From Liaison Notes to Regional Indicators

Text-Based Measurement of Regional Economic Activity from Business Liaison Program Data and a Randomized Evaluation of a Generative-AI Report-Drafting Assistant

Thanatchon Chongmankhong

November 2026

Abstract

Regional output statistics in Thailand are compiled annually and released with a substantial lag, whereas the Business Liaison Program (BLP) of the Bank of Thailand collects timely but qualitative information from firms. This proposal sets out a four-study research program that converts Thai-language BLP interview notes into quantitative regional indicators and builds and evaluates a retrieval-augmented generative-AI assistant that drafts regional economic reports. Study 1 develops a six-topic codebook, compares lexicon, fine-tuned encoder and large-language-model (LLM) coding against an expert gold standard, and constructs diffusion indices weighted by sectoral output shares and adjusted for changes in panel composition through weighted two-way fixed effects. Study 2 tests predictive validity in a pseudo-real-time nowcasting exercise with Clark–West and Diebold–Mariano tests. Study 3 follows design science research to build an assistant that cites a source for every sentence and blocks any number it cannot trace to a data table. Study 4 is a pre-registered randomized crossover experiment testing whether the assistant reduces report-preparation time by more than 20% without degrading blinded quality ratings. An open-source implementation is validated end to end on synthetic data: composition adjustment raises the correlation between the index and latent activity from 0.89 to 0.92 on average, the verifier blocks every injected error, and a power analysis shows that 30 participants completing four tasks each provide power of 0.92 if the true time reduction is 35%. Results on the confidential BLP data are the object of the planned research and are not reported here.

Keywords business liaison · text as data · nowcasting · regional economics · large language models · retrieval-augmented generation · randomized crossover experiment · central banking

JEL classification C53 · C55 · C93 · E32 · E58 · R11

Contents

1 Introduction **4**

2 Related literature **5**
2.1 Central bank liaison and soft information . . . 5
2.2 Text as data and LLM-based measurement . . . 6
2.3 Nowcasting with soft data . . . 6
2.4 Generative AI and the productivity of knowledge work . . . 6
2.5 Retrieval augmentation, hallucination and governance . . . 6
2.6 Theoretical lenses . . . 6

3 Research design and data **7**
3.1 Overall design . . . 7
3.2 Data . . . 7
3.3 Gold-standard annotation . . . 7
3.4 Data governance and ethics . . . 8
3.5 Reproducibility . . . 8

4 Study 1: Measuring business conditions from liaison notes **8**
4.1 Codebook . . . 8
4.2 Scoring methods . . . 9
4.3 Index construction . . . 9
4.4 Composition adjustment . . . 9
4.5 Validity framework . . . 10

5 Study 2: Predictive validity and nowcasting **10**
5.1 Targets . . . 10
5.2 Pseudo-real-time design . . . 11
5.3 Models . . . 11
5.4 Evaluation . . . 11
5.5 Timeliness . . . 11

6 Study 3: Design of a grounded report-drafting assistant **12**
6.1 Requirements and baseline . . . 12
6.2 Architecture . . . 12
6.3 Grounding constraints and the verifier . . . 12
6.4 Model selection and retrieval . . . 13
6.5 Design cycles . . . 13
6.6 Offline evaluation . . . 13

7 Study 4: Randomized crossover experiment on preparation time **13**
7.1 Design . . . 13
7.2 Participants and tasks . . . 14
7.3 Outcomes . . . 14
7.4 Estimation and hypothesis tests . . . 14
7.5 Power analysis . . . 15
7.6 Field pilot . . . 15
7.7 Threats to validity . . . 16

8 Proof of concept on synthetic data **16**
8.1 Synthetic data . . . 16
8.2 Anonymization and coding . . . 16
8.3 Composition adjustment . . . 16
8.4 Nowcasting . . . 17

8.5 Verification of drafts . . . 17
8.6 Dry run of the pre-registered analysis . . . 18

9 Discussion **18**
9.1 Expected contributions . . . 18
9.2 Policy implications . . . 18
9.3 A responsible-AI guideline . . . 19
9.4 Limitations . . . 19

10 Work plan **19**
10.1 Effort allocation . . . 19
10.2 Timeline . . . 20
10.3 Publication plan . . . 20
10.4 Risk register . . . 20

11 Conclusion **21**

A Prompt templates **25**

B Software architecture **25**

C Power analysis grid **25**

D Synthetic data-generating process **26**

1 Introduction

Monetary policy and regional development policy depend on timely information about economic conditions outside the capital. In Thailand, the official measure of regional output—the gross regional and provincial product compiled by the Office of the National Economic and Social Development Council (NESDC)—is annual and released with a substantial lag, so it cannot reveal turning points within the year. Higher-frequency regional indicators cover selected activities such as consumption, investment, manufacturing, tourism and farm income, and are themselves published weeks after the reference period.

Central banks complement official statistics with systematic conversations with firms. Prominent examples are the Federal Reserve's Beige Book, the Bank of England's Agents, the Reserve Bank of Australia's business liaison program and the Bank of Canada's Business Outlook Survey. The Bank of Thailand (BOT) runs a Business Liaison Program (BLP) through its regional offices. Liaison information is timely, captures forward-looking intentions and explains *why* conditions change. Its weakness is its form: it arrives as free text written by many interviewers, the set of firms changes from one round to the next, and turning it into a report requires economists to read, summarize and cross-check a large volume of notes by hand.

Two developments make it possible to address both weaknesses. Text-as-data methods, and more recently large language models (LLMs), can score narrative text at scale (Gentzkow et al., 2019; Hansen and Kazinnik, 2023). Generative AI can draft text, and randomized studies report substantial time savings in writing and other knowledge work (Noy and Zhang, 2023; Dell'Acqua et al., 2023; Brynjolfsson et al., 2025). Central bank reports, however, impose requirements that generic tools do not meet: every statement must be traceable to evidence, every number must be correct, and confidential information about individual firms must never appear in a draft.

The research program is organized around four gaps.

G1. Measurement in a low-resource language. Quantitative work on liaison text has concentrated on English-language programs; Thai lacks validated economic lexicons and coding benchmarks.
G2. Panel composition. Liaison samples are purposive and rotate. A change in the mix of firms can move a raw diffusion index even when no individual firm's conditions change.
G3. Causal evidence in central banking. Experimental evidence on generative AI comes from consulting, customer support, software development and generic writing tasks. We are not aware of randomized evidence for analytical report production inside a central bank.
G4. Auditable generation. Designs that guarantee sentence-level provenance and numerically correct output for official reports are rarely specified and evaluated.

The program answers four research questions.

- **RQ1.** Can BLP interview notes be converted into regional indicators that are *reliable*, in the sense of agreeing with expert coders, and *valid*, in the sense of tracking official measures of activity?
- **RQ2.** Do BLP indicators make regional nowcasts more accurate and more timely than benchmarks that use official data only?
- **RQ3.** Does a grounded generative-AI assistant reduce report-preparation time by more than 20% without lowering report quality or increasing errors?
- **RQ4.** Which safeguards are needed for the responsible use of generative AI in central bank reporting?

The program makes four contributions. First, it provides a measurement framework for Thai liaison text: a codebook, a head-to-head comparison of lexicon, fine-tuned encoder and LLM coding, and an index that removes the effect of changing panel composition. Second, it supplies evidence on the information content

Table 1: Hypotheses and pre-specified targets.

	Hypothesis	Measure	Target
H1	LLM coding agrees with expert coders	Cohen's κ and macro-F1 against the adjudicated gold standard	$\kappa \geq 0.70$ and F1 ≥ 0.80
H2	The BLP index co-moves with activity	Correlation with an official activity composite	$r \geq 0.60$
H3	BLP indicators improve nowcasts	Out-of-sample RMSE relative to an AR benchmark	Ratio ≤ 0.90; Clark–West $p < 0.05$
H4	BLP indicators are timelier	Days available before the official release	≥ 30 days
H5	The assistant saves time	$\tau = 1 - e^{\beta}$ from a log-time mixed model	$\tau > 20\%$, one-sided test
H6	Report quality is not worse	Blinded 10-point quality rubric	Non-inferiority margin 0.5 points
H7	Errors do not increase	Factual and numeric errors per 1,000 words	Rate ratio ≤ 1; no numeric error after verification

of liaison data for regional nowcasting in an emerging economy where official regional data are sparse. Third, it contributes a design artifact—a drafting assistant whose output is traceable sentence by sentence and is blocked when a number cannot be verified—together with its evaluation. Fourth, it delivers pre-registered randomized evidence on the time and quality effects of generative AI in central bank analytical work. An open-source implementation accompanies the proposal.

