# Programme Documentation of the research programme ‘AI-driven knowledge production’: Bundled Management Summaries of the Methodological Reflection Reports of five AI-Generated Dissertations.

Dr. R. Schimmel

21 August 2026

# Preface

This volume brings together the Management Summaries of the five Methodological Reflection Reports accompanying the pentalogy *AI-Driven Knowledge Production*. The pentalogy consists of four field-specific dissertations and a meta-dissertation in which the substantive results and production histories of those dissertations are used as case material. Together, the studies examine AI-driven knowledge production within design-oriented, anthropological, historical and natural-science research, while the meta-dissertation develops the comparative and reflexive framework connecting the individual cases.

The summaries are presented in the order in which the research programme developed. *Polyglot Learning in the AI Era* was the first case, followed by *Governing within Biological Boundaries*. The first two versions of the meta-dissertation *AI-Driven Knowledge Production* were produced on the basis of those two cases. The historical and natural-science dissertations were developed subsequently, after which the meta-dissertation was revised to incorporate the complete pentalogy. Its position as the third document in this volume therefore reflects the chronology of production rather than the final analytical hierarchy of the programme.

All five cases share a deliberate separation between textual production and academic responsibility. The dissertation texts were generated by language models. The original research ideas, research architectures, central conceptual choices and criteria for acceptance originated with the human curator. The curator specified tasks, selected and rejected generated material, maintained structural and terminological consistency, organised criticism and verification, and retained final scientific responsibility.

Language models were assigned different functions within the respective production architectures. These included exploration of unfamiliar domains, theoretical dialogue, drafting, revision, criticism, bibliometric verification, analysis of source use, dataset construction and production of research or instructional materials. Such functions are treated throughout the programme as assigned roles rather than as autonomous forms of authorship or inherent properties of particular models. Their outputs acquired a place within the research only through curatorial assessment and selection.

The Methodological Reflection Reports are therefore not supplementary accounts of how otherwise conventional dissertations happened to be written. They constitute a second analytical layer of the research programme. Whereas the dissertations present the resulting arguments, designs and empirical constructions, the reflection reports reconstruct the processes through which those results were produced. They document successive versions, interaction histories, model assignments, curatorial decisions, production failures, critical interventions and verification procedures. In several cases, production, criticism and verification themselves became objects of investigation.

The cases were not designed to produce uniform outcomes. The design-oriented case developed into an unfinished but methodologically productive interaction between a dissertation, a learning book and the system required to produce that learning book. The anthropological case exposed the possibilities and risks of constructing cross-domain explanations from literature

originating in different disciplines. The meta-dissertation made curatorial practice and high-speed academic production directly observable. The historical case revealed differences between linear text production and recursive analytical tasks, as well as the limitations of language models in forensic reconstruction. The natural-science case used scientific illusionism to examine how a conventionally credible empirical study could be constructed when both its text and its data were synthetic.

To make these differences comparable, every summary follows the same sequence of questions: what was produced; what the reflection report contains; how the dissertation was produced; how human and AI roles were divided; what went wrong; what was remarkable; how reliability was controlled; and what the case contributed to understanding AI-driven knowledge production. The shared structure is intended to facilitate comparison without reducing the methodological specificity of the individual cases.

Reliability is not treated in these reports as a property of a language model or as a guarantee that every generated statement is correct. It is organised through production architecture: human selection, functional separation, adversarial criticism, version control, bibliometric checking, verification of source use and, where applicable, verification of datasets or instructional materials. The reports also show that verification procedures may themselves produce omissions, overstatements or plausible but incorrect assessments. Verification output therefore remains subject to procedural control and human interpretation.

These Management Summaries provide a structured entry point into that documentary material. They do not replace the full reflection reports, their appendices or the underlying interaction records. Quantitative findings are specific to the cases in which they were obtained and should not be read as general performance estimates for language models. Equally, a conceptually developed but unfinished research design should not be confused with a completed empirical study, and synthetically constructed data should not be interpreted as observational findings.

