appropriate instrument for the question we ask for two distinct reasons. First, no public dataset crosses the required dimensions at the borrower level, and regulatory reporting does not permit occupation-level reconstruction. Second, the Safe-Tail Paradox arises from a specific interaction between income and AI exposure across occupations: identifying it cleanly requires varying these dimensions independently, which no available dataset for the French market permits since income, contract type, and occupational task content are endogenously correlated in any real borrower population. The synthetic approach therefore trades empirical realism in selected dimensions for clean identification of the mechanism.

The portfolio is built around a sectoral partition that serves two purposes simultaneously. First, it disciplines the borrower mix to match the aggregate French employment structure: $n = 5{,}000$ borrowers are partitioned into ten broad economic sectors aggregated from the INSEE sectoral classification, with portfolio weights chosen so that sectoral shares track INSEE employment statistics (INSEE, 2024) adjusted for the homeowner composition of the French mortgage book. Second, the same sectoral partition provides the bridge between borrower-level attributes and occupational AI exposure indices, which are published at the SOC major-group level and have no direct French equivalent: we map the task coverage of the Anthropic Economic Index (Massenkoff and McCrory, 2026), which measures observed AI task coverage by occupation, onto our sectoral partition through a SOC-to-NAF correspondence table. The sectoral partition thus simultaneously reproduces the credit risk attributes of each borrower and his occupational AI exposure, preserving their empirical joint distribution without imposing any assumed correlation structure.

Conditional on a sector, each borrower is constructed in three steps. First, gross annual income is drawn from INSEE statistics by NAF sector (INSEE, 2024), and employment contract status is assigned with probability equal to the sectoral share of permanent contracts (*contrats à durée indéterminée*, CDI). Second, loan size is derived from HCSF Decision D-HCSF-2021-7 (35% DTI cap, 25-year maximum maturity) at the Q1 2024 average French mortgage rate, yielding a mean principal of €221,000 in line with ACPR new-production statistics (ACPR, 2024). Third, a logistic scoring model of the form (1) is calibrated to match the ACPR 12-month default rate of 0.44%. The resulting portfolio reproduces the main aggregate features of the

French residential book: mean gross income €46,534, mean LTV 79.9%, mean DTI 32.8%, and a permanent-contract share of 84.2%.

Occupational AI exposure enters the borrower vector through two indices, e_i^{obs} and e_i^{pot}, set equal to the within-sector means delivered by the SOC-to-NAF crosswalk plus a small idiosyncratic dispersion that captures within-sector occupational heterogeneity.[4] Crucially, neither index feeds into the scoring model or the rating partition, implementing Assumption 1.[5] The parameters of the stress equations (7)–(8) are anchored on independent sources: the French regulatory environment for the origination block; the US micro-evidence of Brynjolfsson et al. (2025) for the employment elasticity. Where French micro-estimates are unavailable, we rely on US evidence and verify robustness by re-running the stress scenario with elasticities scaled to ±30% of their baseline values; the Safe-Tail Paradox obtains across the full range, confirming that the result is not an artefact of parameter choice. No parameter is tuned on the post-shock distribution of PDs or capital: the paradox is inherited from the joint distribution of income and AI exposure across occupations, not engineered by the calibration. Appendix B documents the portfolio construction in full and reports the sensitivity analysis.

The Safe-Tail Paradox. Figure 2 confronts the two sides of the paradox: the safest borrowers bear the heaviest AI exposure. Panel A reports the class-level calibrated probability of default $\bar{\pi}_k^{(0)}$, which rises monotonically from 0.21% in the AAA class to 3.82% in the CCC class, by construction of the rating partition. Panel B reports the class-average observed AI exposure $\bar{e}_k^{\text{obs}}$, which declines from 23.0% in the AAA class to 17.6% in the CCC class. The pattern is structural: across 20 simulation seeds, the Spearman rank correlation between $\bar{e}_k$ and k averages -0.98.

The mechanism is transparent in the sectoral composition of each rating class. AAA and AA borrowers are concentrated in Finance & Insurance ($\bar{e}^{\text{obs}} = 28\%$), Information & Technology (36%), and Public Administration (34%), sectors with high incomes, stable contracts, and low default risk, but also the highest observed AI task coverage. CCC borrowers, by contrast,

[4] NAF (*Nomenclature des Activités Françaises*) is the French national activity classification, broadly equivalent to NAICS in the US. SOC (Standard Occupational Classification) is the US occupational classification system used by the Bureau of Labor Statistics.

[5] This matches common practice: the publicly documented IRB methodologies of the main French retail banks (see, e.g., the Pillar 3 disclosures of BNP Paribas, Crédit Agricole SA, Société Générale, and BPCE) do not incorporate an occupational AI exposure variable, relying instead on income, contract type, DTI, LTV, and, for some institutions, broad sector indicators that aggregate over widely heterogeneous task contents.

are drawn disproportionately from Construction ($\bar{e}^{\text{obs}} = 5\%$), Hospitality & Food (6%), and Manufacturing (9%), sectors with lower incomes and higher credit risk, but also the lowest AI exposure.

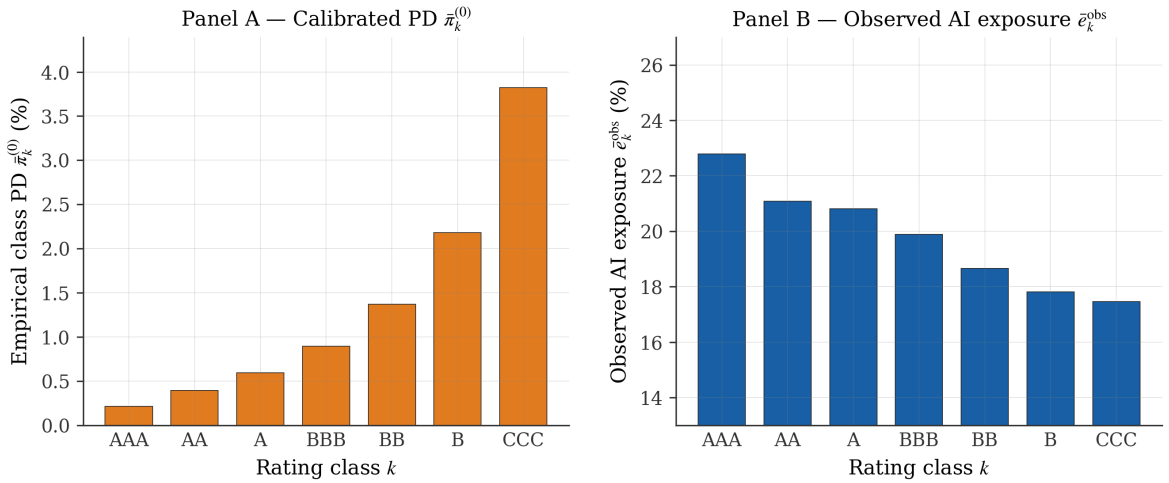

Figure 2. The Safe-Tail Paradox. Panel A displays the class-level calibrated probability of default $\bar{\pi}_k^{(0)}$; Panel B displays the class-average observed AI exposure $\bar{e}_k^{\text{obs}}$. The juxtaposition makes Proposition 1 empirically visible: the highest-rated borrower classes carry the greatest occupational AI exposure. Borrowers are partitioned into $K = 7$ rating classes using the empirical quantiles of $\text{PD}_i^{(0)}$. Synthetic portfolio; calibration details in Appendix B.

5.2 Stress Test Results

At the AI-exposure level already observed in 2026, internalising the shock in the scoring model raises calibrated PDs by 36% in AAA against 8% in CCC, downgrades 6.2% of the AAA cohort to a lower rating, and lifts regulatory capital requirements by 25% in AAA against 4% in CCC. The AI-blind scoring model thus under-prices its safest borrowers four to six times more than its riskiest, the empirical signature of the Safe-Tail Paradox.

