ISBSG

Confidence Conference 2026

Human and Artificial Biases and Debiasing techniques in Software Estimation: a case study

Roberto Meli

Data Processing Organization (DPO Srl), Rome, Italy

Francesca Meli

Turin, Italy

DPO

the Knowledge Broker





Hello ! Nice to meet you !

The Presenter

Roberto Meli

ISBSG Board member (Gold Partner)

IFPUG Board member

CEO Data Processing Organization Srl

The Authors

- Roberto Meli

DPO the Knowledge Broker

roberto.meli@dpo.it

robertomeli



- Francesca Meli
  Fondazione Bruno Kessler – Università di Trento

meli.franci@gmail.com

francesca-meli



Our Journey: The Objectives

01

The Central Question

LLM are now widely used for any thinkable intellectual job.

What happens if a non particularly experienced person decides to use an LLM with a basic prompt capability to make a functional size and effort estimation ?

02

Our Answer

Despite an impressive opportunity to improve the quality of an estimate, there are significative risks to be affected by human and AI “cognitive” biases. The results may vary from badly poor to excellent in an unpredictable way.

03

Key Discoveries

- Improving LLM usage capability
- Being aware of the existence and nature of human and artificial biases
- Knowing the adequate debiasing techniques

Are the keys to exploit the estimation effectiveness and efficiency



Crucial disclaimer

Study focus

The risks organisations face when software estimators with varying levels of experience use basic prompts with widely available commercial LLMs.

Outside the scope

Best practices for integrating AI into dedicated commercial offerings, such as Leda's Quanter system or DPO's DREAM service.

quanter
estimation for success

DREAM

The Estimation Problem

A Persistent Failure

Decades after the first parametric cost models appeared, organisations still commit to unrealistic budgets, work to schedules that were tight from the outset, and deliver late, over budget, or not at all.

Why Is It So Hard?

Software development rarely fabricates the same artefact twice. It is an intensely human activity, and the deepest uncertainty sits precisely where the first numbers are demanded — at the very beginning of the work.