Taken together, the five reports show that the production of complete academic manuscripts can be radically accelerated, but that generation speed does not remove the need for research architecture, criticism, documentary control and accountable judgement. The central question is therefore not whether a language model can produce academically plausible text. It is how the complete production system is designed, which functions are separated, how errors are detected, and where responsibility for the resulting knowledge claims is located.

Dr. Remco Schimmel,

Voorburg,

21 August 2026

# Table of Contents.

# 1. Management Summary - Polyglot Learning in the AI Era: Design and Testing of an AI-Driven Simultaneous Learning Programme for the Acquisition of Related Languages

## 1. What was produced?

*Polyglot Learning in the AI Era: Design and Testing of an AI-Driven Simultaneous Learning Programme for the Acquisition of Related Languages* is the design-oriented case within the pentalogy *AI-Driven Knowledge Production*. The study concerns the simultaneous acquisition of Spanish, Italian and Brazilian Portuguese through pattern recognition, interlingual comparison and AI support. It produced a dissertation, an associated learning book in development, and a production architecture for generating and verifying the required exercise material.

The dissertation reached the stage of a conceptually and methodologically developed research design. The learning programme has not yet been implemented and the learning book has not yet been completed. The study must therefore be distinguished from a completed empirical evaluation of the learning design.

## 2. What's in the Methodological Reflection Report?

The report reconstructs the interdependent development of the dissertation and the learning book. It describes how AI roles were progressively differentiated, how production problems led to increasingly explicit production protocols, and how three control regimes were introduced: adversarial review of the dissertation, verification of references and source use, and verification of the exercise material. A separate chapter, *Musings*, preserves several dialogues that emerged outside the formal reconstruction. The appendices contain the operational evidence base: a complete bibliometric audit, a detailed source-use verification, and the most recent prompts used for bibliographic verification, source-use verification, exercise-material production, and exercise-material verification.

## 3. How was it produced?

The dissertation developed through nine documented phases. The process began with exploration of a field that was largely new to the curator. The research problem, theoretical framework and methodological design were then progressively delimited, structured and revised. The learning book initially followed from the dissertation as the instrument required to operationalise the proposed study. Its development subsequently revealed weaknesses and ambiguities in the research design and therefore also informed further revisions of the dissertation.

The process was highly iterative. AI-generated academic prose could be produced and revised efficiently, but maintaining coherence across the full dissertation required continuous curatorial oversight. Producing the learning material proved more demanding, as its quality depended not only on local textual correctness but also on systematic consistency across languages, lexical items, contexts and exercise structures.

A second AI model was eventually used as an independent reviewer of the dissertation. Its output was not incorporated automatically. The curator evaluated the critique and translated selected points into targeted revision instructions when they revealed relevant theoretical or methodological weaknesses.

## 4. How were human and AI roles divided?

All dissertation text was generated by AI systems. The original research idea, the selection of languages, the core instructional principles and the overall research architecture originated with the human researcher. The curator defined requirements, safeguarded the central argument, selected and rejected generated material, coordinated critique and verification procedures, and retained final scientific responsibility.

AI systems were used for domain exploration, conceptual dialogue, structuring, drafting, revision, instructional design and production of learning materials. A second AI model was assigned a dedicated critical-review function. These functions were not treated as inherent model properties but as role assignments within the production architecture.

The same division of roles was maintained during revision. Rejected passages were not edited directly by the curator but re-entered into the generation cycle. In the future empirical phase, the human researcher will additionally act as research participant and source of auto-ethnographic data.

## 5. What went wrong?

During the early development phases, dissertation text was generated largely at the level of individual paragraphs and sections. The report records that useful local passages could be produced, while the argumentative relationships between sections, document-wide terminology and the overall structure still had to be organised by the curator. Later versions therefore shifted from expansion towards terminological harmonisation, internal cross-referencing and explicit control of coherence across the dissertation.