Post-shock PDs and calibration bias. Figure 3 displays the within-class calibrated PD before and after the shock (left axis) together with the calibration bias $\mathcal{B}_k(\delta)$ (right axis). Under the observed-exposure scenario (Panel A), the AAA class exhibits a bias of 36%, while the CCC class exhibits a bias of only 8%. Under the mid-transition scenario (Panel B), the asymmetry amplifies: $\mathcal{B}_k$ reaches 105% at AAA, against 23% at CCC. The AI shock concentrates most forcefully in the rating classes for which the scoring model implies the lowest baseline default probabilities: the AAA class, nominally the safest, exhibits the largest

calibration bias; the CCC class, nominally the riskiest, the smallest.

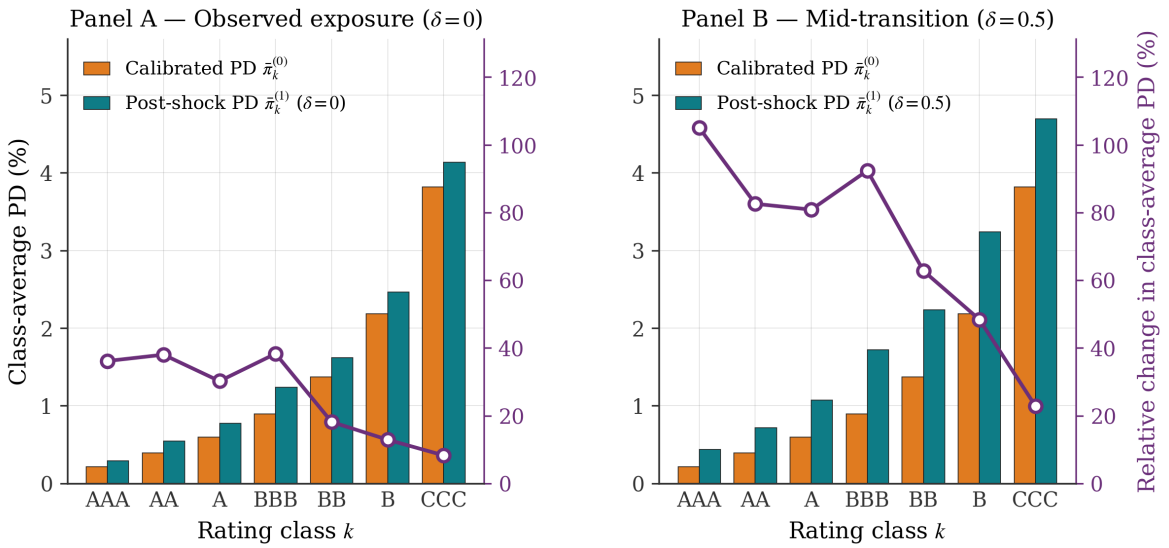

Figure 3. Post-shock class-average PDs and calibration bias. This figure displays the class-average PDs before and after the AI shock (left axis) and relative change in class-average PD (right axis, plum line). Post-shock PDs $\bar{\pi}_k^{(1)}(\delta)$ are computed from equation (10); relative change is $\bar{\pi}_k^{(1)}(\delta)/\bar{\pi}_k^{(0)} - 1$ (in %). Same portfolio and calibration as Figure 2.

Mismeasurement matrices. The mismeasurement matrices $\mathbf{M}(0)$ and $\mathbf{M}(0.5)$, reported in Figures 4 and 5, quantify the rating misclassification of the AI-blind scoring model: each cell $M_{k\ell}(\delta)$ is the share of borrowers the model rates in class k that would instead belong in class ℓ once the AI shock is properly priced; off-diagonal mass therefore measures the size of the mismeasurement, and its concentration above the diagonal measures the share of borrowers the model treats as safer than they truly are. Under the observed-exposure scenario, 6.2% of borrowers rated AAA are already misclassified.

Under the mid-transition scenario, only 47.8% of the AAA cohort would remain there if AI exposure were internalised, with 34.1% belonging in AA and 12.2% in A. As with the calibration bias, the mismeasurement concentrates in the rating classes the scoring model labels as safest, consistent with Proposition 1.

Regulatory capital shortfall. Figure 6 maps the post-shock class-average PDs into the Vasicek–Gordy IRB formula and reports, for each rating class k, the capital shortfall $\Delta K_k(\delta)$. The EAD-weighted aggregate shortfall $\Delta K(\delta)$, reported above each panel, reaches 21.4% under the observed-exposure scenario and 52.1% under the mid-transition scenario. The ratio of

From \ To	AAA	AA	A	BBB	BB	B	CCC	Total	% down	% up
AAA	93.8	2.6	0.1	0.1	0.2	2.4	0.8	100%	6.2%	0.0%
AA	2.9	89.3	4.5	—	0.1	0.5	2.7	100%	7.8%	2.9%
A	—	3.5	89.4	3.5	—	0.3	3.2	100%	7.0%	3.5%
BBB	—	—	3.9	89.2	2.8	0.1	4.0	100%	6.9%	3.9%
BB	—	—	—	4.5	89.0	2.5	4.0	100%	6.5%	4.5%
B	—	—	—	—	3.7	91.3	5.0	100%	5.0%	3.7%
CCC	—	—	—	—	—	9.0	91.0	100%	0.0%	9.0%

Figure 4. Mismeasurement matrix $\mathbf{M}(\delta)$ at $\delta = 0$. Rows are origin classes $\mathcal{C}_k$; columns are classes computed at observed AI exposure $e_i^{(1)}(0) = e_i^{\text{obs}}$. Cell entries are percentage shares. Red entries are downgrades, green entries are upgrades, the diagonal is shaded grey, and em-dashes indicate shares below 0.05%.

From \ To	AAA	AA	A	BBB	BB	B	CCC	Total	% down	% up
AAA	47.8	34.1	12.2	1.1	0.6	2.9	1.3	100%	52.2%	0.0%
AA	0.9	40.8	35.4	17.1	1.3	0.6	3.9	100%	58.3%	0.9%
A	—	1.5	41.2	37.3	15.5	0.7	3.7	100%	57.2%	1.5%
BBB	—	—	1.8	44.0	42.4	7.3	4.6	100%	54.3%	1.8%
BB	—	—	—	1.2	55.3	35.3	8.2	100%	43.5%	1.2%
B	—	—	—	—	1.7	59.3	39.0	100%	39.0%	1.7%
CCC	—	—	—	—	—	3.0	97.0	100%	0.0%	3.0%

Figure 5. Mismeasurement matrix $\mathbf{M}(\delta)$ at $\delta = 0.5$. Same layout as Figure 4. At $\delta = 0.5$, effective exposure is $e_i^{(1)}(0.5) = e_i^{\text{obs}} + 0.5\,(e_i^{\text{pot}} - e_i^{\text{obs}})$, halfway between currently observed and full potential AI adoption.

relative capital changes between AAA and CCC (approximately six) exceeds the corresponding ratio of relative PD changes (approximately four): this amplification reflects the local elasticity of the Vasicek–Gordy formula, which is higher at the low baseline PDs characteristic of the safest classes. The IRB formula thus mechanically magnifies the Safe-Tail Paradox in the units that matter for regulatory buffers.