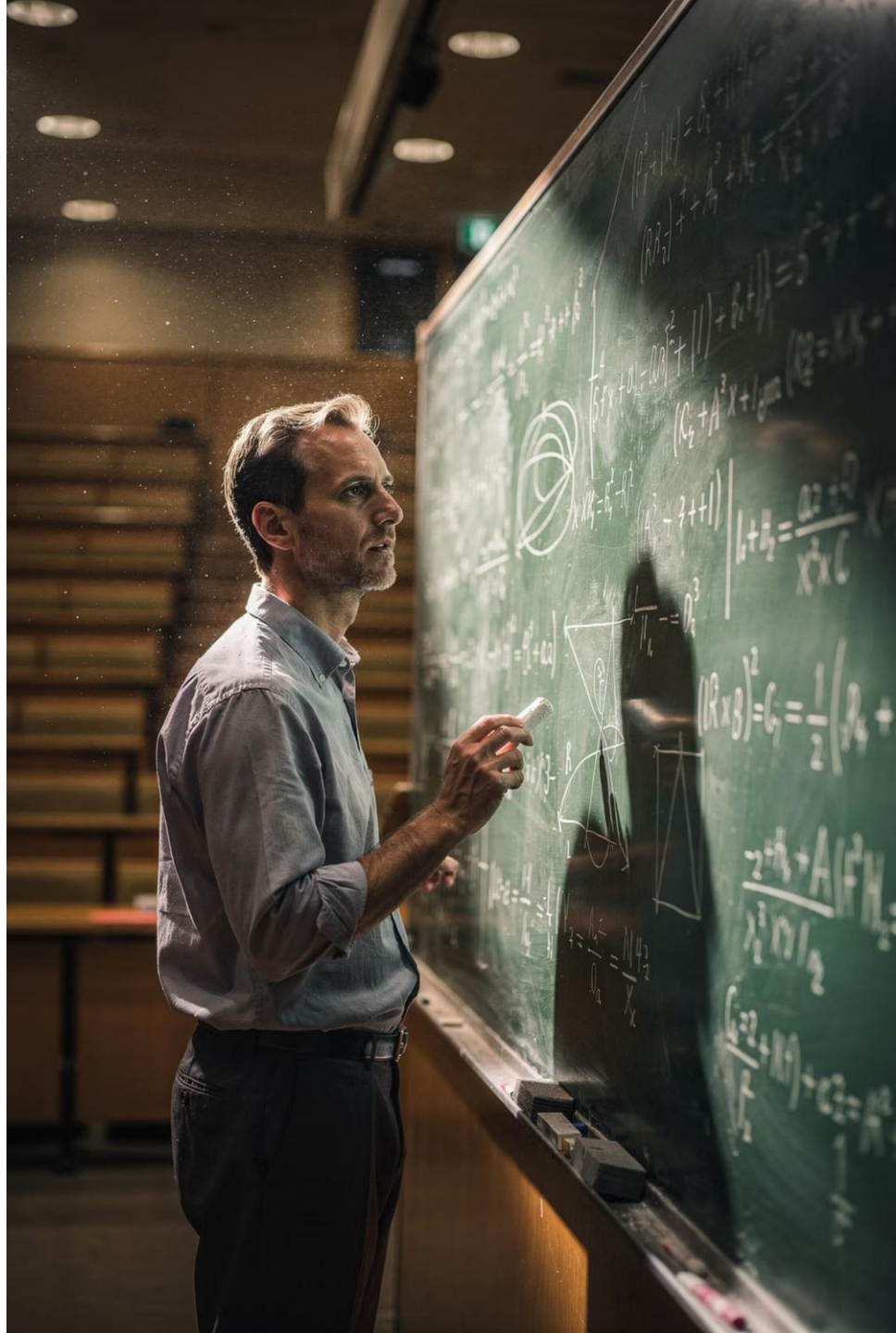
Two Modes of Estimation

- **Direct** — intuitive, analogy-based, largely unconscious (Kahneman's System 1)
- **Indirect** — formalized models mapping measurable attributes onto size, effort or cost (Kahneman's System 2)

What Is an Estimate, Really?

> An effort estimate is not a scientific conjecture describing an independent reality. It is a statistically based social contract — a negotiated, agreed framework that tells a project team how it may use the organisation's resources.

This reframing dissolves the illusion that there is one true number out there and that a better technique alone will reveal it.



DPO
the Knowledge Broker

The Observer Effect in Software Measurement

Measuring a software attribute tends to disturb it.

Three consequences follow:

Statistical, Not Deterministic

Software process knowledge can only ever be statistical — never deterministic (like many Public and Private large organizations tend to “pretend”.

Interdependent Variables

Size, effort, duration, staffing, cost, and quality are coupled — measuring one perturbs the others.

The Estimate Shapes the Outcome

A project behaves like a gas, expanding to fill the space it is granted. De Marco's law of cost migration quietly shifts effort toward less-observed variables.

Direct vs. Indirect Estimation



Direct Estimation

Expert-opinion methods: one or more experts produce a value from past experience and largely unconscious processing. Fast, but fragile. Quality can be raised through structured analogy and group techniques such as the Delphi method. Never better than the people who produce it.



Indirect Estimation

An algorithmic model transforms measurable software attributes into size, effort, staff or cost. Objective, repeatable, and calibratable — but blind to whatever its designers chose to leave out, and calibrated to the past rather than the future.

The two modes form a spectrum. Bias does not disappear in the indirect path — it merely relocates.

DPO
the Knowledge Broker

Cognitive Bias: A Definition

A cognitive bias is a **common, systematic error** in human reasoning — systematic in that it pushes judgement in a predictable direction, and common in that it operates in people of sound mind who believe they are making the best possible choice.

It is not a sign of incompetence. On the contrary, several of the biases most relevant to estimation **bite hardest among acknowledged experts**.