Table 1 states the hypotheses with their quantitative targets. Section 2 reviews the literature, Section 3 sets out the design and data, and Sections 4 to 7 describe the four studies. Section 8 validates the software on synthetic data, Section 9 discusses implications and limitations, and Section 10 presents the work plan.

2 Related literature

2.1 Central bank liaison and soft information

A long literature asks whether the narrative information that central banks collect from firms contains signal beyond official data. For the Federal Reserve's Beige Book, Balke and Petersen (2002) show that quantified district descriptions track current activity, Armesto et al. (2009) use mixed-frequency regressions to measure its information content, and Sadique et al. (2013) relate Beige Book sentiment to economic activity. For the Bank of Canada's Business Outlook Survey, Pichette and Rennison (2011) summarize responses with principal components. Two features of this literature matter here: quantification has relied on human or dictionary scoring of English text, and the composition of the liaison panel has received limited attention. To our knowledge, BLP information in Thailand has been used qualitatively in regional and monetary policy communication but has not been systematically quantified and validated in published work.

2.2 Text as data and LLM-based measurement

Economists build indicators from text with dictionaries (Loughran and McDonald, 2011), keyword counts (Baker et al., 2016), sentiment models (Shapiro et al., 2022) and topic models (Thorsrud, 2020), and use text in machine-learning forecasts (Kalamara et al., 2022); Gentzkow et al. (2019) survey the field. LLMs extend these tools to classification without task-specific training (Hansen and Kazinnik, 2023). For Thai, the building blocks are word segmentation (Phatthiyaphaibun et al., 2023), pretrained encoders (Lowphansirikul et al., 2021) and Thai-capable LLMs (Pipatanakul et al., 2023). Measurement quality is judged with agreement statistics (Cohen, 1960; Landis and Koch, 1977; Krippendorff, 2018).

2.3 Nowcasting with soft data

Nowcasting combines data released at different frequencies and dates (Giannone et al., 2008; Bańbura et al., 2013), using bridge equations, mixed-data sampling (Ghysels et al., 2007) and factor models (Stock and Watson, 2002). Forecast comparisons rely on the Diebold and Mariano (1995) test with the small-sample correction of Harvey et al. (1997), the Clark and West (2007) test for nested models and the model confidence set (Hansen et al., 2011). Low-frequency regional output can be distributed to a higher frequency with related series (Chow and Lin, 1971), and satellite data offer an independent proxy for activity (Henderson et al., 2012).

2.4 Generative AI and the productivity of knowledge work

Randomized evidence points to large but uneven effects. In professional writing tasks, Noy and Zhang (2023) report that access to a chatbot cut completion time by about 40% and raised quality. In customer support, Brynjolfsson et al. (2025) find an increase of about 14% in issues resolved per hour, concentrated among less experienced workers. Consultants studied by Dell'Acqua et al. (2023) worked about 25% faster on tasks inside the "jagged frontier" of model capability but made more errors on a task outside it, and developers using a coding assistant completed a programming task about 56% faster (Peng et al., 2023). Against these estimates a 20% threshold is conservative, but central bank reporting contains verification work that generative AI may not shorten, and errors are costly. The experiment in Section 7 therefore measures time, quality and errors jointly.

2.5 Retrieval augmentation, hallucination and governance

Retrieval-augmented generation conditions a language model on retrieved documents (Lewis et al., 2020); classical BM25 ranking (Robertson and Zaragoza, 2009) and reciprocal rank fusion (Cormack et al., 2009) remain strong components. Retrieval reduces but does not eliminate unsupported statements (Ji et al., 2023), which motivates automated faithfulness evaluation (Es et al., 2024) and the explicit verification layer proposed here. Governance frameworks emphasize risk mapping, measurement and human oversight (National Institute of Standards and Technology, 2023), and the BIS discusses the opportunities and risks of AI for central banks (Bank for International Settlements, 2024).

2.6 Theoretical lenses

Four lenses organize the program. *Soft versus hard information* (Stein, 2002; Petersen, 2004): liaison information is soft—contextual, hard to verify and costly to transmit—and quantification hardens it at the cost

of nuance; Study 1 measures what survives hardening, and the assistant keeps a link to the original text. *The task-based framework* (Acemoglu and Autor, 2011): report production is a bundle of tasks (reading, summarizing, retrieving data, drafting, checking and revising), and generative AI substitutes for some and complements others, which predicts heterogeneous gains across tasks and workers. *Design science research* (Hevner et al., 2004; Peffers et al., 2007) frames the assistant as an artifact that must be built and rigorously evaluated. *Technology acceptance* (Davis, 1989) links perceived usefulness and ease of use to adoption, measured in Study 4.

3 Research design and data

3.1 Overall design

The program follows a pragmatist mixed-methods design with two tracks (Figure 1). The measurement track turns notes into indicators (Study 1) and tests their predictive validity (Study 2). The report-production track builds an assistant that drafts from the notes and the new indicators (Study 3) and estimates its causal effect on preparation time (Study 4). Safeguards for data governance cut across both tracks.

3.2 Data

Table 2 lists the data sources. The BLP notes are the core input and are confidential; all processing of firm-level text takes place inside the BOT's secure environment.

3.3 Gold-standard annotation

A stratified sample of at least 1,200 note segments, drawn by region, sector and year, is coded independently by two economists using the codebook in Table 3; a third economist adjudicates disagreements. The code-

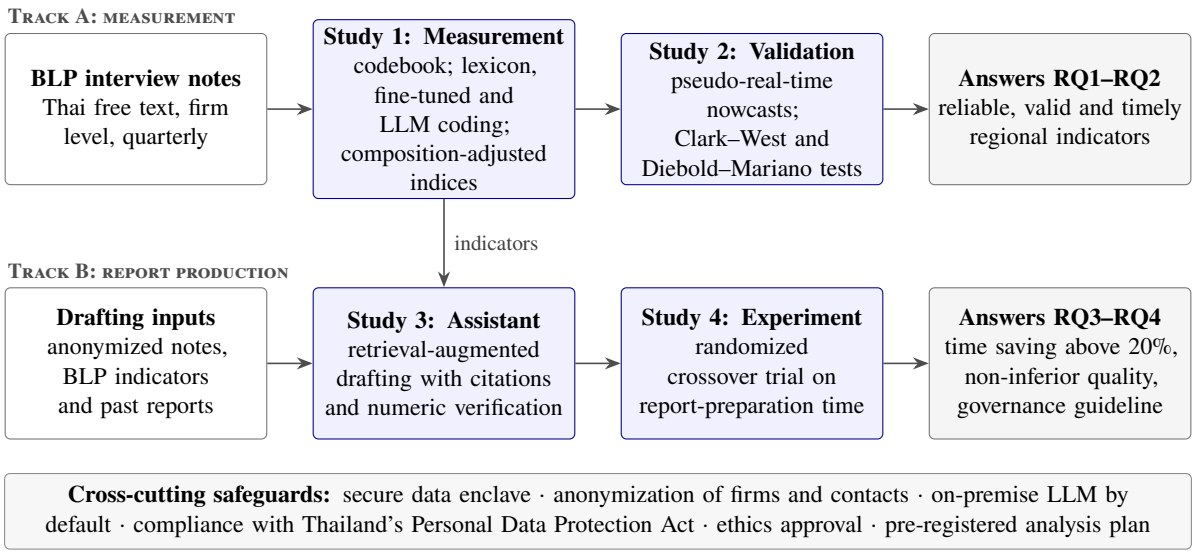


Figure 1: Research framework. Track A converts liaison notes into indicators and tests them; Track B uses the notes and indicators to draft reports and tests the effect on preparation time. Shaded boxes are the four studies.

Table 2: Data sources.

Source	Frequency	Used in	Role and caveats
BLP interview notes (Thai text)	Each round	Studies 1, 3	Core input; confidential, processed in the data enclave only
Structured BLP ratings, if recorded	Each round	Study 1	Benchmark for text-based scores
BOT regional economic indicators	Monthly	Study 2	Nowcasting targets and benchmarks
NESDC gross regional product	Annual	Study 2	Long-run target; few observations, pooled across regions
NSO Labour Force Survey	Quarterly	Study 2	Validity check for the employment index
VIIRS night-time lights	Monthly	Study 2	Independent alternative measure of activity
Past regional economic reports	Monthly, quarterly	Study 3	Style reference and key points for evaluation
Time and edit logs	Per session	Study 4	Primary and secondary outcomes

book is revised until human–human agreement reaches $\kappa \geq 0.75$. The adjudicated sample is split 60/20/20 into training, development and test sets for the supervised model; all reported comparisons use the test set.