The aggregate shortfall at $\delta = 0$ does not correspond to an adverse future scenario: it merely repositions the portfolio at the AI exposure level that has already materialized. A bank

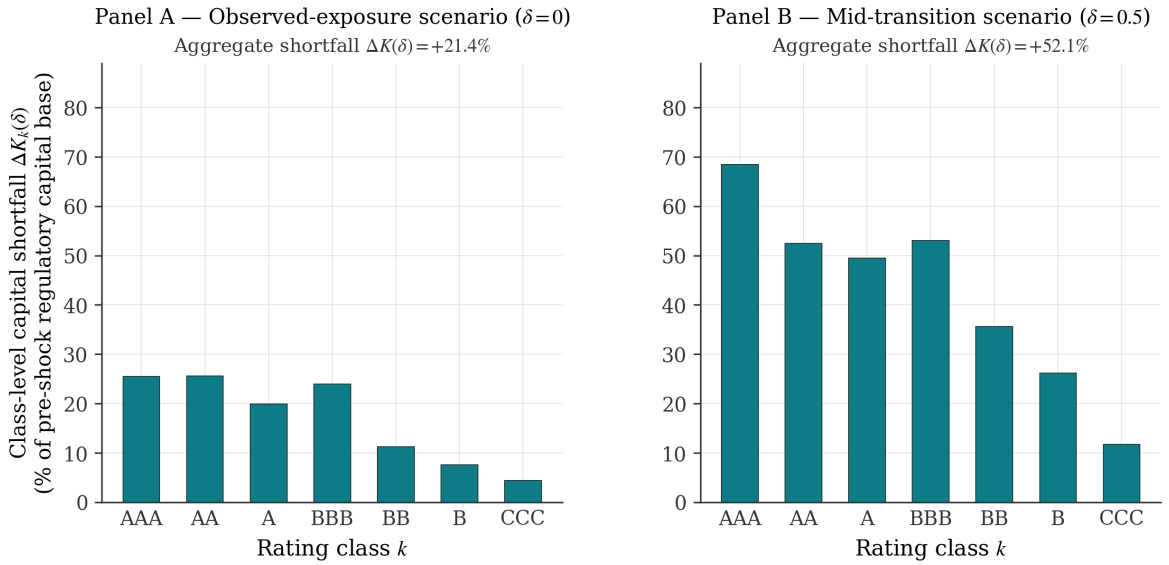

Figure 6. AI-induced mismeasurement of regulatory capital. This figure displays the class-level capital shortfall $\Delta K_k(\delta)$ expressed as a percentage of the pre-shock regulatory capital base. The EAD-weighted aggregate shortfall $\Delta K(\delta)$ is reported above each panel. Same portfolio and calibration as Figure 2.

whose provisioning and capital allocation implicitly assume $e_i^{(0)} = 0$ holds 21% less regulatory capital than would be required under a correct pricing of currently observable AI exposure (as of March 2026). The shortfall is therefore not a forecast of future stress, but a measure of the gap between the lender's scoring model and the environment of March 2026. Moreover, this shortfall is concentrated in the safest segment of the portfolio, precisely the segment in which the IRB formula implies the thinnest capital buffers: the classes with the lowest calibrated PDs absorb the largest relative capital revision, the opposite of what a risk-based ordering of buffers is designed to produce. The Safe-Tail Paradox thus compresses the relative protection of the safest classes while leaving the overall loss distribution under-provisioned.

These aggregate magnitudes could, in principle, be detected by a macro-prudential exercise that shocks the portfolio by its EAD-weighted mean PD change. Such an exercise would flag an aggregate revision comparable in scale to ours: under a uniform +80% PD shock applied to every borrower at $\delta = 0.5$, the IRB formula mechanically implies an aggregate capital increase of +47%, not far from the +52% we obtain. What a uniform shock cannot recover is the cross-sectional pattern that defines the paradox: applied uniformly, the same aggregate PD change produces an AAA-to-CCC relative-capital ratio of 1.5, against the factor of six generated by the borrower-level shock. The distinguishing feature of the Safe-Tail Paradox is not the size of

the aggregate shortfall but its sign pattern along the rating scale, and that pattern is invisible to any stress test that does not model AI exposure at the borrower level.

6 Conclusion

Internal rating systems for retail mortgage portfolios contain no variable capturing occupational AI exposure. In a stable labor market, this omission has no measurable cost. When a technological shock concentrates along occupational lines, as emerging evidence on AI labor market adjustment suggests it is beginning to do, the omission generates a structural paradox: the borrowers most exposed to AI-related labor income disruption are precisely those the scoring model classifies as safest.

The policy implication is direct. Supervisory stress tests conditioned on aggregate variables cannot detect this pattern by construction. Closing this gap does not require new data: loan-level occupational records are already maintained by mortgage lenders, and occupation-level AI exposure indices are publicly available. What is required is a conceptual shift toward frameworks that explicitly model the cross-sectional heterogeneity of technological exposure within the borrower population, rather than the aggregate transmission of macro shocks alone. The framework developed in this paper is one concrete realisation of that shift: a tractable, calibrated borrower-level stress test that combines loan-level occupational records with publicly available AI exposure indices.

While our analysis focuses on residential mortgages, the mechanism extends naturally to other retail credit products. Student loans are a particularly salient case: the occupational sorting that produces the Safe-Tail Paradox begins precisely at the educational stage, and the graduates most exposed to AI substitution are often those whose credentials commanded the highest earning premia at origination. More broadly, any scoring model that uses income or educational attainment as a proxy for creditworthiness inherits the same blind spot, wherever the correlation between AI exposure and the classification variables holds.

The Safe-Tail Paradox points to a general limitation of risk measurement systems calibrated on historical default frequencies in a stable environment. When the source of a new risk is correlated with the variables used for classification, the system does not merely fail to measure that risk: it systematically misallocates it. Given the scale of AI's expected effects on

employment and labor income across the full wage distribution, this misallocation may prove to be one of the most consequential blind spots in contemporary prudential regulation.

Bibliography

Acemoglu, D. and Restrepo, P. (2022). Tasks, automation, and the rise in U.S. wage inequality. *Econometrica*, 90(5):1973–2016.

Acharya, V. V., Berger, A. N., and Roman, R. A. (2018). Lending implications of U.S. bank stress tests: Costs or benefits? *Journal of Financial Intermediation*, 34:58–90.

ACPR (2024). Le financement de l'habitat en 2023. Autorité de Contrôle Prudentiel et de Résolution, Analyses et Synthèses, 160.

Aghion, P. and Bouverot, A. (2024). IA: Notre ambition pour la France. Rapport de la Commission de l'Intelligence Artificielle.

Aghion, P., Bunel, S., Jaravel, X., Mikaelsen, T., Roulet, A., and Søgaard, J. (2025). How different uses of AI shape labor demand: Evidence from france. *AEA Papers and Proceedings*, 115:62–67.

Autor, D. H., Levy, F., and Murnane, R. J. (2003). The skill content of recent technological change: An empirical exploration. *Quarterly Journal of Economics*, 118(4):1279–1333.

Azar, J., Giné, M., and Sanz-Espín, J. (2025). AI is already eroding wages: Quasi-experimental evidence from occupational exposure. SSRN Working Paper.

BCBS (2018). Stress testing principles. Basel Committee on Banking Supervision, Guidelines.

Brynjolfsson, E., Chandar, B., and Chen, R. (2025). Canaries in the coal mine? Six facts about the recent employment effects of artificial intelligence. Stanford Digital Economy Lab.

Budnik, K., Ponte Marques, A., Giglio, C., Grassi, A., Durrani, A., Figueres, J. M., Konietschke, P., Le Grand, C., Metzler, J., Población García, F. J., Shaw, F., Groß, J., Sydow, M., Franch, F., Georgescu, O.-M., Ortl, A., Trachana, Z., and Chalf, Y. (2024). Advancements in stress-testing methodologies for financial stability applications. ECB Occasional Paper No. 348.

EBA (2017). Guidelines on PD estimation, LGD estimation and the treatment of defaulted exposures. European Banking Authority, EBA/GL/2017/16.

Eloundou, T., Manning, S., Mishkin, P., and Rock, D. (2024). GPTs are GPTs: Labor market impact potential of LLMs. *Science*, 384(6702):1306–1308.

Engberg, E., Koch, M., Lodefalk, M., and Schroeder, S. (2025). Artificial intelligence, tasks, skills, and wages: Worker-level evidence from Germany. *Research Policy*, 54(8):105285.

Farmer, J. D., Kleinnijenhuis, A. M., Schuermann, T., Wyman, O., and Wetzer, T. (2022). *Handbook of Financial Stress Testing*. Cambridge University Press.

Federal Reserve (2019). Dodd-Frank act stress test 2019: Supervisory stress test methodology. Board of Governors of the Federal Reserve System.

