

A Scientific Framework for Using LLMs as Measurement Instruments in Social Science

Ana Pastor-Merino, Xavier Martínez-Barbero, Pablo de Pedraza, Josep Domenech

eSMART Research Center, Universitat Politècnica de València

King's Business School, London - May 2026

20 min presentation + 10 min Q&A



A 20-minute roadmap

The talk moves from the measurement problem to a concrete framework and protocol.

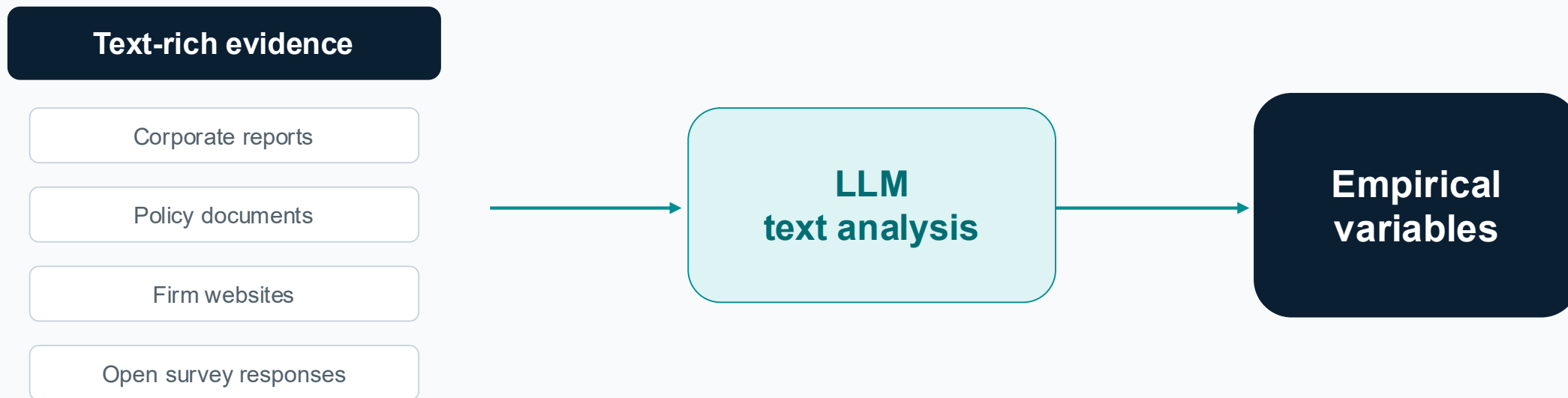


Main question

How can LLM-based text outputs be designed, evaluated and calibrated so that they can be interpreted as empirical measures?

Text is rich evidence - but not yet a measure

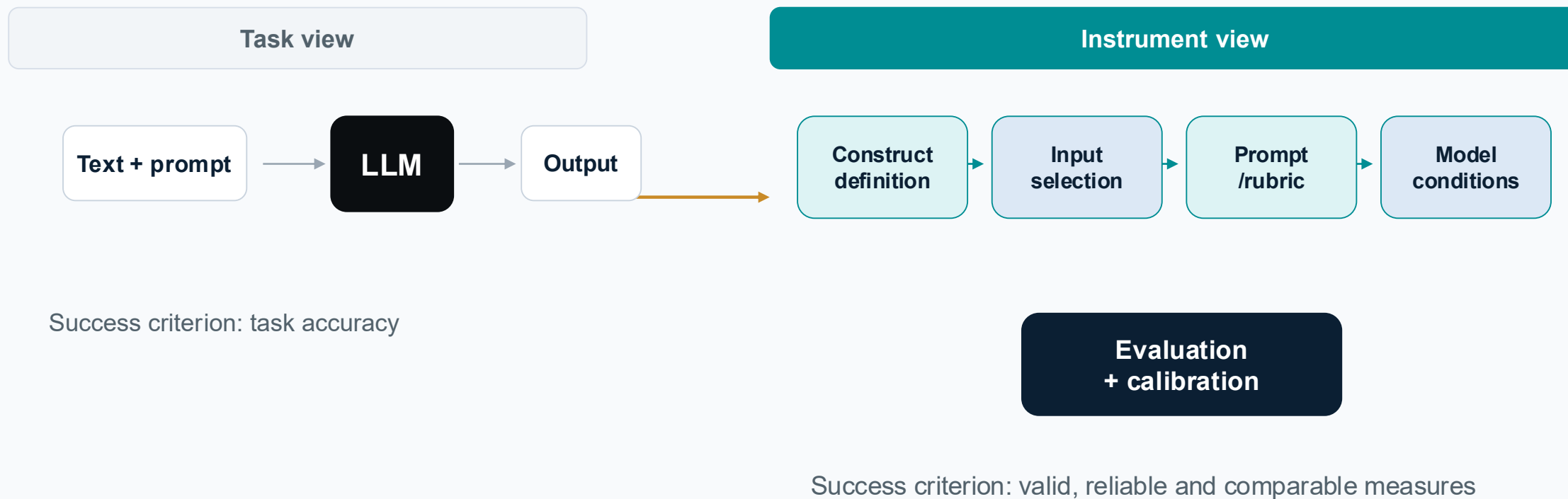
Unstructured documents can inform hard-to-observe constructs, if the transformation is standardized.