3.4 Data governance and ethics

Firm names, contact names, telephone numbers and e-mail addresses are masked before any modeling, first by exact matching against the BLP contact database and then by pattern rules and named-entity recognition; masking recall is audited on a human-checked sample. LLMs run on premises by default. An external API may be used only with masked data, under an approved data-processing agreement. All processing complies with Thailand's Personal Data Protection Act B.E. 2562 (2019) and the BOT's confidentiality rules. Study 4 requires ethics approval and informed consent, individual results are never used for performance assessment, and the hypotheses and analysis code are pre-registered before data collection (Nosek et al., 2018).

3.5 Reproducibility

The code, configuration files and a synthetic data generator are released with the proposal; random seeds, model weights and prompt versions are fixed and logged. Results on confidential data can therefore be reproduced inside the enclave, and every software component can be inspected outside it.

4 Study 1: Measuring business conditions from liaison notes

4.1 Codebook

Each note is coded on the six topics in Table 3. A topic receives a direction relative to the previous quarter—up (+1), unchanged (0) or down (−1)—or is marked as not mentioned. Statements about competitors, the government or world markets are not attributed to the interviewed firm, and a topic that contains both increases and decreases is coded as unchanged.

Table 3: Codebook. Each topic is coded as up, unchanged, down or not mentioned, relative to the previous quarter.

Topic	Definition	Index role
Sales and demand	Sales, orders, revenue, customer and tourist numbers, occupancy	Activity (BLP-RAI)
Employment	Headcount, hiring, working hours	Activity (BLP-RAI)
Investment	Capital spending, machinery, capacity and branch expansion by the firm	Activity (BLP-RAI)
Selling prices	Prices charged by the firm	Price pressure
Costs	Input costs: raw materials, wages, energy	Cost pressure
Outlook	Expected business conditions over the coming quarter	Leading indicator

4.2 Scoring methods

Four scorers are compared on the gold-standard test set.

- **M1. Lexicon.** A Thai economic lexicon assigns each clause to a topic by keyword priority and to a direction by counting increase and decrease terms, with negation handling ("not yet raised" is unchanged). It is transparent and serves as the baseline.
- **M2. Fine-tuned encoder.** A Thai pretrained encoder (Lowphansirikul et al., 2021) with one four-class head per topic, trained on the gold standard.
- **M3. LLM structured extraction.** An instruction-tuned LLM returns one label per topic under a JSON schema, with the codebook and worked examples in the prompt (Appendix A). On-premise and API models are compared.
- **M4. Ensemble.** A majority vote over M1–M3.

The comparison reports accuracy, macro-F1, Cohen's κ and Krippendorff's α by topic, together with cost and processing time per 1,000 notes.

4.3 Index construction

Let $s_{ikt} \in \{-1, 0, +1\}$ denote the score of firm i on topic k in quarter t, where firm i belongs to region r and sector $s(i)$. The diffusion index, or net balance, is

$$DI_{rkt} \;=\; 100\left(p^{+}_{rkt} - p^{-}_{rkt}\right) \;=\; \frac{100}{N_{rkt}} \sum_{i \in I_{rkt}} s_{ikt}, \tag{1}$$

where p^{+} and p^{-} are the shares of firms reporting an increase and a decrease. Because the liaison sample is purposive, sectoral shares in the sample differ from the structure of the regional economy. The sector-weighted index re-weights sector-level net balances by the sector's share w_{rs} in regional output, renormalized over the sectors $\mathcal{S}_{rt}$ observed in quarter t:

$$I^{w}_{rkt} \;=\; \frac{\sum_{s \in \mathcal{S}_{rt}} w_{rs}\, DI_{rskt}}{\sum_{s \in \mathcal{S}_{rt}} w_{rs}}. \tag{2}$$

4.4 Composition adjustment

Firms differ persistently in how optimistic they sound, and the set of firms interviewed changes over time. If firms that enter the panel are systematically more optimistic, a raw index drifts upward even when no

Table 4: Validity framework for the BLP indices.

Property	Test	Criterion
Reliability	κ and α between each scorer and the adjudicated gold standard	$\kappa \geq 0.70$ (H1)
Content validity	Expert review of the codebook against report sections	All report topics covered
Convergent validity	Correlation with official activity composites	$r \geq 0.60$ (H2)
Discriminant validity	Sector indices correlate more with their own sector data (for example tourism) than with unrelated series	Expected ordering holds
Predictive validity	Pseudo-real-time nowcasting (Study 2)	RMSE ratio ≤ 0.90 (H3)
Robustness	Leave-one-sector-out, alternative weights, alternative scorers	Conclusions unchanged

firm's conditions improve. The composition-adjusted index separates persistent firm effects from common quarterly movements with a weighted two-way fixed-effects model,

$$s_{ikt} = \alpha_{ik} + \lambda_{rkt} + \varepsilon_{ikt}, \qquad \omega_{it} = \frac{w_{rs(i)}}{n_{rs(i)t}}, \tag{3}$$

estimated by weighted least squares with weights ω_{it}, so that each sector contributes in proportion to its output share. The quarter effects λ_{rkt} are identified from changes within firms that are interviewed more than once and are computed by alternating projections, which converge for connected panels. They are normalized to the mean of the weighted raw index,

$$I^{\text{adj}}_{rkt} = 100\left(\hat{\lambda}_{rkt} - \bar{\hat{\lambda}}_{rk}\right) + \bar{I}^{w}_{rk}, \tag{4}$$

so the adjusted index retains the net-balance scale. The composite BLP Regional Activity Index (BLP-RAI) averages the adjusted sales, employment and investment indices. The cross-sectional standard deviation of firm scores provides an uncertainty index, and the series are seasonally adjusted with X-13ARIMA-SEATS where seasonality is detected.

Fixed effects remove time-invariant differences between firms but not selection on current conditions—for example, distressed firms declining an interview. Participation is therefore modeled as a function of past scores, and the indices are re-estimated with inverse-probability weights as a sensitivity check.

4.5 Validity framework

Table 4 lists the validity tests and pass criteria. Predictive validity is the subject of Study 2.

5 Study 2: Predictive validity and nowcasting

5.1 Targets

Three targets are used: (i) quarterly composites of the BOT's monthly regional indicators; (ii) annual gross regional product, pooled across regions to increase the number of observations; and (iii) quarterly regional product distributed from annual totals with the Chow and Lin (1971) method. When the BLP index itself serves as a related series in (iii), evaluation uses only targets that do not depend on it, to avoid circularity.

Table 5: Nowcasting models.

Code	Specification	Purpose
M0	AR(1) with region fixed effects	Benchmark
M1	M0 + BLP-RAI	Main test of H3
M2	M1 + lagged outlook index	Value of forward-looking statements
M3	Bridge equation or U-MIDAS with monthly indicators + BLP	Comparison with current practice
M4	Dynamic factor model	Advanced benchmark
M5	Panel pooled across regions	Statistical power with short histories
M6	Elastic net and gradient boosting	Robustness to functional form

5.2 Pseudo-real-time design

At the end of quarter t the information set contains the BLP indices for quarter t and official data released by that date, following each series' publication calendar. Models are re-estimated on an expanding window, and out-of-sample evaluation starts in 2018 so that it includes the COVID-19 shock. First-release vintages are used where they are available.

5.3 Models

Table 5 lists the competing models. The core specification adds the BLP index to an autoregressive benchmark with region fixed effects,

$$y_{rt} = c + \rho\, y_{r,t-1} + \beta\, \mathrm{RAI}_{rt} + \gamma\, \mathrm{OUT}_{r,t-1} + \mu_r + u_{rt}, \tag{5}$$

where OUT is the outlook index; M1 sets $\gamma = 0$ and M0 sets $\beta = \gamma = 0$.

5.4 Evaluation

Accuracy is measured by the RMSE relative to M0, the mean absolute error and directional accuracy. Nested comparisons use the Clark and West (2007) statistic, which adjusts the loss differential for the noise introduced by estimating additional parameters,

$$f_t = \left(y_t - \hat{y}_t^0\right)^2 - \left[\left(y_t - \hat{y}_t^1\right)^2 - \left(\hat{y}_t^0 - \hat{y}_t^1\right)^2\right], \tag{6}$$

tested one-sided with heteroskedasticity- and autocorrelation-consistent standard errors. Non-nested comparisons use the Diebold–Mariano test with the small-sample correction of Harvey et al. (1997), and the model confidence set summarizes all models (Hansen et al., 2011). Event windows around the 2020 pandemic shock and the 2024 floods in northern Thailand show how quickly each model detects turning points.

5.5 Timeliness

Timeliness is the number of days between the date the BLP index can be computed and the release of the corresponding official figure, reported as a distribution across quarters; H4 requires a median gain of at least 30 days.