Felten, E. W., Raj, M., and Seamans, R. (2023). How will language modelers like ChatGPT affect occupations and industries? SSRN Working Paper.

Fédération Bancaire Française (2026). Le financement des particuliers. Fédération Bancaire Française, Études économiques, March 2026.

Ghosal, S., Sen, A., and van der Vaart, A. W. (2000). Testing monotonicity of regression. *Annals of Statistics*, 28(4):1054–1082.

Gordy, M. B. (2003). A risk-factor model foundation for ratings-based bank capital rules. *Journal of Financial Intermediation*, 12(3):199–232.

Handa, K., Tamkin, A., McCain, M., Huang, S., Durmus, E., Heck, S., Mueller, J., Hong, J., Ritchie, S., Belonax, T., Troy, K. K., Amodei, D., Kaplan, J., Clark, J., and Ganguli, D. (2025). Which economic tasks are performed with AI? Evidence from millions of Claude conversations. Anthropic.

HCSF (2019). Diagnostic des risques dans le secteur de l'immobilier résidentiel. Haut Conseil de stabilité financière.

HCSF (2021). Décision D-HCSF-2021-7 relative aux conditions d'octroi de crédits immobiliers. Journal Officiel de la République Française. Haut Conseil de Stabilité Financière; Amended by D-HCSF-2023-2, 29 June 2023.

Humlum, A. and Vestergaard, E. (2025). Still waters, rapid currents: Early labor market transformation under generative AI. NBER, Working Paper 33777.

INSEE (2024). Emploi, chômage, revenus du travail, Edition 2024.

Kochhar, R. (2023). Which U.S. workers are more exposed to AI on their jobs? Pew Research Center Report.

Leitner, Y. and Williams, B. (2023). Model secrecy and stress tests. *Journal of Finance*, 78(2):1055–1095.

Massenkoff, M., Lyubich, E., McCrory, P., Appel, R., and Heller, R. (2026). Anthropic economic index report: Learning curves. Anthropic.

Massenkoff, M. and McCrory, P. (2026). Labor market impacts of AI: A new measure and early evidence. Anthropic.

Parlatore, C. and Philippon, T. (2025). Designing stress scenarios. *Journal of Finance*, 80(2):833–873.

Shapiro, J. and Zeng, J. (2024). Stress testing and bank lending. *Review of Financial Studies*, 37(4):1265–1314.

Vasicek, O. (2002). Loan portfolio value. *Risk*, 15(12):160–162.

Appendix A: AI Exposure–Income Relationship

This appendix provides empirical support for Assumption 2 of the main text, which posits that AI exposure is increasing in wage across the occupational distribution.

A.1 Data

We document the AI exposure–Income relationship using two complementary AI exposure indices and a nationally representative wage benchmark for the US. The primary AI exposure index is the *observed exposure* measure of Massenkoff and McCrory (2026), released alongside the paper in the Anthropic Economic Index repository.[6] For each six-digit Standard Occupational Classification (SOC) occupation j, $e_j^{\text{obs}} \in [0,1]$ is the time-weighted share of the occupation's O*NET tasks (U.S. Department of Labor) that are *AI-covered*, where a task is counted as covered if (i) an LLM (alone or augmented with software tools) could reduce the time required to complete it by at least half (Eloundou et al., 2024) and (ii) it shows sufficient work-related Claude usage in the Anthropic Economic Index (Massenkoff et al., 2026). Task-level coverage indicators are then averaged to the occupation level using O*NET time-fraction weights. The measure therefore captures the share of an occupation's task portfolio plausibly being delegated to AI in practice, not the broader share that is theoretically automatable.

As a comparison benchmark, we use the theoretical β exposure of Eloundou et al. (2024), which measures the share of tasks that LLMs could plausibly accelerate according to a human- and GPT-4-coded rubric. The rubric assigns each O*NET task one of three labels: E0 (no time reduction or quality loss), E1 (direct exposure: time reduced by at least 50% using a current LLM such as ChatGPT or the OpenAI Playground), or E2 (LLM-plus exposure: a 50% time reduction is feasible only with additional software or tooling built on top of the LLM). The occupation-level β measure is then defined as the within-occupation share of tasks rated E1, plus one-half of the share rated E2By construction, $\beta_j \in [0,1]$ for each occupation j, and both human and GPT-4 annotation pipelines yield broadly consistent rankings. Because the rubric scores potential time savings rather than measured deployment, β is therefore best read as a measure of *potential* exposure.

Wage and employment data are drawn from the Bureau of Labor Statistics (BLS) Occupational Employment and Wage Statistics survey (OEWS, May 2021, national dataset), which reports annual mean wages and employment counts at the six-digit SOC level. The three sources are merged on SOC codes, yielding a cross-section of $m = 729$ occupations after dropping occupations with missing exposure scores or wage data.[7] For each occupation j, let e_j denote the AI exposure index and w_j the annual mean wage. The unit of observation is the occupation,

[6] File `job_exposure.csv`.

[7] Dropped occupations are concentrated in military-specific (SOC codes 55-XXXX) and residual *All other* categories, which represent a negligible share of civilian employment.

and all regressions are weighted by occupation-level employment to ensure representativeness of the employed population.

A.2 Nonparametric Estimation of $\mathbb{E}[e_i \mid w_i]$

To estimate the conditional mean function $h(w) = \mathbb{E}[e_i \mid w_i = w]$ without imposing functional-form restrictions, we use a Nadaraya–Watson kernel estimator,

$$\hat{h}(w) = \frac{\sum_{j=1}^{N} K_b(w - w_j)\, e_j\, \omega_j}{\sum_{j=1}^{N} K_b(w - w_j)\, \omega_j}, \tag{16}$$

where $K_b(\cdot) = K(\cdot/b)/b$ is an Epanechnikov kernel with bandwidth b selected by leave-one-out cross-validation, and ω_j denotes the employment weight of occupation j. The bandwidths selected by cross-validation are $\hat{b}_{\mathrm{O}} = \$10{,}490$ for the observed-exposure measure and $\hat{b}_{\mathrm{E}} = \$3{,}709$ for the Eloundou theoretical benchmark (both in nominal annual wage units).

Figure 1 plots $\hat{h}(w)$ over the support of the wage distribution, together with pointwise 95% confidence bands constructed by the wild bootstrap with 999 replications. For the theoretical β benchmark, the estimated conditional mean rises monotonically over essentially the full support of the wage distribution, from approximately 15% in the lowest wage decile to over 50% in the top decile. For the Massenkoff—McCrory observed-exposure index, the relationship is positive on average. Both the Ghosal statistic and the employment-weighted rank regression reject the null of a non-increasing relationship. Although the relationship is not strictly monotonic in levels, this rejection is primarily driven by the strongly increasing segment over the lower-to-middle wage range and is not offset by the mild decline observed at the very top of the distribution. Observed exposure rises steeply from the bottom of the wage distribution, peaks around the middle ($w \approx \$60$–$\$80$k, corresponding to mid-skill cognitive occupations such as programmers, data-entry workers, customer-service representatives, and analysts), and declines mildly at the very top, where high-earning physicians, executives, and craft specialists show comparatively little observed Claude usage. Across the wage support, the *observed*-exposure curve lies systematically below the *theoretical* benchmark, consistent with the finding of Massenkoff and McCrory (2026) that realised AI usage tracks but does not fully exploit theoretical capability.

A.3 Monotonicity Test

We test the null hypothesis H_0: h is non-increasing on a set of positive measure against the alternative H_1: h is strictly increasing, using the nonparametric monotonicity test of Ghosal

et al. (2000). In its studentised finite-difference form, the test statistic is

$$T_m = \sup_{w<w'} \frac{\hat{h}(w') - \hat{h}(w)}{\hat{\sigma}(w, w')}, \tag{17}$$

where $\hat{\sigma}(w, w')$ is a consistent estimator of the standard deviation of $\hat{h}(w') - \hat{h}(w)$ obtained from the wild bootstrap. Under H_0, T_m converges in distribution to the supremum of a standardised Gaussian process; critical values are approximated from the bootstrap distribution of the centred statistic. Table 1 reports the results for both AI exposure indices.