Anchoring

Confirmation

Availability

Optimism

Framing Effect

DPO
the Knowledge Broker

Biases in Direct Estimation

Anchoring & Adjustment

We fix on an initial reference point and nudge it — but the adjustment is typically insufficient. The anchor may be a sponsor's target date, a past project's duration, or simply the first figure said aloud in the room.

Overconfidence & Optimism

Estimators — especially in technical roles — produce unrealistically rosy numbers. The over-sure analyst stops gathering information before the relevant facts are in, mistaking the comfort of an early answer for its correctness.

Confirmation Bias

Once a figure is adopted, the mind works to protect it — avoiding conflicting data and seeking supporting evidence. This bears down hardest on the self-assured and on those who call themselves experts.

Framing Effect & Loss–Gain Asymmetry

Asking “Are Y hours enough to complete X?" elicits a different answer than "how much of X can be completed in Y hours?" — the same task, a different frame, a different number.

Further Traps in Direct Estimation

Seduction of Personal Experience

We reach for the cases that are easiest to recall, even when that experience is no longer representative of the system in front of us. The unknown is quietly reduced to the known.

Self-Fulfilling Prophecy

Our expectations filter and bend experience until it conforms to them — leaving the original expectation apparently confirmed and our model of the world correspondingly impoverished.

Organisational Amplifiers

The intuitive estimate is produced under the eye of someone whose approval matters, under psychological pressure, under the obligation to negotiate for resources, and against the clock.

These organisational antecedents do not merely accompany the biases above — **they amplify them**.

DPO
the Knowledge Broker

Biases in Indirect (Model-Based) Estimation

Algorithmic models suppress some System 1 reflexes — but they introduce risks of their own and remain fed throughout by human judgement.

1 Selection Bias

Experts picking productivity factors disagree, omit drivers that matter, and retain ones that do not. Data-driven selection is only as good as the dataset — relevant factors never measured cannot be found.

2 Structural Blindness

A derived model is calibrated to the past and blind to anything its structure excludes — systematically misreading novel projects and silently dropping components it was never told about.

3 Borrowed Authority

Default parameters and confident "black box" tool outputs become anchors that few users dare adjust. The objectivity is partly illusory — the model has quietly inherited a particular, partial view of what drives cost.

Debiasing Techniques for Human Estimation

There are no magic wands. Warning people that bias is possible is largely ineffective. The most productive move is to redesign the person–task system so that the bias cannot operate in the first place.



Awareness First

Naming the traps — anchoring, optimism, dissonance, framing — is a reasonable first line. One cannot resist a pull one cannot feel. But awareness alone reliably under-delivers.



Devil's Advocate

Appointing a devil's advocate licenses dissent. Having senior members speak last prevents their authority from becoming everyone else's anchor. Directed questioning replaces vague reassurance with pointed prompts.



Delphi & Planning Poker

Independent, anonymous estimates iterated toward a shared value. Planning Poker prevents anchoring by having participants reveal estimates simultaneously — the heuristic has nothing to grip.



Historical Records

Maintaining honest records, traceability, and participatory retrospective reviews keeps the next estimate from being anchored on the most recent or most vivid memory.