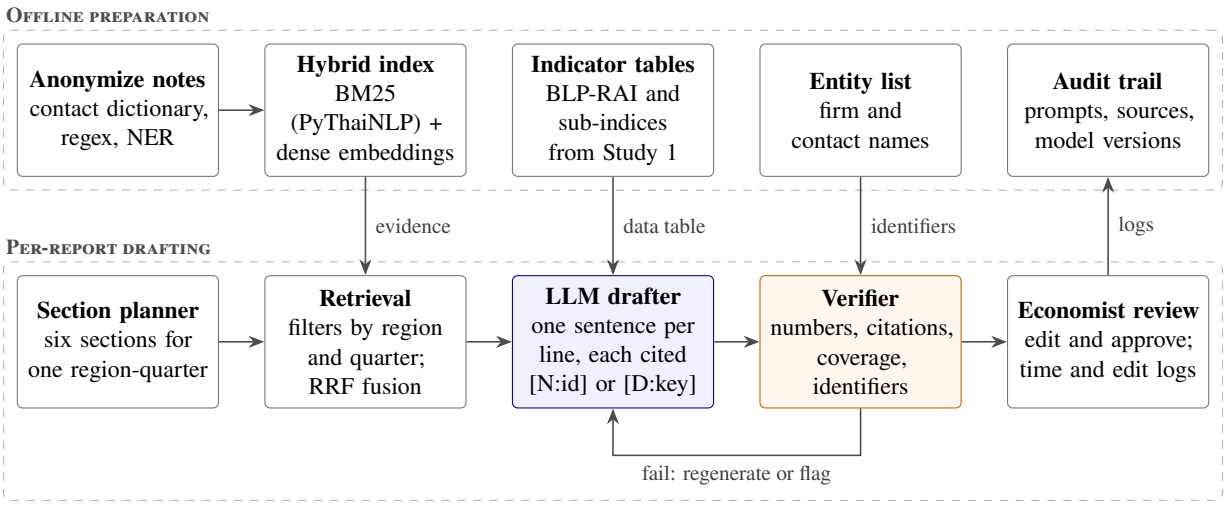


Figure 2: Architecture of the drafting assistant. Only drafts that pass verification reach the economist; failing drafts are regenerated or flagged. Every prompt, retrieved source and model version is written to the audit trail.

6 Study 3: Design of a grounded report-drafting assistant

6.1 Requirements and baseline

Semi-structured interviews and a focus group with 8–12 report writers establish requirements. A time-and-motion study over two report cycles measures the time spent reading notes, summarizing, retrieving data, drafting, checking numbers and revising. The baseline locates where the assistant can save time and defines the tasks used in Study 4.

6.2 Architecture

Figure 2 shows the architecture. An offline layer masks identifiers, indexes the notes for hybrid retrieval and stores the indicator tables from Study 1. For each report, a planner splits the report into six sections; for each section, the retriever returns evidence filtered by region and quarter, the drafter writes cited sentences, and the verifier decides whether the draft may reach the economist.

6.3 Grounding constraints and the verifier

The drafter must satisfy four constraints, and the verifier enforces the first three mechanically.

- **C1. Provenance.** Each sentence is written on its own line and ends with at least one citation, either to a note (`[N:id]`) or to a row of the data table (`[D:key]`). Citations must point to evidence actually supplied for that section.
- **C2. Numeric grounding.** A number is accepted only if it matches a value in the data table (level, previous value, change or absolute change), a number in the cited evidence, a contextual number (quarter, year in the Gregorian or Buddhist calendar), or the count of notes cited in the same sentence.
- **C3. Confidentiality.** The draft may not contain any name from the entity list or any pattern that resembles a firm name, contact name, telephone number or e-mail address.
- **C4. Abstention.** When evidence is insufficient, the drafter must say so rather than generalize.

Table 6: Offline evaluation of the assistant.

Metric	Definition	Target
Faithfulness	Share of statements supported by the cited source, judged by experts	≥ 95%
Numeric accuracy	Share of numbers that match the data table	100% (otherwise blocked)
Citation precision	Share of citations that support their sentence	≥ 90%
Key-point coverage	Share of key points in the published report that the draft covers	≥ 80%
Confidentiality leakage	Identifiers appearing in a draft	Zero

A draft that violates C1–C3 is blocked rather than passed with a warning; its sentence-level citation coverage must reach at least 95%.

6.4 Model selection and retrieval

On-premise open-weight models with Thai capability (Pipatanakul et al., 2023) are the default because notes never leave the organization, model versions can be frozen and costs are fixed. API models are evaluated on masked data as a benchmark. Retrieval combines BM25 over PyThaiNLP tokens with dense multilingual embeddings and fuses the two rankings by reciprocal rank fusion (Robertson and Zaragoza, 2009; Cormack et al., 2009).

6.5 Design cycles

The assistant is developed in three design-science cycles (Peffers et al., 2007): a retrieval-and-citation prototype evaluated offline; the addition of the verifier and an editing interface, tested in think-aloud sessions with three to five users; and a final revision after which the system, prompts and model weights are frozen for Study 4.

6.6 Offline evaluation

Table 6 lists the offline metrics. Three variants are compared to isolate the contribution of each component: an extractive template without an LLM, an LLM without retrieval and the full retrieval-augmented system.

7 Study 4: Randomized crossover experiment on preparation time

7.1 Design

Each participant drafts four reports, two with the assistant and two without, in a randomized crossover design (Jones and Kenward, 2014) shown in Figure 3. Participants are stratified by experience and randomized to the sequence ABAB or BABA, where A denotes AI-assisted drafting. The order of the four tasks follows a Williams Latin square (Williams, 1949), which balances first-order carryover and ensures that every task appears equally often under both conditions.

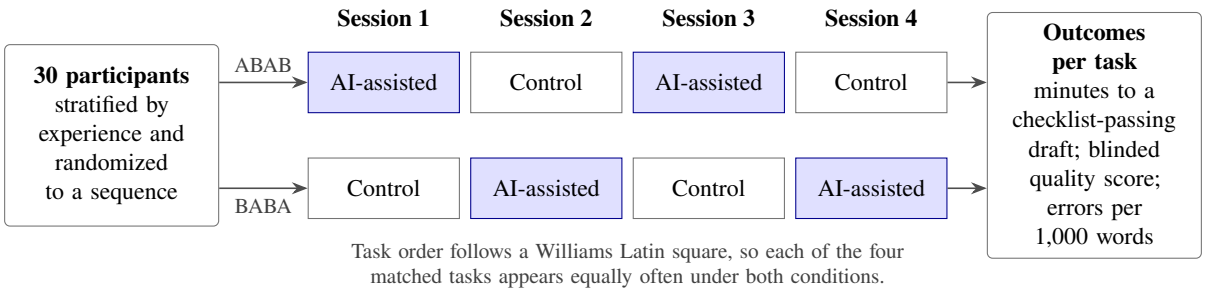


Figure 3: Randomized crossover design. Each participant completes four matched drafting tasks, alternating between AI-assisted and control conditions; the two sequences mirror each other.

Table 7: Outcomes of the experiment.

Level	Variable	Measurement
Primary	Minutes to a checklist-passing draft	Automatic time log of active working time
Secondary	Quality	10-point rubric (accuracy, completeness, coherence, insight), two blinded raters
Secondary	Errors	Numeric errors and unsupported statements per 1,000 words
Secondary	Workload and usability	NASA-TLX (Hart and Staveland, 1988); System Usability Scale (Brooke, 1996), target ≥ 70
Exploratory	Reliance on the assistant	Edit distance between the AI draft and the submitted draft

7.2 Participants and tasks

Participants are BOT economists and analysts who prepare regional reports, recruited from the regional offices and the head office with a target of 30. If fewer are available, the design moves to eight tasks per participant (two per session), and trained research analysts may join as a separate stratum analyzed in a pre-specified subgroup. Each task is to draft the six-section report for one region and quarter from a fixed packet of masked notes and data tables. Packets are matched on the number of notes and the complexity of the data and pilot-tested for difficulty. A task ends when the draft passes a standard completion checklist, which prevents speed from being bought with incompleteness.

7.3 Outcomes

Table 7 lists the outcomes. Quality is rated by two economists who do not know the condition under which a draft was produced.