Table 1. Monotonicity tests: AI exposure over the wage distribution. Columns (1) and (3) report the nonparametric test of Ghosal et al. (2000). Columns (2) and (4) report the employment-weighted OLS, rank–rank regression, and R^2. All specifications use $m = 729$ occupations. Bootstrap p-values based on 999 replications. HC1-robust standard errors in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$. Employment-weighted. NP: Nadaraya–Watson estimator, bandwidth selected by leave-one-out cross-validation. Rank–rank: Spearman-type OLS on fractional ranks of e_j and w_j.

	Observed exposure		Theoretical exposure	
	NP test	OLS / Rank–rank	NP test	OLS / Rank–rank
	(1)	(2)	(3)	(4)
T_m	4.14	—	77.45	—
Bootstrap p-value	0.003	—	< 0.001	—
OLS slope (per \$10k)	—	0.008^{**}	—	0.026^{***}
	—	(0.004)	—	(0.004)
Rank–rank slope	—	0.319^{***}	—	0.502^{***}
	—	(0.088)	—	(0.062)
R^2 (OLS)	—	0.02	—	0.17

The nonparametric test rejects the null of non-increasing exposure for both indices: $T_m = 4.14$, bootstrap $p = 0.003$ for the observed measure; $T_m = 77.45$, bootstrap $p < 0.001$ for the theoretical β. The employment-weighted OLS slope implies that a \$10,000 increase in annual mean occupation wage is associated with a 0.8 percentage point increase in the Massenkoff–McCrory observed-exposure index ($R^2 = 0.02$, significant at the 5% level) and a 2.6 percentage point increase in the Eloundou index ($R^2 = 0.17$, significant at the 1% level). The low R^2 on the observed index reflects the non-monotone shape documented in Figure 1: a linear summary misses the mid-wage hump, but rank–rank regressions continue to identify a positive unconditional relationship. A one-decile increase in wage rank is associated with a 3.2 and 5.0 decile increase in exposure rank for the observed and theoretical indices respectively. These results are robust to alternative bandwidth choices.

Appendix B: Portfolio Construction and Calibration

This appendix documents the construction of the synthetic mortgage portfolio used in Section 5 to quantify the Safe-Tail Paradox and implement the borrower-level AI stress test. The portfolio is designed to be *computationally reproducible* and to match the joint distribution of borrower characteristics that drive internal rating systems for residential mortgage books, rather than to be empirically realistic on every individual dimension. French publicly available benchmarks are used as calibration targets wherever possible. All algorithms are deterministic conditional on a single integer seed, so the simulated portfolio can be regenerated. Notation follows the main text; quantities introduced only in this appendix are indexed with the prefix B.

B.1 Design Principles and Data Sources

The portfolio is a cross-section of $n = 5{,}000$ residential mortgage loans, each described by a vector $\mathbf{x}_i = (w_i, \mathrm{DTI}_i, \mathbf{z}_i)$ of observable characteristics and a pair $(e_i^{\mathrm{obs}}, e_i^{\mathrm{pot}})$ of latent occupational AI exposure indices, where $\mathbf{z}_i = (\mathrm{LTV}_i, \mathrm{perm}_i)$ collects additional borrower-level controls. LTV_i denotes the loan-to-value ratio at origination, measuring collateral coverage, and $\mathrm{perm}_i \in \{0, 1\}$ indicates whether the borrower holds a permanent employment contract. Four principles guide the construction.

(i) Observability partition. Variables in $\mathbf{x}_i$ are those that enter a standard bank scoring model and are jointly observed by the lender at origination. The pair $(e_i^{\mathrm{obs}}, e_i^{\mathrm{pot}})$ is measurable at the occupational level from public indices but is not used by banks: it implements Assumption 1 of the main text.

(ii) Sectoral channel. Borrower occupation is summarized by one of ten economic sectors indexed by $s \in \mathcal{S} = \{1, \ldots, 10\}$. Sectors jointly determine income, contract type, and both AI exposure indices. This parsimonious structure is sufficient to reproduce the AI exposure–income relationship documented in Section 2, while allowing sector-level presentation of the results.

(iii) Calibration to French aggregates. Wherever a parameter can be anchored on a publicly available French source; model-internal parameters that are not directly observable are set to match summary statistics of the French residential mortgage book. Table 2 summarizes the public sources used for calibration.

(iv) No free calibration on outcomes. Parameters of the stress equations (7)–(9) are set once based on independent sources, not on the portfolio distribution of $\mathrm{PD}_i^{(1)}$. This ensures that the Safe-Tail Paradox is inherited from the structure of the portfolio rather than engineered by parameter choice.

Table 2. Calibration sources. Each row lists the quantity used in the portfolio construction, the publicly available source from which it is taken, and the value used in the simulation.

Quantity	Source	Value used
Regulatory ceilings:		
Max debt-service ratio	HCSF (2021) Decision D-HCSF-2021-7	35%
Max loan maturity	HCSF (2021) Decision D-HCSF-2021-7	25 years
French residential mortgage book:		
12-month default rate (end-Q4 2023)	ACPR (2024), Housing financing in 2023, n° 160	0.44%
Doubtful loan ratio (31 December 2023)	ACPR (2024), Housing financing in 2023, n° 160	0.97%
Average loan principal (annual 2023)	ACPR (2024), Housing financing in 2023, n° 160	€196,800
Average LTV at origination (2023)	ACPR (2024), Housing financing in 2023, n° 160	78.8%
Share of outstanding covered by a guarantee	ACPR (2024), Housing financing in 2023, n° 160	97.0%
French labor market:		
Share of permanent contracts (CDI), 2023	INSEE (2024), Emploi, chômage, revenus du travail.	73.2%
Share of services-sector employment, 2024	INSEE (2025), Estimations d'emploi 2024	80.3%
Net monthly wage by sector, private sector, 2023	INSEE (2024) Première, no. 2020, Août 2024	Tab. 3
Net-to-gross conversion ratio, salaried employment	Standard French payroll convention	0.77
Unemployment-benefit replacement rate (λ^{UI})	UNEDIC (2026), Aide au Retour à l'Emploi (ARE)	57%
Basel III IRB framework - CRR2:		
Asset correlation, retail residential mortgage	Basel III CRR2 Art. 154(3)	$\rho = 0.15$
Confidence level (single-factor ASRF)	BCBS *Explanatory Note* (2005); CRR2	99.9%
Loss given default (uniform, illustrative)	Above CRR2 Art. 164 retail residential floor of 10%	LGD = 20%
AI exposure indices:		
Observed AI exposure e_s^{obs}	Massenkoff and McCrory (2026), March 2026 release	Tab. 3
Potential AI exposure e_s^{pot}	Eloundou et al. (2024)	Tab. 3

B.2 Sectoral Structure

Each borrower i is assigned to a sector $s_i \in \mathcal{S}$ with probability $p_s = \phi_s$, the portfolio weight reported in Table 3. The mapping from French aggregate employment shares to portfolio weights is calibrated to two moments. First, the portfolio reproduces the broad sectoral split between services and the secondary sector, with 80.1% of weight allocated to services against 80.3% in INSEE, 19.9% to industry and construction against 17.7%, and a deliberate omission of the agricultural sector (2.0% in INSEE), which is essentially absent from urban residential mortgage books.