The gap: many LLM workflows optimize task performance, but do not specify the measurement system behind the output.

The shift: from task execution to measurement design

Model outputs are not direct observations; they are shaped by design choices.



Measurement systems decompose decisions and error

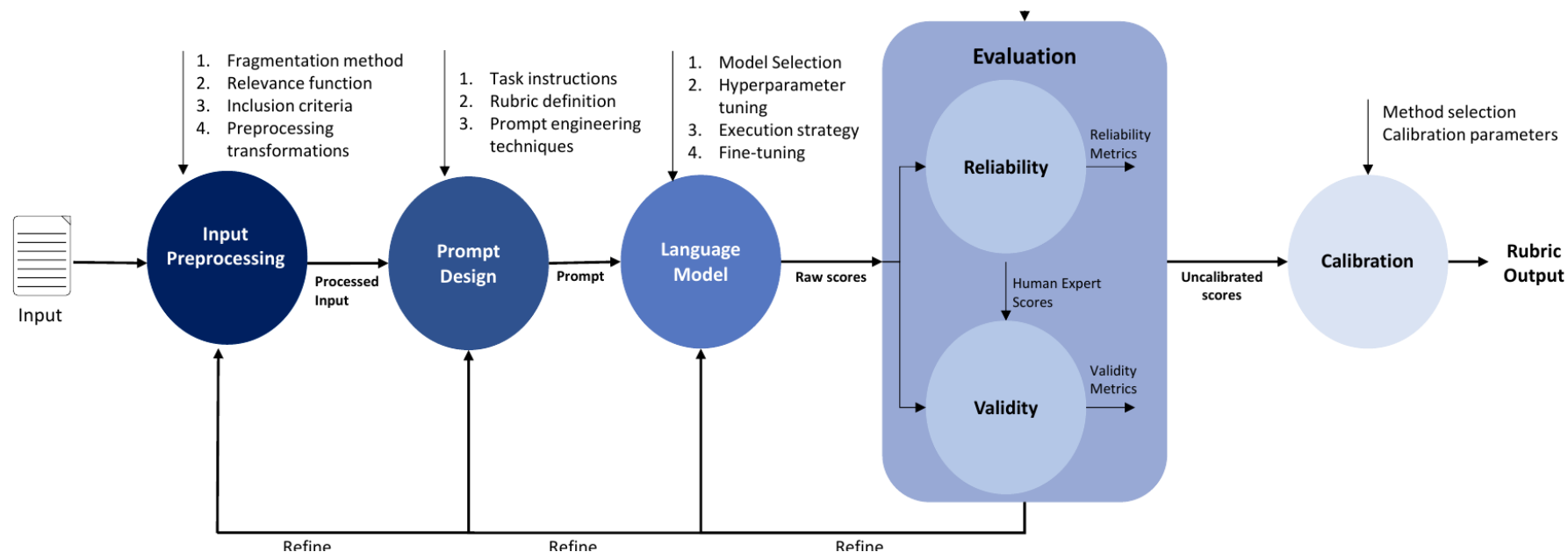
The framework borrows from survey methodology, psychometrics and content analysis.

Tradition	Evidence	Instrument	Implementation	Quality	Adjustment
Established measurement	Data processing Coding units	Questionnaire Rubric	Mode / scorer conditions	Reliability Validity	Calibration Scaling
LLM-based measurement	Input preprocessing	Prompt + rubric	Model execution	Evaluation metrics	Calibration parameters

Core implication: quality is an emergent property of the full pipeline - not of one model call.

A five-stage framework for LLM-based measurement

A modular pipeline for transforming text into valid, reliable and interpretable empirical variables.



The LLM is one component of the measurement system - not the measurement system itself.

Stages 1-2: define the evidence and the instrument

Before the model runs, the construct must be operationalized and the input must be curated.