7.4 Estimation and hypothesis tests

Preparation time is analyzed on the log scale with a linear mixed model,

$$\log T_{ijk} \;=\; \mu + \beta\,\mathrm{AI}_{ij} + \gamma_{\text{session}} + \delta_{\text{task}} + u_i + \varepsilon_{ijk}, \tag{7}$$

where u_i is a participant random effect. The proportional time reduction is $\tau = 1 - e^{\beta}$, and H5 is the one-sided test

$$H_0 : \beta \geq \ln 0.8 \qquad \text{against} \qquad H_1 : \beta < \ln 0.8. \tag{8}$$

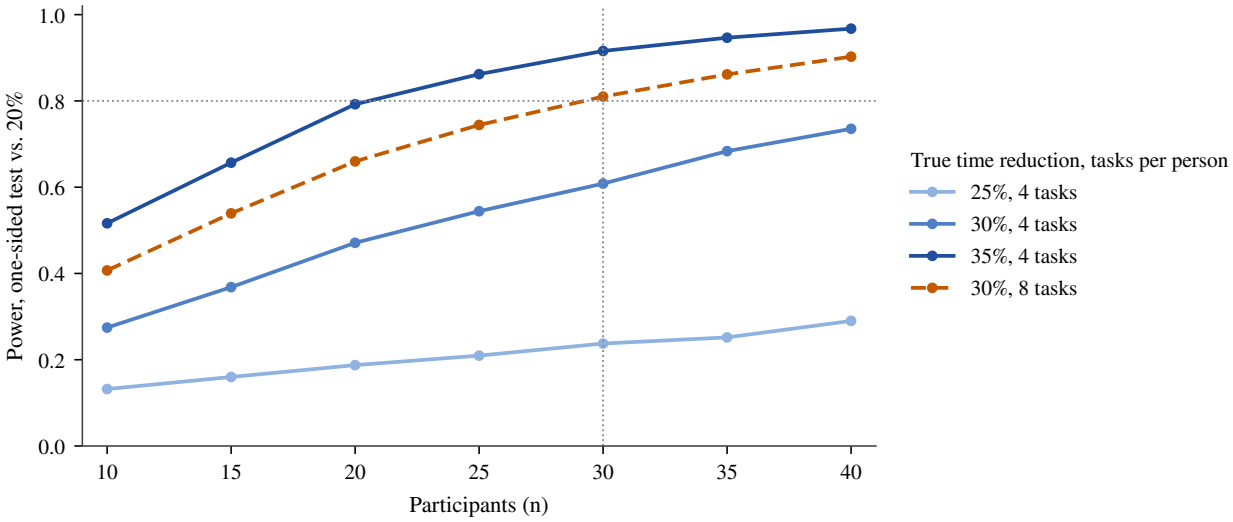


Figure 4: Power of the one-sided test that preparation time falls by more than 20%, by number of participants, true reduction and tasks per participant (4,000 simulations per point). Dotted lines mark power of 0.80 and the planned sample of 30 participants.

H5 is supported when the lower one-sided 95% confidence bound for τ exceeds 20%. Quality is analyzed with the same model, and H6 is a non-inferiority test with margin 0.5 points, $H_0 : \theta \leq -0.5$ against $H_1 : \theta > -0.5$, reported following the CONSORT extension for non-inferiority trials (Piaggio et al., 2012). Error counts are modeled with a Poisson regression with an offset for document length and standard errors clustered by participant (H7). A first-order carryover term (whether the previous session used the assistant) tests for carryover; robustness checks use a Wilcoxon signed-rank test on participant-level differences and per-protocol estimates. Heterogeneity by experience is a pre-specified secondary analysis, motivated by the larger gains for less experienced workers found by Brynjolfsson et al. (2025).

7.5 Power analysis

Power is computed by simulating Equation (7) under the crossover design, with standard deviations on the log scale of 0.15 for task effects, 0.30 for residuals and 0.15 for individual differences in the AI effect, and a learning effect of -0.03 per session; these values will be updated with pilot data. Figure 4 shows that testing against a 20% threshold, rather than against zero, is demanding. With four tasks per participant, 30 participants provide power of 0.92 if the true reduction is 35% but only 0.61 if it is 30%. Eight tasks per participant raise power at a 30% reduction to 0.81. The plan is therefore 30 participants with four tasks each, moving to eight tasks if the pilot suggests a smaller effect or more residual variation. For quality, 30 participants give power of 0.83 to establish non-inferiority when the true difference is zero and the within-person standard deviation is one point.

7.6 Field pilot

After the experiment, the assistant is used in two regular report cycles. Time logs are compared with the baseline time-and-motion study. Because the comparison is not randomized, the pilot supports external validity and adoption analysis rather than causal claims.

Table 8: Threats to validity in Study 4.

Threat	Safeguard
Learning and carryover	Crossover with mirrored sequences, Williams square, test for first-order carryover
Unequal task difficulty	Matched and pilot-tested packets; task fixed effects
Speed bought with quality	Completion defined by checklist; non-inferiority test on blinded quality
Novelty and Hawthorne effects	Training session before the experiment; field pilot over two cycles
Small sample	Mixed model; option of eight tasks per participant
Limited external validity	Field pilot with real report cycles

7.7 Threats to validity

Table 8 lists the main threats and the corresponding safeguards.

8 Proof of concept on synthetic data

All numbers in this section come from synthetic data generated by the companion code. They show that the pipeline is implemented correctly and behaves as intended; they are not evidence about BLP data or about the hypotheses.

8.1 Synthetic data

The generator (Appendix D) produces three regions over 56 quarters (2012Q1–2025Q4), with 45 interviews per region and quarter and 80% quarterly retention of firms, giving 7,560 notes from 1,714 firms. Latent national, regional and cost factors drive firm responses, with a pandemic shock in 2020Q2 and a slower recovery in tourism. Firms that join later are more optimistic, which creates composition bias, and the sample's sector mix differs from regional output shares. Notes are written in Thai from phrase templates and include mixed statements, negations and distractor sentences about competitors, the government and world markets, together with fictitious firm names, contact names and telephone numbers.

8.2 Anonymization and coding

After dictionary and pattern masking, no identifier remained in any of the 7,560 notes. The lexicon scorer reached accuracy of 0.970, macro-F1 of 0.971 and $\kappa = 0.958$ against the generated labels; agreement was lowest for investment ($\kappa = 0.843$) because distractor sentences about competitors' expansion are misattributed to the firm. Two simulated human coders agreed with $\kappa = 0.787$ and $\alpha = 0.802$. Because synthetic notes are built from templates, these figures are an upper bound for the lexicon on real notes, which is why Study 1 compares it with fine-tuned and LLM coding.

8.3 Composition adjustment

Figure 5 compares the raw and adjusted activity indices with the latent activity that generated the data. The raw index drifts upward relative to the truth as optimistic firms enter the panel; the adjusted index removes

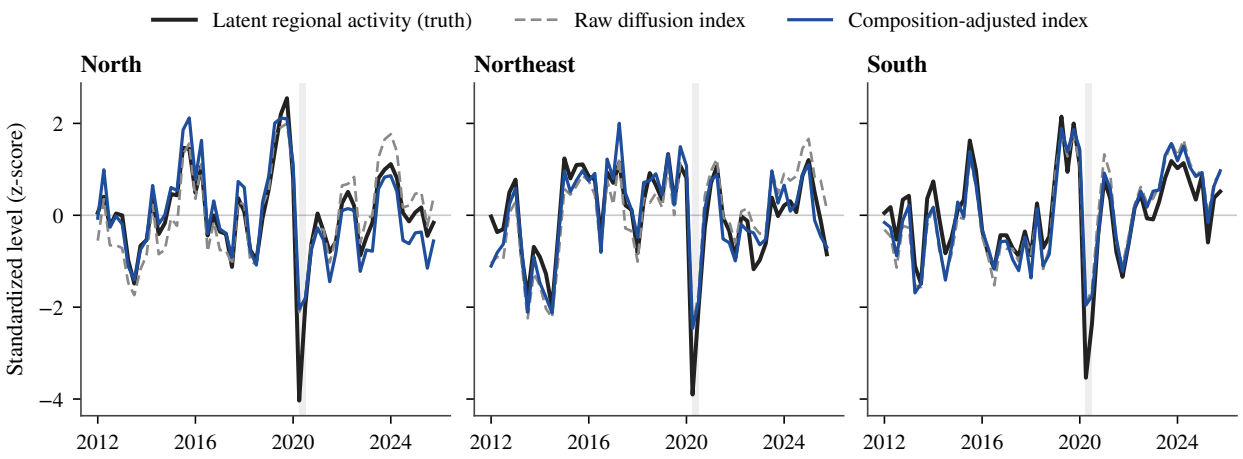


Figure 5: Raw and composition-adjusted BLP activity indices against latent regional activity in synthetic data, 2012Q1–2025Q4. All series are standardized; the shaded band marks 2020Q2.

Table 9: Out-of-sample nowcast accuracy on synthetic data, 96 nowcasts, 2018Q1–2025Q4.