Second, within services, the portfolio is tilted toward non-market services (Public Administration, Healthcare and Social, Education): these three sectors jointly receive 45% of portfolio weight, against an INSEE non-market services share of 30.4%, an over-weighting of 14.6 percentage points. Symmetrically, market services (Finance, Legal, Tech, Retail and Hospitality) receive 35% of portfolio weight against 50.0% in INSEE. The mechanism producing this tilt is the higher homeownership rate among salaried workers in stable employment, a positive selection documented by the HCSF (2019): French mortgage borrowers are concentrated in the most stable socio-professional categories, with 54% classified as *cadres* or *professions intermédiaires* versus 34% in the non-indebted active population, and they exhibit a 4.5% unemployment rate against 15.8% in the non-indebted active population, a 11.3 percentage point differential that closely matches the magnitude of the CDI tilt embedded in the calibration. The three over-weighted sectors all exhibit CDI shares above 0.86 (Public Administration 0.95, Educa-

tion 0.92, Healthcare 0.86), and the implied portfolio CDI share is 84.2%, approximately 11 percentage points above the 73.2% CDI share in the French active population (INSEE, 2024).[8]

Table 3. Sector calibration. The table displays the ten economic sectors used to generate borrower-level attributes. ϕ_s is the portfolio weight; $\bar{w}_s$ is gross annual mean income; π_s^{perm} is the share of permanent-contract (CDI) employment; e_s^{obs} and e_s^{pot} are the observed and potential AI exposure indices; $\delta_s^{\max}$ is the sector-level maximum displacement rate at $\delta = 1$. The severity cap $\delta_s^{\max}$ rescales the effective shock intensity at the borrower level.

s	Sector	ϕ_s	$\bar{w}_s$ (€)	π_s^{perm}	e_s^{obs}	e_s^{pot}	$\delta_s^{\max}$
1	Finance & Insurance	0.058	65,000	0.94	0.284	0.943	0.52
2	Legal & Consulting	0.048	61,000	0.87	0.204	0.890	0.48
3	Information & Tech	0.072	61,000	0.88	0.358	0.943	0.42
4	Public Administration	0.185	43,600	0.95	0.343	0.780	0.32
5	Healthcare & Social	0.180	41,300	0.86	0.130	0.520	0.19
6	Education & Library	0.085	39,700	0.92	0.182	0.617	0.17
7	Retail & Trade	0.125	34,300	0.72	0.269	0.500	0.12
8	Hospitality & Food	0.048	30,800	0.68	0.050	0.220	0.10
9	Manufacturing	0.130	46,900	0.83	0.080	0.240	0.09
10	Construction	0.069	38,000	0.78	0.030	0.120	0.05

Table 3 reports the gross annual mean income $\bar{w}_s$, the share of permanent-contract employment π_s^{perm}, the observed AI exposure e_s^{obs}, the potential exposure e_s^{pot}, and the sector-specific severity cap $\delta_s^{\max}$. Gross annual incomes are obtained from INSEE net monthly wages for 2023 via $\bar{w}_s = \bar{w}_s^{\text{net, mo}} \times 12/0.77$, where 0.77 is the standard French net-to-gross conversion ratio for private-sector salaried employment (INSEE, 2024).

B.3 Borrower-Level Attribute Generation

For each borrower $i = 1, \ldots, n$, sector assignment s_i is drawn from a multinomial distribution with probabilities proportional to the portfolio weights ϕ_s; all subsequent attributes are conditional on s_i. Let $U_{i,j}$ and $Z_{i,j}$ denote independent uniform $[0, 1]$ and standard normal draws.

Gross income is log-normally distributed around the sector mean $\bar{w}_{s_i}$, with a minimum value of 15,000 €

$$w_i = \max\left\{15{,}000,\ \bar{w}_{s_i} \cdot \exp(0.30 \cdot Z_{i,1})\right\}, \tag{B.1}$$

with dispersion calibrated to the within-sector coefficient of variation of INSEE private-sector wages. Contract type is drawn from a Bernoulli with success probability $\pi_{s_i}^{\text{perm}}$.

[8] The HCSF data are reported by socio-professional category rather than by economic sector and do not allow a direct verification of the 14.6 pp non-market-services tilt: within market services and within industry, sub-sectoral weights are illustrative rather than structurally identified, and the specific sectoral weights should be read as one plausible calibration consistent with the two moments above.

Loan size is a uniform multiple of income, $L_i = w_i \cdot (3.5 + 3.0 \cdot U_{i,1})$, consistent with the HCSF (2021) Decision D-HCSF-2021-7 imposing a 35% DTI cap and a 25-year maximum maturity at the 2024 average French mortgage rate of 3.5%, yielding a mean simulated principal of €221,000, close to the average loan amount of €196,800 (within 12%) reported in ACPR (2024).[9]

Property value is set to $V_i = L_i/[0.65 + 0.30 \cdot U_{i,2}]$, so that the loan-to-value ratio $\mathrm{LTV}_i = L_i/V_i$ is uniform on $[0.65, 0.95]$, with mean 0.80 consistent with the average LTV at origination of 78.8% reported by ACPR (2024).

The debt-service ratio is computed as

$$\mathrm{DTI}_i = \frac{0.0055\, L_i}{w_i/12}, \tag{B.2}$$

where $0.0055 \times L_i$ is the monthly debt-service payment and $w_i/12$ is monthly income. The debt-service ratio is truncated to $[0.12, 0.40]$: the lower threshold excludes implausibly low debt burdens and the upper threshold enforces a soft ceiling above the HCSF 35% cap, allowing the 20% of borrower files that the HCSF Decision authorises to exceed the cap to remain in the portfolio.

Occupational AI exposure indices are drawn as

$$e_i^{\mathrm{obs}} = e_{s_i}^{\mathrm{obs}} + 0.07 \cdot Z_{i,3}, \qquad e_i^{\mathrm{pot}} = e_{s_i}^{\mathrm{pot}} + 0.07 \cdot Z_{i,4}, \tag{B.3}$$

bounded to $[0.02, 0.99]$ and $[0.05, 0.99]$ respectively, with $e_i^{\mathrm{pot}} \geq e_i^{\mathrm{obs}} + 0.02$ enforced to preserve the definitional ordering between observed and potential exposure.

B.4 Scoring Model and Pre-shock PD

The bank's scoring model is calibrated following equation (1) of the main text as

$$\mathrm{PD}_i^{(0)} = \Lambda\big(\alpha + \beta_w\, w_i + \beta_{\mathrm{DTI}}\, \mathrm{DTI}_i + \beta_{\mathrm{LTV}}\, \mathrm{LTV}_i + \beta_{\mathrm{perm}}\, \mathbf{1}\{\mathrm{perm}_i\} + u_i\big), \tag{B.4}$$

with $\Lambda(\cdot)$ the logistic CDF. The exposure indices e_i^{obs} and e_i^{pot} enter neither this equation nor its estimation, consistent with Assumption 1. The scoring residual $u_i \sim \mathcal{N}(0, \sigma_u^2)$ with $\sigma_u = 0.50$ captures unobserved borrower heterogeneity not absorbed by the four observable covariates $(w_i, \mathrm{DTI}_i, \mathrm{LTV}_i, \mathbf{1}\{\mathrm{perm}_i\})$, generating within-rating-class dispersion of individual PDs around the class average. The calibrated parameters of equation (B.4) are reported in Table 4.

The portfolio exhibits the following origination summary statistics: mean gross annual income €46,534 (median €43,388); mean LTV 79.9%; mean DTI 32.8%; share of permanent contracts

[9] The loan-to-income multiplier is uniform on $[3.5, 6.5]$ with mean 5.0. The monthly debt-service payment on a loan of size L_i at 3.5% over 25 years equals $0.0055 \times L_i$, so the mean DTI is $0.0055 \times 12 \times 5.0 \approx 33\%$, compatible with the HCSF 35% ceiling.

Table 4. Scoring coefficients. Calibrated values of the logistic scoring model in equation (B.4). Signs are constrained by the main text. The intercept α implements a through-the-cycle (TTC) PD calibration following EBA (2017).