Stage 1

Input preprocessing

Fragmentation
method

Relevance
function

Inclusion
criteria

Transformations
/ tagging

Output: curated evidence aligned with the
construct

Stage 2

Prompt design

Task
instructions

Rubric
definition

Selected
input

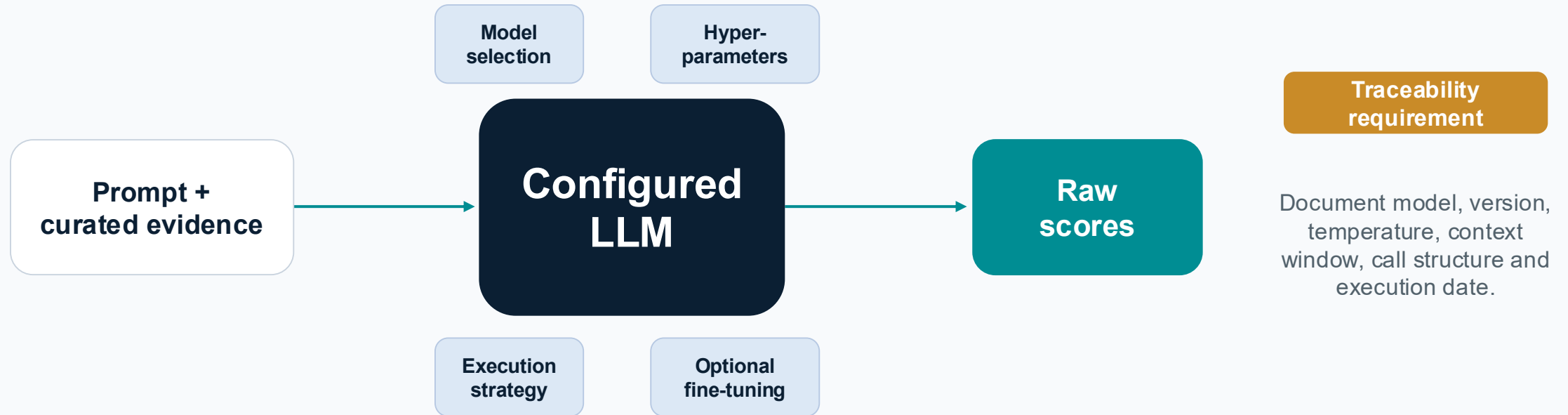
Output
format

Output: a reproducible measurement instrument

Key idea: prompt wording is not a technical detail. It is part of construct operationalization.

Stage 3: model execution is part of the measurement design

The same prompt can behave differently under different model and inference conditions.



Design objective: balance performance, scalability and reproducibility.

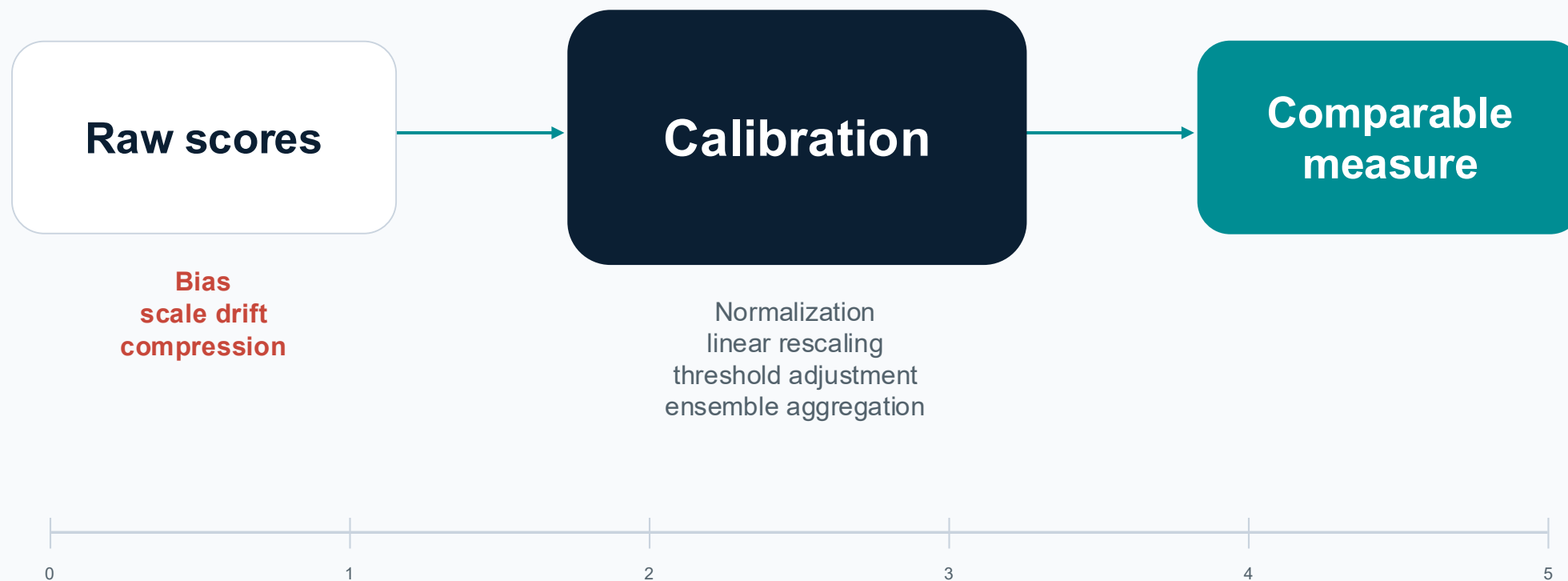
Stage 4: evaluation creates measurement evidence