Model	RMSE	MAE	RMSE ratio	Clark–West t (p)	Diebold–Mariano (p)
M0: AR(1)	1.681	1.165	1.000	–	–
M1: + BLP-RAI	0.986	0.750	0.587	3.10 (0.001)	2.69 (0.008)
M2: + lagged outlook	0.901	0.716	0.536	3.03 (0.001)	2.80 (0.006)

the drift. Correlation with latent activity rises from 0.862 to 0.919 in the Northeast, from 0.895 to 0.905 in the North and from 0.901 to 0.921 in the South. Sector weighting alone does not help in this data-generating process (0.858, 0.895 and 0.897), because sectors share the regional factor except during the tourism shock.

8.4 Nowcasting

The pseudo-real-time exercise produces 96 nowcasts (three regions, 2018Q1–2025Q4). Adding the BLP index lowers the RMSE by 41% relative to the autoregressive benchmark, and adding the lagged outlook index lowers it by 46% (Table 9 and Figure 6). Both improvements are significant under the Clark–West and Diebold–Mariano tests. The outlook index correlates 0.64 with contemporaneous growth and 0.81 with growth one quarter ahead, consistent with its role as a leading indicator. These gains reflect the synthetic data-generating process and set no expectation for real data.

8.5 Verification of drafts

With the extractive baseline in place of an LLM, the assistant drafted all six sections of a report for the North region in 2025Q4: 26 sentences, every one cited, and every section passed verification. As a negative control, a sentence containing a fictitious firm name, an invented number and a citation to a non-existent data row was appended to one section. The verifier blocked the section, reporting the two unverifiable numbers, the invalid citation and the firm name.

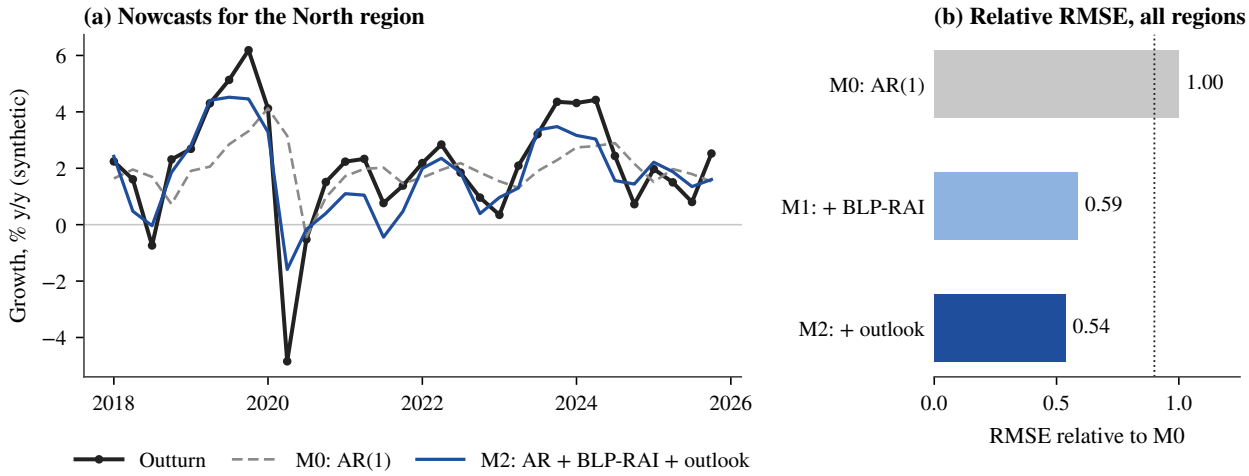


Figure 6: Pseudo-real-time nowcasts on synthetic data. Panel (a) shows the North region; panel (b) pools all regions, and the dotted line marks the H3 target ratio of 0.90.

8.6 Dry run of the pre-registered analysis

The analysis code was run on a simulated experiment with 30 participants and a true reduction of 30%. The randomization balanced tasks across conditions (14 to 16 observations per task and condition). The estimated reduction was 27.8% (95% confidence interval 20.2–34.7%), with a lower one-sided bound of 21.5% and $p = 0.022$ against the 20% threshold. The quality difference was −0.19 points (non-inferiority $p = 0.053$), the error rate ratio 0.92 and the carryover test $p = 0.11$. That a single draw lands this close to both thresholds illustrates why the sample-size decision in Figure 4 matters.

9 Discussion

9.1 Expected contributions

If the hypotheses hold, the program will show that liaison text in a low-resource language can be turned into indicators that are reliable, valid and timely, and that a generative-AI assistant can shorten report production without sacrificing accuracy. If they fail, the program still documents where the pipeline breaks: in coding agreement, in composition, in predictive content or in the time cost of verification. Either outcome is informative for central banks that are deciding how to use these tools.

9.2 Policy implications

For the BOT, the indicators can feed a regional monitoring dashboard and the material prepared for monetary policy meetings, with uncertainty indices signaling when regional conditions diverge. The assistant can shift economists' time from summarizing notes toward analysis and from routine drafting toward judgment.

9.3 A responsible-AI guideline

RQ4 is answered with a guideline distilled from Studies 3 and 4: keep a human author responsible for every report; require a source for every statement; verify every number mechanically; mask identifiers before modeling; freeze and log model versions, prompts and sources; evaluate before deployment; and monitor reliance on the assistant for signs of automation bias.

9.4 Limitations

The BLP sample is purposive, and fixed effects cannot remove selection on current conditions. Histories are short, which limits the precision of nowcasting comparisons. LLMs change between versions, so results are specific to the frozen model. Synthetic validation shows that the software works, not that the hypotheses hold. The experiment measures drafting under controlled conditions, and its external validity depends on the field pilot. Finally, results from one central bank may not transfer to programs with different interview protocols.

10 Work plan

10.1 Effort allocation

Effort is split 60% academic design and 40% code (Figure 7). Academic design covers the research questions, measurement design, identification, experimental design, statistical analysis and writing; code covers the data pipeline, models, the assistant, logging tools and reproducibility. Code is concentrated in Study 1 and Study 3, where the artifacts are built; the introduction, literature and discussion carry no code.

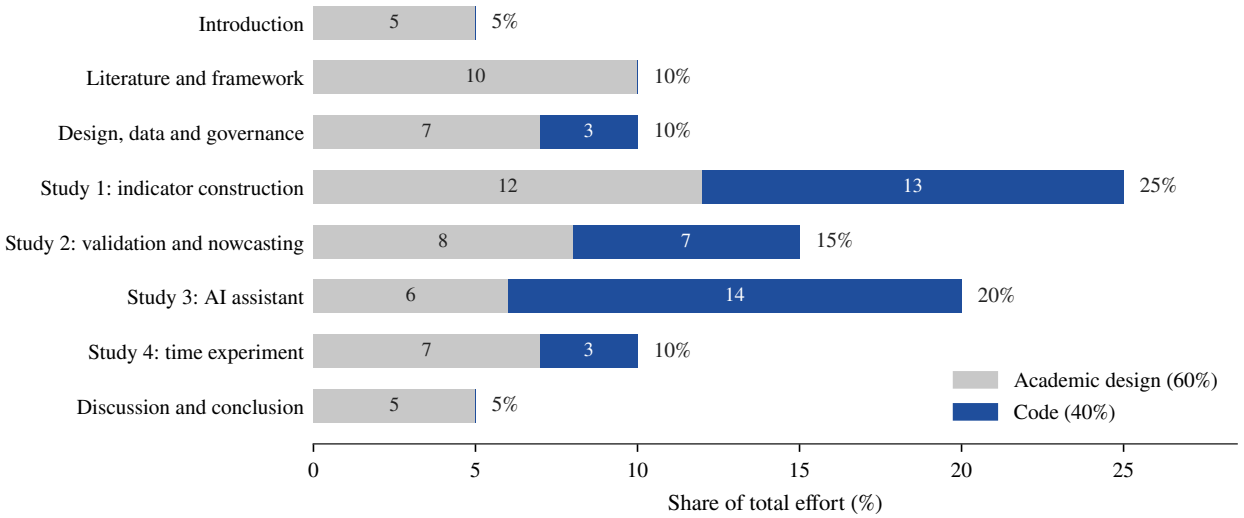


Figure 7: Planned allocation of effort by component. Bar segments show academic design and code; labels give percentage points of total effort.