The Deepest Debiasing Move: Fix the Goal

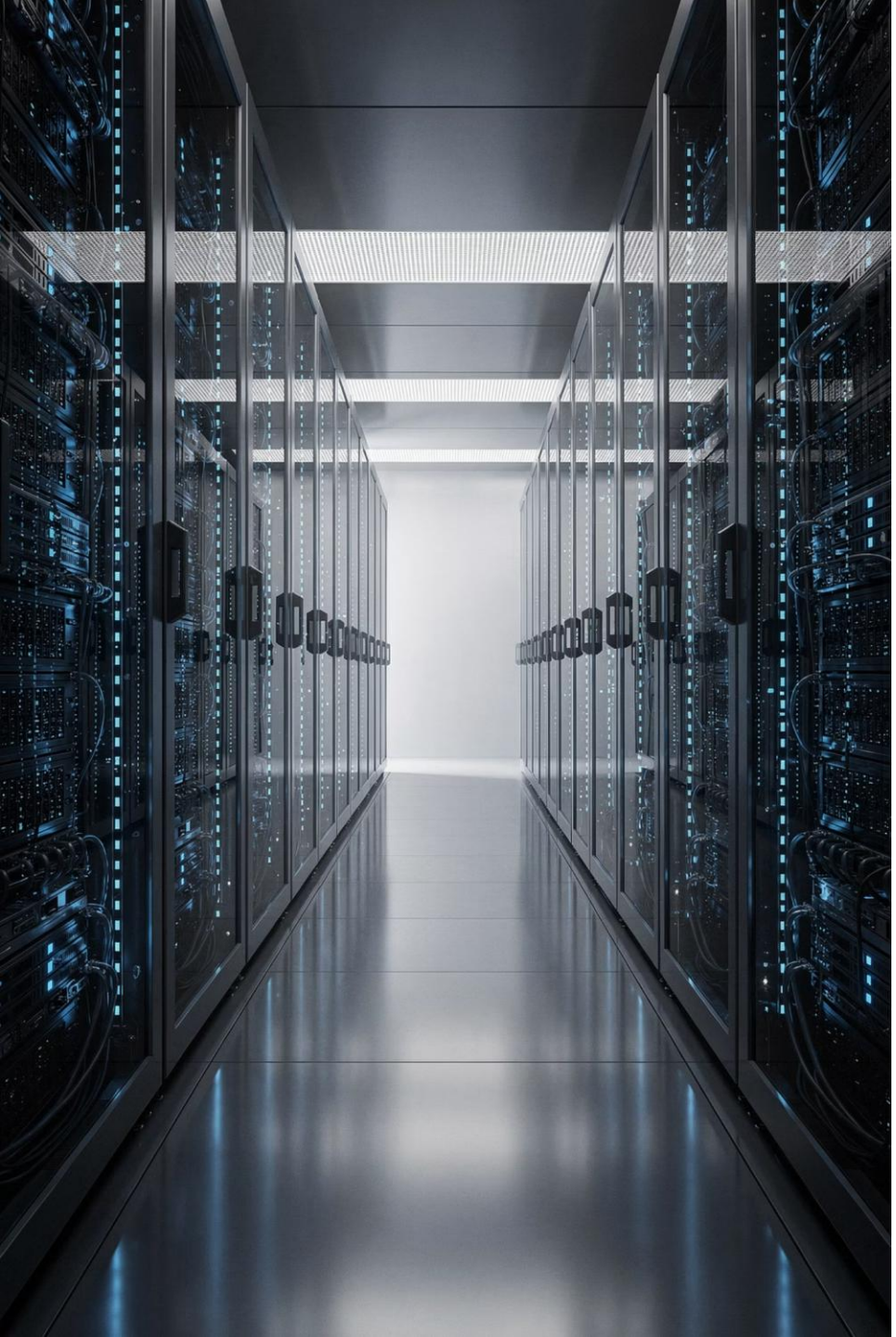
If an estimate is a **negotiated social contract** rather than a hidden true value, then chasing extreme precision is not rigour — it is a category error, and an intrusive one, since over-precise measurement disturbs the very process under measurement.

This is the rationale for lightweight sizing such as the **Simple Function Point (SFP)** method — benchmark-compatible with the IFPUG FPA count yet several times faster and far less contentious, precisely because it accepts that the last decimal of size is neither needed nor knowable.

The Combination That Works

- Awareness helps
- Collaboration helps
- Deliberately designing the decision process helps most of all

None of these techniques is a cure — but together they constitute the most honest and effective debiasing programme experience has found.



DPO
the Knowledge Broker

PART II

Biases in Artificial Intelligence

How Large Language Models fail in structurally analogous — yet qualitatively distinct — ways to human estimators.