Reliability and validity determine whether outputs can be interpreted as indicators.



Stage 5: calibration makes scores comparable

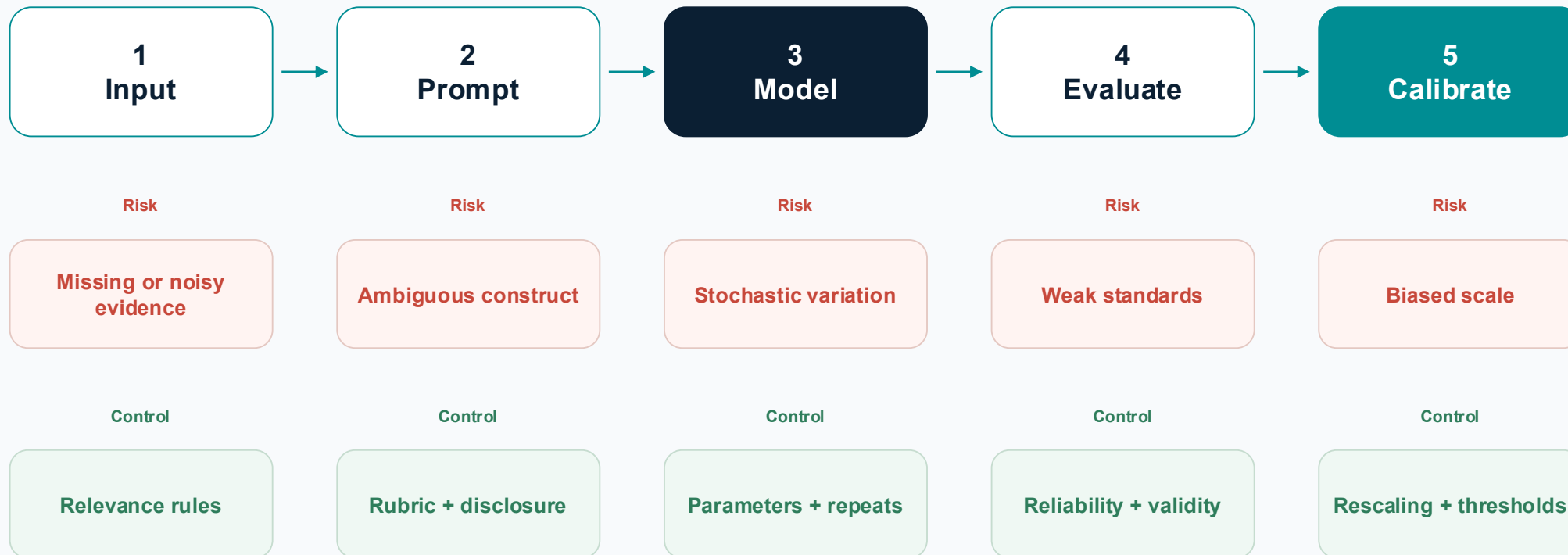
Even reliable systems can show systematic overestimation, underestimation or central compression.



Calibration aligns the output scale with empirical benchmarks or reference distributions.