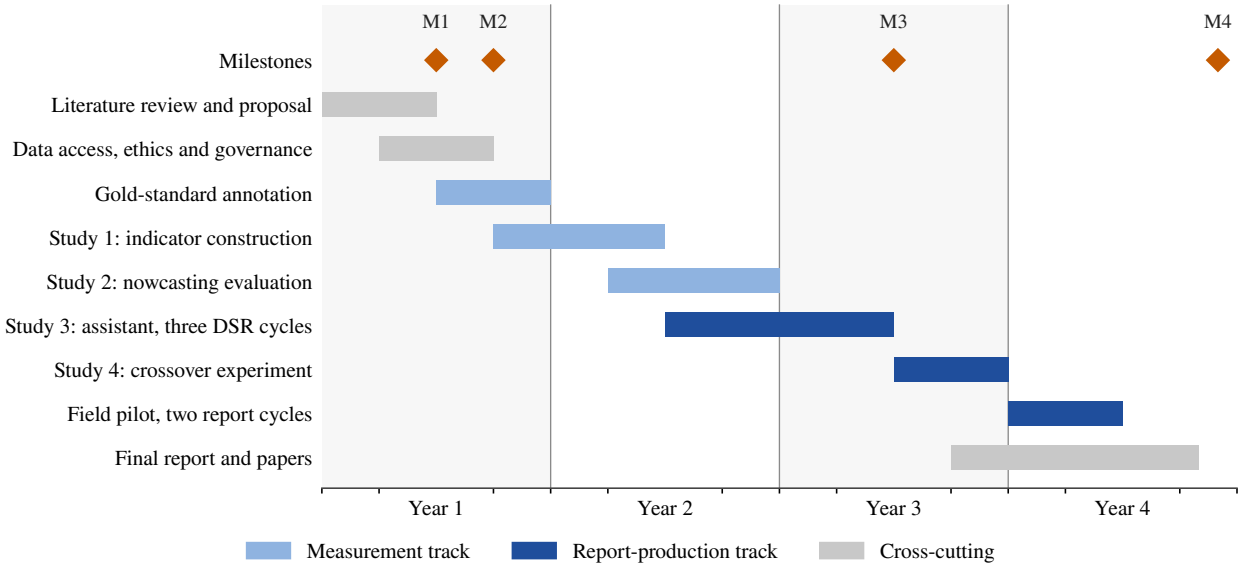


Figure 8: Four-year work plan. Bars show work packages by track; diamonds mark milestones M1–M4.

Table 10: Publication plan.

Paper	Working title	Target outlets	Year
1	Measuring business conditions from Thai liaison notes	International Journal of Central Banking; central bank working paper series	2
2	Liaison data and regional nowcasting in an emerging economy	International Journal of Forecasting; Journal of Forecasting	2–3
3	A grounded drafting assistant for central bank reports	Decision Support Systems; Government Information Quarterly	3
4	Generative AI and report preparation: a randomized crossover experiment	Management Science; Journal of Economic Behavior & Organization	4

10.2 Timeline

Figure 8 shows the four-year plan. The milestones are M1, approval of the research proposal (month 6); M2, ethics approval and the data-access agreement (month 9); M3, pre-registration of Study 4 (month 30); and M4, the final report (month 47).

10.3 Publication plan

Each study is written as a self-contained paper (Table 10), and a final report combines them with an introduction and a concluding synthesis.

10.4 Risk register

Table 11 lists the main risks with their mitigations.

Table 11: Risk register.

Risk	Likelihood	Impact	Mitigation and fallback
Firm-level notes unavailable or delayed	Medium	High	Build indices from published BLP summaries and regional report texts; run Study 4 on masked or public-text packets
Short history of digitized notes	High	Medium	Pool regions; parsimonious models; shrinkage
Inadequate Thai LLM quality	Medium	High	Verifier blocks failures; fine-tune on-premise; extractive fallback
Too few participants	Medium	High	Eight tasks per participant; analyst stratum
Confidentiality incident	Low	High	Enclave only; masking audits; no external API without approval
Model drift during the experiment	Medium	Medium	Freeze weights and prompts; log versions in the audit trail

11 Conclusion

The Business Liaison Program collects some of the most timely information available about Thailand's regional economies, but its value is limited by its form and by the effort needed to turn it into reports. This proposal sets out how to measure that information rigorously, how to test whether it improves regional nowcasts, how to build an assistant whose drafts can be audited sentence by sentence, and how to test causally whether that assistant saves more than a fifth of the time needed to prepare a report. The accompanying software already runs end to end; the next stage of the research will apply it to the BLP data.

Code and data availability

The companion code (`https://github.com/thanatchon36/blp-research`) implements every study, includes a synthetic data generator and regenerates all data-driven figures and numbers in this paper with `make figures`. BLP data are confidential and remain within the Bank of Thailand's secure environment; only synthetic data are distributed.

Declaration on the use of generative AI

A generative-AI assistant was used to draft parts of the text, code and figures under the author's direction. The author reviewed the material and takes full responsibility for the content, including the verification of all references.

References

Daron Acemoglu and David Autor. Skills, tasks and technologies: Implications for employment and earnings. In Orley Ashenfelter and David Card, editors, *Handbook of Labor Economics*, volume 4B, pages 1043–1171. Elsevier, 2011.

Michelle T. Armesto, Rubén Hernández-Murillo, Michael T. Owyang, and Jeremy Piger. Measuring the information content of the Beige Book: A mixed data sampling approach. *Journal of Money, Credit and Banking*, 41(1):35–55, 2009.

Scott R. Baker, Nicholas Bloom, and Steven J. Davis. Measuring economic policy uncertainty. *Quarterly Journal of Economics*, 131(4):1593–1636, 2016.

Nathan S. Balke and D'Ann Petersen. How well does the Beige Book reflect economic activity? Evaluating qualitative information quantitatively. *Journal of Money, Credit and Banking*, 34(1):114–136, 2002.

Marta Bańbura, Domenico Giannone, Michele Modugno, and Lucrezia Reichlin. Now-casting and the real-time data flow. In Graham Elliott and Allan Timmermann, editors, *Handbook of Economic Forecasting*, volume 2A, pages 195–237. Elsevier, 2013.

Bank for International Settlements. Artificial intelligence and the economy: Implications for central banks. In *Annual Economic Report 2024*. BIS, Basel, 2024.

John Brooke. SUS: A "quick and dirty" usability scale. In Patrick W. Jordan, B. Thomas, Bernard A. Weerdmeester, and Ian L. McClelland, editors, *Usability Evaluation in Industry*, pages 189–194. Taylor & Francis, London, 1996.

Erik Brynjolfsson, Danielle Li, and Lindsey Raymond. Generative AI at work. *Quarterly Journal of Economics*, 140(2):889–942, 2025.

Gregory C. Chow and An-loh Lin. Best linear unbiased interpolation, distribution, and extrapolation of time series by related series. *Review of Economics and Statistics*, 53(4):372–375, 1971.

Todd E. Clark and Kenneth D. West. Approximately normal tests for equal predictive accuracy in nested models. *Journal of Econometrics*, 138(1):291–311, 2007.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

Gordon V. Cormack, Charles L. A. Clarke, and Stefan Büttcher. Reciprocal rank fusion outperforms Condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 758–759, 2009.

Fred D. Davis. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 13(3):319–340, 1989.

Fabrizio Dell'Acqua, Edward McFowland, Ethan R. Mollick, Hila Lifshitz-Assaf, Katherine Kellogg, Saran Rajendran, Lisa Krayer, François Candelon, and Karim R. Lakhani. Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality. Working Paper 24-013, Harvard Business School, 2023.

Francis X. Diebold and Roberto S. Mariano. Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3):253–263, 1995.

Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. RAGAS: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158, 2024.

Matthew Gentzkow, Bryan Kelly, and Matt Taddy. Text as data. *Journal of Economic Literature*, 57(3): 535–574, 2019.

Eric Ghysels, Arthur Sinko, and Rossen Valkanov. MIDAS regressions: Further results and new directions. *Econometric Reviews*, 26(1):53–90, 2007.

Domenico Giannone, Lucrezia Reichlin, and David Small. Nowcasting: The real-time informational content of macroeconomic data. *Journal of Monetary Economics*, 55(4):665–676, 2008.

Anne Lundgaard Hansen and Sophia Kazinnik. Can ChatGPT decipher Fedspeak? SSRN Working Paper, 2023.

Peter R. Hansen, Asger Lunde, and James M. Nason. The model confidence set. *Econometrica*, 79(2): 453–497, 2011.

Sandra G. Hart and Lowell E. Staveland. Development of NASA-TLX (task load index): Results of empirical and theoretical research. In Peter A. Hancock and Najmedin Meshkati, editors, *Human Mental Workload*, volume 52 of *Advances in Psychology*, pages 139–183. North-Holland, 1988.

David Harvey, Stephen Leybourne, and Paul Newbold. Testing the equality of prediction mean squared errors. *International Journal of Forecasting*, 13(2):281–291, 1997.

J. Vernon Henderson, Adam Storeygard, and David N. Weil. Measuring economic growth from outer space. *American Economic Review*, 102(2):994–1028, 2012.

Alan R. Hevner, Salvatore T. March, Jinsoo Park, and Sudha Ram. Design science in information systems research. *MIS Quarterly*, 28(1):75–105, 2004.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55 (12):1–38, 2023.