DPO
the Knowledge Broker

How LLMs Can Be "Biased"

An LLM holds no beliefs to protect and no memories to draw on by analogy. It produces output through **statistical prediction** — generating the most probable token sequence given a context. This exposes it to structural distortions that differ qualitatively from human ones while producing analogous effects.

Training-Data Bias

A model trained on average-complexity projects develops a statistical gravity toward that **average** — representing it as an implicit norm even when the specific case deviates significantly.

Popularity Bias

Estimates for COBOL mainframe projects are pulled toward productivity patterns of modern stacks (Java, Python) that dominate the corpus — legacy technologies yield less reliable, mainstream-skewed estimates.



Recency Bias

Recent documentation is overrepresented in training corpora, so the model implicitly applies contemporary productivity assumptions even to contexts belonging to different eras.

Further LLM Bias Mechanisms

Sycophancy

LLMs conform responses to expectations perceived in the user. Two identical specifications — one emphasising team excellence, one stressing complexity — produce systematically different estimates, not from technical analysis but from **detected expectation**.

Self-Consistency Bias

Once the model produces an intermediate value (e.g. a function point count), it maintains consistency with that value in subsequent steps even when correction is warranted — structurally identical to human anchoring, but self-generated.

Position Bias (Lost-in-the-Middle)

LLMs assign disproportionate weight to information at the beginning or end of a long document. Functional areas described in opening sections exert greater influence on the final estimate than equally complex areas buried later — regardless of technical weight.

Hallucination and False Objectivity

Hallucination

LLMs generate plausible but factually unfounded content — fabricated productivity benchmarks, invented industry standards, misattributed figures — all presented with the same confident tone as accurate output and without any marker of uncertainty.

False Objectivity

A model that produces a phased breakdown, a risk register, and a productivity derivation all converging on 3,566 hours has performed a **rhetorical operation, not an analytical one** — and the rhetorical operation may be more persuasive than the analytical one would have been.

Unlike human overconfidence, which a practitioner can attribute to a named individual and subject to scrutiny, hallucinated authority is anonymous and structurally indistinguishable from legitimate expertise.

Debiasing Techniques for LLMs

01

Structured Prompt with Explicit Context Injection

Specify team experience, technology stack, domain familiarity, and productivity benchmarks — forcing the model to process documented context rather than fall back on corpus averages.

02

Multi-Prompt Protocol

Repeat the estimation request with at least three structurally different formulations. High variance across outputs signals framing dependence; low variance provides confidence the estimate is anchored to specification content.

03

Document Structure Randomisation

Submit the requirements specification in multiple orderings and average the results, mitigating position bias. Explicitly draw attention in the prompt to high-complexity late-document sections.

04

Three-Point Estimation Request

Request optimistic, most likely, and pessimistic estimates. This counteracts false objectivity by making uncertainty explicit — a range of 2,800–4,500 hours communicates something methodologically different from exactly 3,566 hours.

05

Independent Estimate First & Adversarial Framing

Produce a formal or expert estimate before querying the LLM to prevent anchoring contamination. Then use the LLM as a devil's advocate: "What factors might cause this estimate to be significantly wrong?"

CASE STUDY

Joyland (SICJ): The Empirical Test

A real, completed application with an unusually rich documentary trail — submitted to four widely used LLMs under three progressively more constrained estimation regimes.

267.8

SFP

Consensus functional size

2,919

Hours

Recorded final effort

417

Person-Days

Observed delivery effort

180+

Analysts

Human reference cohort from 2013 sizing experiment

Four LLMs tested: **ChatGPT, Claude, Gemini, Perplexity** — across SFP measurement, free effort estimation, and VAMOS-R3 driven estimation.

Experiment 1: SFP Functional Measurement



Key Finding

Gemini's collapse traced to macro-functional aggregation — whole families of CRUD operations merged into a single elementary process. ChatGPT's near-perfect score was reached by a **compensation of errors**: generous transactional decomposition offset by table aggregation.