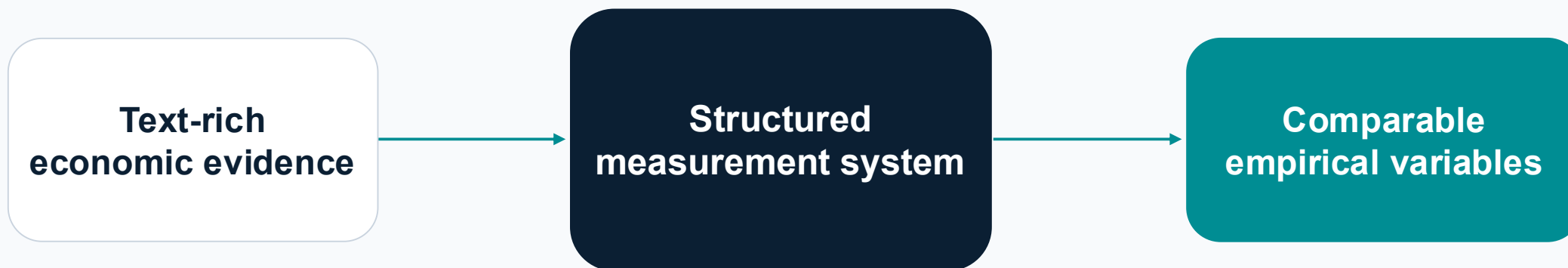
Each stage turns error risk into a control lever

The value of the framework is making hidden choices visible, auditable and improvable.



Why this matters for economic measurement

It enables reproducible variables from text-rich evidence sources that are difficult to use with conventional instruments.



Examples

Corporate disclosures
firm websites
policy documents
qualitative responses

Measurement gains

Transparent variables
control over measurement error
repeatable protocols

Research use

Hard-to-observe constructs
organizational practices
technological capabilities

The broader shift: from analyzing text to measuring from text.

A transparent reporting protocol for LLM-based measures

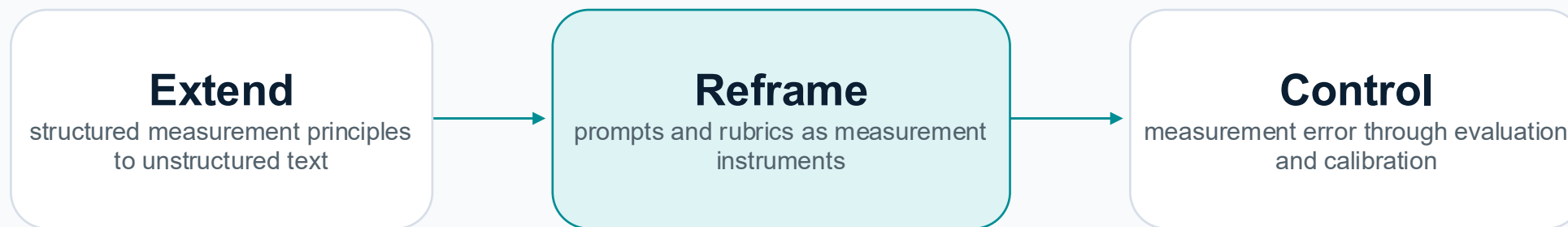
A practical checklist for making text-derived variables reproducible and auditable.

- 1 Define the construct and score scale
- 2 Document the input universe and selection rules
- 3 Disclose the prompt/rubric or a reproducible specification
- 4 Record model, version, parameters and execution strategy
- 5 Report reliability and validity evidence
- 6 Report calibration method and uncertainty handling

Use this as a replication table in papers, appendices or statistical documentation.

Main contribution: measuring from text

The paper provides a methodological foundation for rigorous LLM-based text measurement.



Future research directions

Empirically compare alternative measurement designs

Test robustness across models and contexts

Standardize calibration and uncertainty reporting

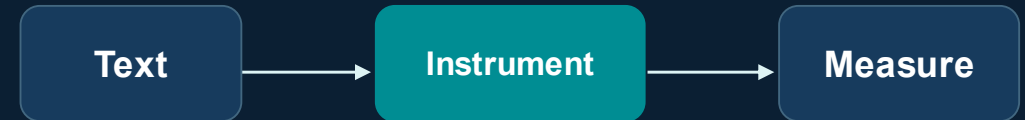
Integrate text-based measures with statistical and causal inference frameworks

Thank you

Questions & discussion

A foundation for treating text not only as data to be analyzed, but as evidence to be measured.

Xavier Martínez-Barbero
xamarbar@upv.es
eSMART Research Center, UPV



Selected references

Core works cited in the paper and useful for discussion.

Biemer (2010). Total survey error: design, implementation and evaluation.

Groves et al. (2009). Survey Methodology.

Krippendorff (2018). Content Analysis: An Introduction to Its Methodology.

Messick (1995). Validity of psychological assessment.

Gentzkow, Kelly & Taddy (2019). Text as Data.

Grimmer & Stewart (2013). Text as data: promise and pitfalls.

Ziems et al. (2024). Can Large Language Models Transform Computational Social Science?

Foisy et al. (2025). Prompting the Machine: Introducing an LLM Data Extraction Method for Social Scientists.