Coefficient	Value	Sign	Calibration rationale
α	-6.20	—	Set by bisection so that $\text{avg}_i\, \text{PD}_i^{(0)} \approx 0.77\%$, i.e. 1.75× the realised 12-month default rate of 0.44% (ACPR, 2024); a TTC buffer that reflects long-run averages including the 2008–2009 and 2020 stress periods.
β_w	-2.2×10^{-5} per €	< 0	At mean income $\bar{w} \approx e46{,}500$: $\beta_w \bar{w} = -1.02$ log-odds.
β_{DTI}	$+4.50$	> 0	Borrower at the HCSF 35% ceiling: $\beta_{\text{DTI}} \cdot 0.35 = +1.58$ log-odds.
β_{LTV}	$+1.70$	> 0	LTV 60% → 95%: $\beta_{\text{LTV}} \cdot 0.35 = +0.60$ log-odds.
β_{perm}	-0.90	< 0	CDI vs. CDD differential: odds ratio $\exp(-0.90) \approx 0.41$, i.e. default odds divided by ~ 2.4 for permanent-contract borrowers.
σ_u	0.50	—	Standard deviation of $u_i \sim \mathcal{N}(0, \sigma_u^2)$.

84.2%; mean $\text{PD}_i^{(0)}$ 0.77% (median 0.55%). Table 5 compares these moments line by line with the corresponding French residential mortgage book benchmarks reported by ACPR (2024), providing an external check on the joint distribution generated by the calibration.

Table 5. Validation: simulated portfolio versus ACPR 2023 benchmarks. Portfolio statistics at origination ($n = 5{,}000$) versus French residential mortgage book benchmarks at end-2023.

Statistic	Portfolio	ACPR 2023	Source
Mean LTV at origination	79.9%	78.8%	ACPR (2024), p. 3
Mean DTI / taux d'effort	32.8%	30.7%	ACPR (2024), p. 3
Mean loan principal	€221,000	€196,800	ACPR (2024), p. 3
Mean $\text{PD}^{(0)}$ vs. default rate	0.77%	0.44%	ACPR (2024), p. 3
Share at fixed rate	100%	99%	ACPR (2024), p. 3

Four of the five benchmarks line up within statistical tolerance: simulated mean LTV exceeds the ACPR figure by 1.1 percentage point, mean DTI by 2.1 percentage points, the share of fixed-rate loans by 1 percentage point, and the ratio between the simulated mean $\text{PD}^{(0)}$ and the realised default rate is 1.75×, in line with the TTC buffer described in Table 4. The single material deviation is the average loan principal, which exceeds the ACPR benchmark by 12%: this is a mechanical consequence of using the 25-year HCSF maximum maturity in the loan-size formula of Section B.4 rather than the 22.3-year observed average of ACPR (2024), and could be eliminated at the cost of weakening the link to the regulatory ceiling. The CDI share of the portfolio (84.2%) is not directly comparable to an ACPR benchmark, since the

report does not break out borrowers by contract type; it remains, however, consistent with the positive employment selection documented by the HCSF (2019), as discussed in Section B.3.

B.5 Rating Class Assignment

Borrowers are partitioned into $K = 7$ rating classes, reflecting a standard level of granularity in internal rating systems that balances discriminatory power and sample size within classes. Thresholds $0 < \theta_1 < \cdots < \theta_6 < 1$ are defined as empirical quantiles of the pre-shock PD distribution:

$$\theta_k = Q_{\mathrm{PD}^{(0)}}(q_k), \qquad (q_1, \ldots, q_6) = (0.20, 0.42, 0.63, 0.80, 0.92, 0.98), \tag{B.5}$$

so that rating class C_k is populated in the proportions $(20\%, 22\%, 21\%, 17\%, 12\%, 6\%, 2\%)$. These proportions approximate the observed distribution of S&P ratings for large European residential mortgage books.

The PD assigned to rating class C_k is the empirical within-class average of the model-implied individual PDs:

$$\bar{\pi}_k^{(0)} = \frac{1}{|C_k|} \sum_{i \in C_k} \mathrm{PD}_i^{(0)}, \qquad \bar{\pi}_k^{(1)}(\delta) = \frac{1}{|C_k|} \sum_{i \in C_k} \mathrm{PD}_i^{(1)}(\delta). \tag{B.6}$$

The class partition is fixed at origination on the basis of $\mathrm{PD}_i^{(0)}$ and is not updated by the stress: comparisons between $\bar{\pi}_k^{(0)}$ and $\bar{\pi}_k^{(1)}(\delta)$ are within-class and capture the average within-class revision of default probabilities induced by the AI shock.

B.6 Parameters of the Stress Equations

This section documents the calibration of equations (7)–(9). Each parameter is anchored on an independent empirical source.

Employment channel (eq. 8). The probability of displacement is given by

$$p_i^{\mathrm{emp}}(\delta) = \Lambda(\gamma_0 + \gamma_1\, e_i^{(1)}\tilde{\delta}_i + \gamma_2\, (e_i^{(1)}\tilde{\delta}_i)^2), \quad (\gamma_0, \gamma_1, \gamma_2) = (-3.2, 2.8, 1.6). \tag{B.8}$$

The three coefficients pin down the level, slope, and curvature of the displacement-risk function over the composite covariate $z_i \equiv e_i^{(1)}\tilde{\delta}_i \in [0, 1]$. They are jointly identified by three empirical anchors evaluated at three reference points of the composite-exposure grid: $z = 0$ (no shock), $z = 0.5$ (mid-scenario), and $z = 1$ (saturation).

Moment 1 (baseline at $z = 0$). The AI channel is inactive and p_i^{emp} collapses to the background rate of involuntary separations from French salaried employment, which the DARES estimates at approximately 3.9% per year, net of scheduled fixed-term contract endings. Solving $\Lambda(\gamma_0) = 0.039$ delivers $\gamma_0 = \ln(0.039/0.961) = -3.20$.

Moment 2 (mid-scenario at $z = 0.5$). The composite value $z = 0.5$ corresponds to a representative high-exposure worker under significant AI diffusion (e.g., top-quintile occupational exposure $e \approx 0.65$ combined with mid-transition severity $\tilde{\delta} \approx 0.77$). At this point we anchor $p_i^{\text{emp}} = 20\%$, a magnitude consistent with the -16% employment decline reported by Brynjolfsson et al. (2025) for $z \approx 0.45$ at the end of the 2022–2024 LLM diffusion window. The moment condition $\Lambda(\gamma_0 + 0.5\,\gamma_1 + 0.25\,\gamma_2) = 0.20$ yields, after substitution of γ_0, the linear constraint

$$0.5\,\gamma_1 + 0.25\,\gamma_2 = 1.80. \tag{B.9}$$

Moment 3 (saturation at $z = 1$). At maximal composite exposure, Eloundou et al. (2024) estimate that approximately 80% of the tasks performed by workers in the top exposure quintile are LLM-automatable; net of an allowance for non-substitutable tasks and labor-reallocation friction, the model imposes $p_i^{\text{emp}}(1) = 77\%$. The condition $\Lambda(\gamma_0 + \gamma_1 + \gamma_2) = 0.77$ yields the second linear constraint

$$\gamma_1 + \gamma_2 = 4.40. \tag{B.10}$$

The 2×2 linear system formed by (B.9)–(B.10) admits the unique solution $(\gamma_1, \gamma_2) = (2.80, 1.60)$, which together with $\gamma_0 = -3.20$ delivers the parameter triplet reported in (B.8). The calibrated curve reproduces the three target moments exactly: $\Lambda(-3.20) = 3.9\%$ at the baseline, $\Lambda(-1.40) = 19.8\%$ at $z = 0.5$, and $\Lambda(+1.20) = 76.9\%$ at saturation. The convexity ratio $\gamma_2/\gamma_1 \approx 0.57$ delivers the accelerating displacement risk in the exposure dimension: the marginal effect $\partial p^{\text{emp}}/\partial z$ rises from 0.11 at $z = 0$ to 0.70 at $z = 0.5$.