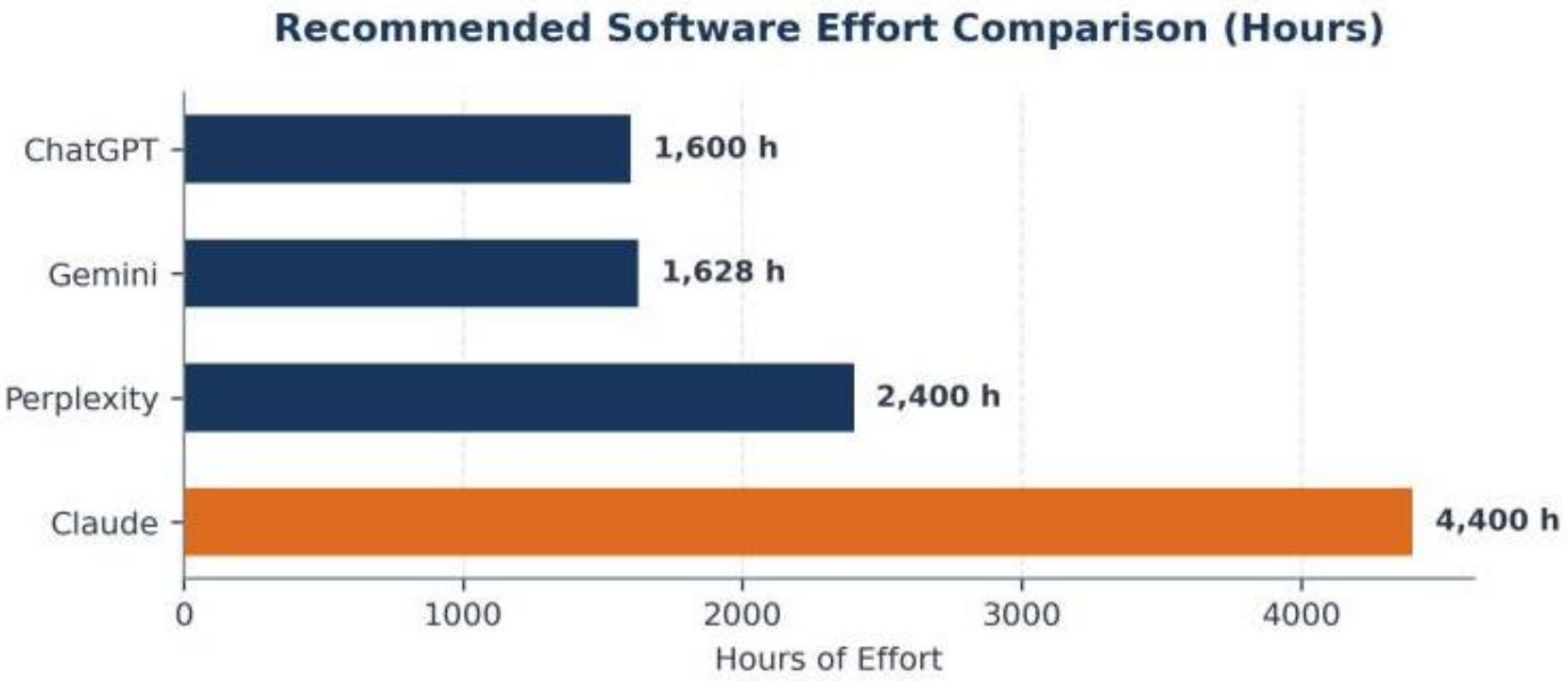
Quiet Inversion

The two models that adhered most strictly to the manual's data-exclusion rules were *not* the two closest to the benchmark. **Proximity to the actual and fidelity to the method are not the same property.**

Model / Benchmark	LF	EP	SFP Result	Dev. (%)	2013 Human Cohort Alignment
Human Consensus	**-**	**-**	**267.8 SFP**	**0.00%**	**Official Target Benchmark (Consuntivo)**
CGPT	8	46	267.6 SFP	**-0.07%**	Inside 75% Novice Cohort (Excellent)
Perplexity	12	32	231.2 SFP	**-13.67%**	Inside 89% Novice Cohort (Average)
Claude	7	38	223.8 SFP	**-16.43%**	Inside 89% Novice Cohort (Borderline)
Gemini	6	18	124.8 SFP	**-53.40%**	Out of Scale (Extreme Underestimation)