Byron Jones and Michael G. Kenward. *Design and Analysis of Cross-Over Trials*. Chapman and Hall/CRC, Boca Raton, FL, 3rd edition, 2014.

Eleni Kalamara, Arthur Turrell, Chris Redl, George Kapetanios, and Sujit Kapadia. Making text count: Economic forecasting using newspaper text. *Journal of Applied Econometrics*, 37(5):896–919, 2022.

Klaus Krippendorff. *Content Analysis: An Introduction to Its Methodology*. SAGE, Thousand Oaks, CA, 4th edition, 2018.

J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474, 2020.

Tim Loughran and Bill McDonald. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *Journal of Finance*, 66(1):35–65, 2011.

Lalita Lowphansirikul, Charin Polpanumas, Nawat Jantrakulchai, and Sarana Nutanong. WangchanBERTa: Pretraining transformer-based Thai language models. arXiv:2101.09635, 2021.

National Institute of Standards and Technology. Artificial intelligence risk management framework (AI RMF 1.0). Technical Report NIST AI 100-1, NIST, 2023.

Brian A. Nosek, Charles R. Ebersole, Alexander C. DeHaven, and David T. Mellor. The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115(11):2600–2606, 2018.

Shakked Noy and Whitney Zhang. Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654):187–192, 2023.

Ken Peffers, Tuure Tuunanen, Marcus A. Rothenberger, and Samir Chatterjee. A design science research methodology for information systems research. *Journal of Management Information Systems*, 24(3):45–77, 2007.

Sida Peng, Eirini Kalliamvakou, Peter Cihon, and Mert Demirer. The impact of AI on developer productivity: Evidence from GitHub Copilot. arXiv:2302.06590, 2023.

Mitchell A. Petersen. Information: Hard and soft. Working paper, Kellogg School of Management, Northwestern University, 2004.

Wannaphong Phatthiyaphaibun et al. PyThaiNLP: Thai natural language processing in Python. In *Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS)*, 2023.

Gilda Piaggio, Diana R. Elbourne, Stuart J. Pocock, Stephen J. W. Evans, and Douglas G. Altman. Reporting of noninferiority and equivalence randomized trials: Extension of the CONSORT 2010 statement. *JAMA*, 308(24):2594–2604, 2012.

Lise Pichette and Lori Rennison. Extracting information from the Business Outlook Survey: A principal-component approach. *Bank of Canada Review*, pages 21–28, 2011. Autumn issue.

Kunat Pipatanakul et al. Typhoon: Thai large language models. arXiv:2312.13951, 2023.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009.

Shibley Sadique, Francis In, Madhu Veeraraghavan, and Paul Wachtel. Soft information and economic activity: Evidence from the Beige Book. *Journal of Macroeconomics*, 37:81–92, 2013.

Adam Hale Shapiro, Moritz Sudhof, and Daniel J. Wilson. Measuring news sentiment. *Journal of Econometrics*, 228(2):221–243, 2022.

Jeremy C. Stein. Information production and capital allocation: Decentralized versus hierarchical firms. *Journal of Finance*, 57(5):1891–1921, 2002.

James H. Stock and Mark W. Watson. Macroeconomic forecasting using diffusion indexes. *Journal of Business & Economic Statistics*, 20(2):147–162, 2002.

Leif Anders Thorsrud. Words are the new numbers: A newsy coincident index of the business cycle. *Journal of Business & Economic Statistics*, 38(2):393–409, 2020.

E. J. Williams. Experimental designs balanced for the estimation of residual effects of treatments. *Australian Journal of Scientific Research, Series A*, 2:149–168, 1949.

A Prompt templates

The prompts are written in Thai in the code; English translations follow.

Coding prompt (M3). *System:* You code interview notes from a central bank business liaison program. Answer only with JSON that follows the schema. *User:* Codebook, relative to the previous quarter and only for the interviewed firm: sales (sales, orders, revenue, customers, tourists, occupancy); employment (headcount, hiring, hours); investment (capital spending, machinery, capacity or branch expansion by the firm itself); price (the firm's selling prices); cost (the firm's input costs); outlook (expected business conditions). Labels: up, flat, down, not_mentioned. Rules: a topic with both increases and decreases is flat; statements about competitors, the government or world markets do not count; "not yet raised" is flat. Example and note follow.

JSON schema. An object with the six topics as required keys, each a string restricted to {up, flat, down, not_mentioned}, and no additional properties.

Drafting system prompt. You assist central bank economists in drafting a formal, concise and neutral regional economic report in Thai. Use only DATA and EVIDENCE. Write one sentence per line and end every line with at least one citation [D:key] or [N:id]. Copy every number exactly from DATA; do not compute, round or estimate. Never name firms or people or give identifying details. If evidence is insufficient, say so and cite the table checked.

B Software architecture

Table 12: Modules of the companion code and the sections they implement.

Module	Function	Section
`src/blp/synthetic.py`	Synthetic notes, gold labels and target series	8, D
`src/blp/preprocess.py`	Masking of identifiers; Thai clause splitting	3.4
`src/blp/scoring.py`	Lexicon (M1), LLM (M3) and ensemble (M4) scorers	4.2
`src/blp/reliability.py`	Cohen's κ, Krippendorff's α, macro-F1	4.5
`src/blp/indicators.py`	Diffusion, sector-weighted and composition-adjusted indices	4.3–4.4
`src/blp/nowcast.py`	Expanding-window evaluation; Clark–West and Diebold–Mariano tests	5
`src/blp/assistant/`	LLM clients, hybrid retrieval, drafter, verifier	6
`src/blp/experiment/`	Crossover randomization, power simulation, analysis, time logging	7
`run_pipeline.py`	Runs every study on synthetic data	8
`scripts/make_figures.py`	Regenerates all data-driven figures and numbers	All

C Power analysis grid

Table 13: Simulated power of the one-sided test of a time reduction above 20% (4,000 simulations per cell).

	4 tasks per participant		6 tasks per participant		8 tasks per participant	
Participants	30%	35%	30%	35%	30%	35%
10	0.27	0.52	0.35	0.61	0.41	0.69
15	0.37	0.66	0.48	0.79	0.54	0.86
20	0.47	0.79	0.57	0.88	0.66	0.94
25	0.54	0.86	0.67	0.93	0.74	0.97
30	0.61	0.92	0.73	0.97	0.81	0.99
35	0.68	0.95	0.79	0.98	0.86	1.00
40	0.74	0.97	0.83	0.99	0.90	1.00

D Synthetic data-generating process

A national activity factor and a cost factor follow AR(1) processes, $g_t = 0.6\, g_{t-1} + e_t$ with $e_t \sim N(0, 0.55^2)$ and $\pi_t = 0.7\, \pi_{t-1} + v_t$ with $v_t \sim N(0, 0.45^2)$, with a shock of -3.0 in 2020Q2 and -1.0 in 2020Q3. Regional factors are $f_{rt} = \lambda_r g_t + u_{rt}$ with loadings $(1.0, 0.8, 1.2)$ for the North, Northeast and South and AR(1) idiosyncratic terms with coefficient 0.5 and standard deviation 0.4. Tourism receives an additional shock of -1.5 for seven quarters from 2020Q2. Regional activity is $a_{rt} = f_{rt} + \sum_s w_{rs}\tau_{st}$, and the official-data proxy is $y_{rt} = 2 + 1.8\, a_{rt} + \eta_{rt}$ with $\eta_{rt} \sim N(0, 0.6^2)$.

Each region interviews 45 firms per quarter; a firm stays in the panel with probability 0.8, and entrants in quarter t draw an optimism effect $\alpha_i \sim N(0.8\, t/(T-1) - 0.4,\ 0.5^2)$. With $a_{it} = f_{rt} + \tau_{s(i)t}$, latent firm responses are $z^{\text{sales}} = a_{it} + \alpha_i + \epsilon$, $z^{\text{emp}} = 0.6\, a_{it} + 0.7\, \alpha_i + \epsilon$, $z^{\text{inv}} = 0.7\, a_{it} + 0.8\, \alpha_i + \epsilon$, $z^{\text{price}} = 0.7\, \pi_t + 0.3\, a_{it} + 0.3\, \alpha_i + \epsilon$, $z^{\text{cost}} = 0.9\, \pi_t + \epsilon$ and $z^{\text{out}} = 0.6\, a_{i,t+1} + 0.3\, a_{it} + \alpha_i + \epsilon$, with $\epsilon \sim N(0, 0.7^2)$ (0.8^2 for investment). Responses above 0.55 are coded up and below -0.55 down. Each topic is mentioned with probability 0.85 (0.90 for the outlook), activity topics are written as mixed statements with probability 0.08, and a distractor sentence is added with probability 0.25.