The path from the observed to the potential frontier is indexed by the global scenario severity $\delta \in [0, 1]$ and rescaled at the sector level by $\delta_s^{\max}$ (Tab. 3): $\delta = 0$ freezes exposure at the level observed in 2026, $\delta = 0.5$ corresponds to mid-transition, and $\delta = 1$ reaches the full-transition potential. The borrower's post-scenario occupational exposure is the linear interpolation $e_i^{(1)} = e_i^{\text{obs}} + \delta\,(e_i^{\text{pot}} - e_i^{\text{obs}})$ defined in the main text, and the borrower's effective severity $\tilde{\delta}_i$ translates the global δ into a sector-specific severity reflecting the maximum displacement rate $\delta_s^{\max} \in [0, 1]$ reachable in sector s at $\delta = 1$:

$$\tilde{\delta}_i = \delta \cdot \frac{\delta_{s_i}^{\max}}{e_{s_i}^{\text{pot}}}. \tag{B.11}$$

Normalisation by e_s^{pot} ensures that a borrower at maximum potential exposure in sector s reaches the sector-specific cap $z_i = \delta_s^{\max}$ at $\delta = 1$. Cognitive sectors carry $\delta_s^{\max} \in [0.42, 0.52]$ (Finance, Legal, Tech), whereas low-routine sectors carry $\delta_s^{\max} \leq 0.10$ (Manufacturing, Construction, Hospitality). At identical global severity, the effective severity reaching individual borrowers varies substantially with sectoral assignment.

Wage channel (eq. 9). The unemployment replacement rate is set to $\lambda^{\text{UI}} = 0.57$, corresponding to the floor of the ARE (Aide au Retour à l'Emploi) scheme administered by France

Travail. The wage compression parameter is calibrated at $\mu = 0.70$, reflecting the upper bound of task substitutability in highly exposed occupations as documented by Eloundou et al. (2024). Idiosyncratic wage dispersion is modeled as $\sigma_\varepsilon = 0.03 \cdot w_i^{(0)}$, capturing moderate heterogeneity in individual outcomes.

The exposure threshold $\bar{e} = 0.35$ separates the substitution regime, in which AI compresses wages of highly exposed workers, from the augmentation regime, in which AI raises the productivity of low-exposed workers. The threshold $\bar{e} = 0.35$ separates occupations with substantial observed AI delegation from the rest. In the unconditional occupational distribution it sits at the 94th percentile; conditional on the 45% of occupations with positive observed exposure, it sits at the 85th percentile, just above the 8th-decile boundary at 0.29. The choice is illustrative and operationalises the cognitive–routine boundary discussed in the broader task-automation literature (Autor et al., 2003; Acemoglu and Restrepo, 2022), rather than being directly anchored to a specific quantile.

Secondary adjustments. Post-shock debt-service-to-income and loan-to-value are updated mechanically:

$$\mathrm{DTI}_i^{(1)} = \min\{0.65,\ \mathrm{DTI}_i^{(0)} \cdot (w_i^{(0)}/w_i^{(1)})\}, \tag{B.12}$$

$$\mathrm{LTV}_i^{(1)} = \min\{0.98,\ \mathrm{LTV}_i^{(0)} \cdot (1 + 0.02 \cdot \delta \cdot \max\{e_i^{(1)} - \bar{e}, 0\})\}. \tag{B.13}$$

Substituting $(w_i^{(1)}, \mathrm{DTI}_i^{(1)}, \mathrm{LTV}_i^{(1)}, \mathrm{perm}_i^{(1)})$ into (B.4) produces $\mathrm{PD}_i^{(1)}(\delta)$. Borrowers who lose their job have $\mathrm{perm}_i^{(1)} = 0$.

B.7 Regulatory Capital

The IRB capital charge per unit of exposure at default at PD level π is the single-factor Vasicek–Gordy formula prescribed by CRR2 Art. 154 for retail residential mortgages,

$$\kappa(\pi) = \mathrm{LGD} \cdot \left[\Phi\left(\frac{\Phi^{-1}(\pi) + \sqrt{\rho}\,\Phi^{-1}(0.999)}{\sqrt{1-\rho}} \right) - \pi \right], \tag{B.14}$$

where Φ denotes the standard normal cumulative distribution function, $\rho = 0.15$ is the asset correlation prescribed by CRR2 Art. 154 for retail residential exposures, and 0.999 is the IRB confidence level. The maturity adjustment of the corporate IRB formula does not apply to retail exposures, and the Basel II scaling factor of 1.06 has been removed under Basel III. The loss-given-default is set to LGD = 20% uniformly across borrowers, sitting between the Basel III regulatory floor for retail residential mortgages ($\mathrm{LGD}_{\min} = 10\%$ under CRR2 Art. 164 prior to the 2025 output-floor phase-in) and the realised LGD magnitudes reported in ACPR (2024) for uncollateralised recovery scenarios. The uniform specification is adopted for transparency: none of the results reported in Section 5 depend on a borrower-level LGD function, and using a flat LGD ensures that variation in κ across rating classes and severity levels reflects variation in PD alone.

The class-level capital charge is computed using the class-average empirical PDs defined in (B.6): $\kappa(\bar{\pi}_k^{(0)})$ for the pre-shock capital and $\kappa(\bar{\pi}_k^{(1)}(\delta))$ for the post-shock capital. The aggregate charges K_0 and $K_1(\delta)$ are obtained by EAD-weighting per equations (16)–(17) of the main text. Under the present calibration, $\kappa(\bar{\pi}_k^{(0)})$ ranges from approximately 0.67% of EAD in class AAA ($\bar{\pi}_{\text{AAA}}^{(0)} = 0.21\%$) to 4.44% of EAD in class CCC ($\bar{\pi}_{\text{CCC}}^{(0)} = 3.64\%$), both values in the range typical of retail residential mortgage books under IRB.

B.8 Validation: Empirical Check of Proposition 1

Table 6 reports class-level pre-shock statistics. The class-average AI exposure $\bar{e}_k^{\text{obs}}$ is weakly decreasing in k (with a tie between BBB and BB at 19.5%), from 23.0% in AAA to 17.6% in CCC, while $\bar{\pi}_k^{(0)}$ increases strictly. The sample correlation between e_i^{obs} and $\text{PD}_i^{(0)}$ is -0.09. The nonparametric monotonicity test of Ghosal et al. (2000), applied with the class index as the running variable (so that the alternative becomes $\bar{e}_k^{\text{obs}}$ strictly decreasing in k), rejects the null of a non-decreasing relationship at the 5% level.

Table 6. Empirical pre-shock statistics by rating class. n_k: number of loans; $\bar{\pi}_k^{(0)}$: mean pre-shock PD; $\bar{e}_k^{\text{obs}}$: mean observed AI exposure; $\bar{w}_k$: mean gross annual income; LTV_k, DTI_k: mean loan-to-value and debt-to-income ratios.

Class	n_k	$\bar{\pi}_k^{(0)}$	$\bar{e}_k^{\text{obs}}$	$\bar{w}_k$ (€)	LTV_k	DTI_k
AAA	1,000	0.213%	23.0%	61,352	0.77	0.27
AA	1,100	0.399%	21.6%	48,796	0.79	0.31
A	1,050	0.606%	19.6%	44,184	0.80	0.33
BBB	850	0.880%	19.5%	40,141	0.81	0.34
BB	600	1.297%	19.5%	38,084	0.82	0.36
B	300	2.062%	18.3%	36,439	0.83	0.38
CCC	100	3.635%	17.6%	33,469	0.84	0.39

B.9 Limitations

Three limitations deserve mention. First, the portfolio abstracts from interest-rate heterogeneity and loan vintage: all loans are priced at an implicit 3.5% rate corresponding to the Q1 2024 French market, and vintage effects could be added without altering the mechanism. Second, the sectoral aggregation combines occupations with heterogeneous exposures: within-sector dispersion is captured by $\sigma_e = 0.07$, and a finer granularity would only strengthen the within-class concentration of risk. Third, the stress equations are calibrated to French labor market benchmarks and should be read as illustrative of the order of magnitude rather than structural estimates. None of these affects the direction of the results in Section 5: the Safe-Tail Paradox is a feature of the joint distribution of income and AI exposure across occupations, reproduced by any calibration consistent with Assumptions 1–2.