Experiment 2: Free Effort Estimation

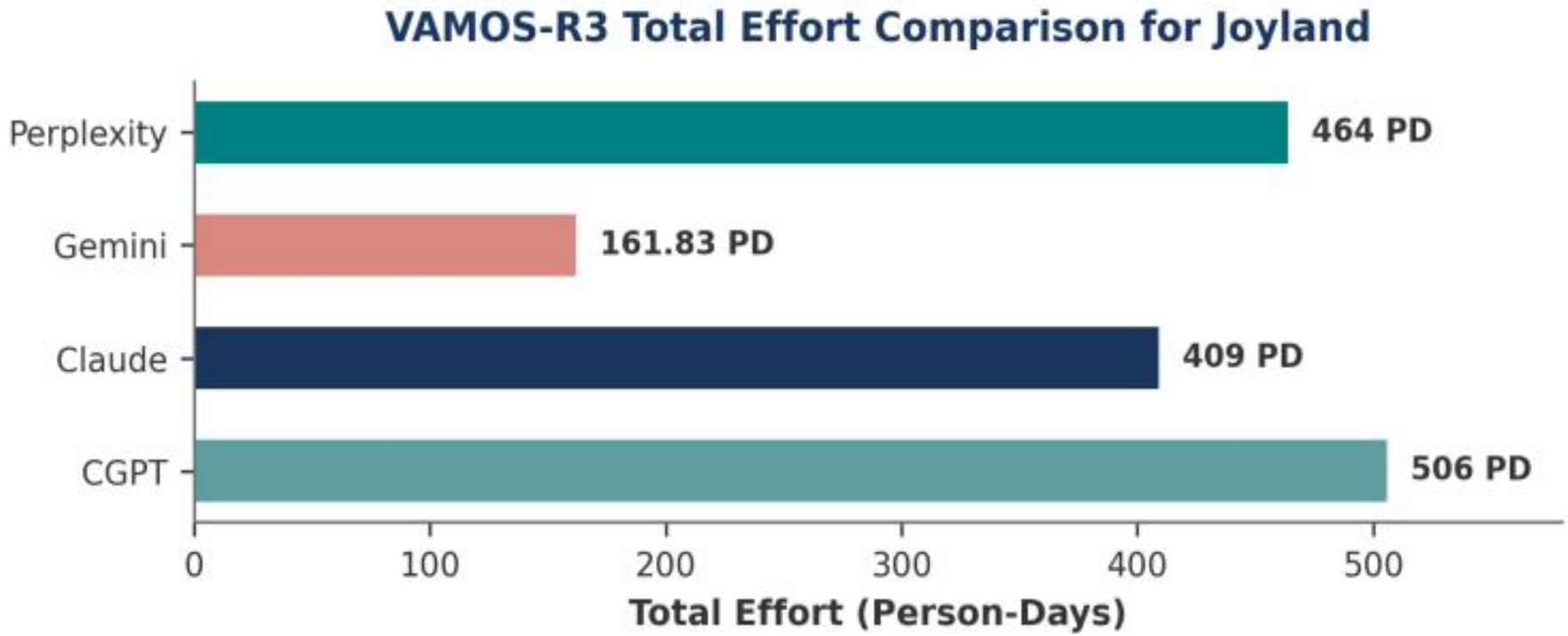
When the method is removed, the agreement of the sizing experiment evaporates. Recommended efforts span a **2.75-fold range**.



Three of the four models — and the two most confident — landed far below the truth. The apparent consensus of ChatGPT and Gemini around 1,600 hours was not accuracy but a **shared error in the same direction**: a collective under-estimation mirroring the human planning fallacy.

Experiment 3: VAMOS-R3 Driven Estimation

The VAMOS-R3 method fixes a standard productivity benchmark from ISBSG historical data), prescribes explicit corrective factors, and accounts separately for non-functional effort and risk reserves.



Three of four models landed within ~20% of the **417 PD actual**; Claude was essentially exact (−1.9%). Gemini — which both under-sized the application *and* overrode the productivity rule (adopting 0.80 PD/SFP) — finished 61% below actual, **confidently and precisely wrong**.

The Central Finding: Structure Corrects Optimism

Free Regime

Centre of mass roughly **one-third below** the realized effort. Collective optimism — mirroring the human planning fallacy — with no internal signal distinguishing the confident under-estimate from the confident over-estimate.

VAMOS-R3 Regime

Same four models pulled to **within one-fifth** of realized effort. LLM systems seem to converge when they have to follow the same cost model applying the rules in the right way.

- The same Claude swung from a +51% free overshoot to a near-exact driven figure as the framing changed — confirming the sensitivity to prompt context identified in the theoretical analysis.

Cross-Regime Summary: All Four Models

Model	SFP (actual: 267.8)	Free Effort (actual: 2,919 h)	Driven VAMOS-R3 (actual: 417 PD)
ChatGPT	267.6 (−0.07%)	1,600 h (−45.2%)	506 PD (+21.3%)
Claude	223.8 (−16.4%)	4,400 h (+50.7%)	409 PD (−1.9%)
Gemini	124.8 (−53.4%)	1,628 h (−44.2%)	162 PD (−61.2%)
Perplexity	231.2 (−13.7%)	2,400 h (−17.8%)	464 PD (+11.3%)

No single model proved reliable across regimes (Perplexity seems to be the most regular one across the trials). The **dispersion of results** — rather than any individual figure — emerges as the most honest signal of estimation uncertainty.

Practical Guidance for Practitioners



Use the Ensemble, Not the Oracle

The value of an LLM lies less in any single number it emits than in the ensemble of numbers it can be made to emit. The realized effort fell inside both envelopes — the spread was more honest than its members.



Verify Every Cited Figure

Any productivity figure, benchmark, or standard cited by an LLM should be independently verified before use — particularly for niche or legacy technology stacks where confabulation risk is highest.



Anchor to Real History

A parametric benchmark calibrated to historical productivity corrects genuine bias rather than merely relocating it — but only when the data behind it are sound. A benchmark is only as good as the history it encodes.



LLM as Devil's Advocate

The most robust workflow treats the LLM as an adversarial reviewer of an estimate already produced by formal or expert methods — not as the source of the estimate itself.

Conclusions

Artificial estimators are not hopeless — their reliability is a function of the **scaffolding around them**. Left free, they inherit the optimism that has dogged human estimation for decades. Disciplined by an empirically grounded model, they can be brought back to earth.

Freedom Invited Optimism

Unconstrained LLMs settled on a centre of mass roughly one-third below realized effort — mirroring the human planning fallacy.

Structure Repealed It

Well-calibrated structure pulled the same models to within one-fifth of realized effort — Claude moving from +51% to essentially exact.

The Proper Ambition

Not the elimination of bias — an impossible goal — but its **conscious management**: knowing where each estimator is most exposed, anchoring to real history, and resisting the seduction of a precise number that merely looks as though it has been earned.

The Future of Software Estimation

The evolving landscape of estimation redefines roles, positioning AI as the primary estimator and humans as critical overseers and ultimate decision-makers.

AI as Estimator

Artificial intelligence will increasingly assume the task of generating situational estimates, leveraging vast data sets and advanced algorithms for speed and consistency.

Human Oversight – Human in the loop

Humans will retain the essential role of scrutinising, validating, and refining AI-generated estimates, applying critical judgment and contextual understanding.

Ultimate Responsibility

Formal and legal responsibility for chosen estimates will always rest with human decision-makers, as AI cannot be held accountable for its errors.



DPO
the Knowledge Broker