The main production problem emerged in the learning book. Generating large volumes of interrelated multilingual material under multiple simultaneous constraints proved difficult to control. Individually acceptable items could lose cross-item comparability, shift in meaning across languages, or fail to fulfil their intended instructional function. Iterative production therefore revealed not only execution errors but also previously implicit requirements in the learning design itself.

Independent AI-based critique introduced a second-order problem. The reviewer model could identify relevant weaknesses but also produced critiques that were overly general, overly categorical, or insufficiently aligned with the research design. Critical review therefore required curatorial interpretation rather than direct implementation.

Subsequent audits confirmed that overall theoretical coherence could coexist with local documentary failures. Thirteen of the 98 bibliographic records could not be verified in the form in which they were cited. The source-use verification further identified attribution errors, conceptual compression, and insufficient differentiation between literature-derived findings and newly constructed syntheses.

## 6. What was remarkable or extraordinary?

The chapter *Musings* emerged from the need for a space in which ideas were not immediately converted into dissertation prose. It contains dialogues on the risk that a system used to produce and support the study may also appear to validate its own outputs; the tension between explicit grammar instruction and pattern-based learning; the operational conditions for AI-supported scaffolding; and an unsolicited model-generated reflection triggered by the single-word prompt *reflections*. The chapter is explicitly optional, and the final model-generated passage has no methodological status. The dialogues are preserved largely in their original form.

The development of the learning book also altered the epistemological status of the project. What began as the production of an instructional artefact evolved into the design of a production system capable of generating that artefact under controlled conditions. The learning book and its production architecture therefore developed as two interdependent design objects.

A further distinctive feature of this report is the combination of an early AI-generated dissertation trajectory with verification procedures developed and refined at later stages. The appendices therefore document not only the original production process but also the progressively increased level of methodological control within the broader research programme.

## 7. How was reliability restored or controlled?

Reliability was governed through three distinct but interrelated control regimes.

The first was **adversarial control of dissertation content**. Production and critique were separated by assigning an independent review function to a second AI model. This created structured opposition to the assumptions and theoretical commitments embedded in the primary generation process. The curator determined which critiques were methodologically relevant and how accepted critiques were integrated into the research design.

The second was **verification of references and source use**. Bibliometric verification assessed whether cited publications could be identified as existing documents. Source-use verification addressed a different question: whether those publications actually supported the argumentative

functions assigned to them in the dissertation. This procedure distinguished local attribution from global argumentative support and separated literature-derived claims from hypotheses, analogies, design decisions and newly constructed syntheses. The appendices provide the corresponding evidence registers and integrated assessments.

The thirteen untraceable references were not corrected in either the Dutch or English versions of the dissertation. They were retained as evidence of the generative process itself. The audit therefore increased transparency while avoiding retrospective reconstruction of an error-free text.

The third regime was **verification of exercise-material validity**. Production and verification were treated as separate pipelines. The production protocol defined the constraints for generating multilingual items, while the verification protocol evaluated the resulting dataset against those constraints and required all findings to be traceable to observable evidence. The final sections of Appendix C present both protocols in parallel. Both remain provisional, as further scaling of the learning book may reveal additional constraints or failure modes.

Together, these three regimes address distinct objects: argumentative coherence, documentary validity, and instructional validity. None is reducible to the others.

## 8. What did this case teach about AI-driven knowledge production?

The case demonstrates that full AI generation of dissertation text does not constitute autonomous knowledge production. The role of the human researcher shifts from authoring text to designing research architecture, specifying tasks, selecting outputs, evaluating critique, conducting verification, and assuming scientific responsibility for the resulting claims.

The development of the learning book further shows that rapid and low-cost generation does not eliminate the need to define what counts as a valid artefact. When design criteria are incomplete, iteration does not merely correct outputs; it also makes implicit requirements explicit. The central challenge is therefore not only generation, but the specification and governance of the generative space.

The case also shows that different activities require different degrees of generative constraint. Exploration and critique benefit from openness and variability. Controlled production and verification require strict constraints, stable criteria and traceable evidence. These functions must therefore be institutionally and procedurally separated.

Finally, the audits demonstrate that generative recombination can simultaneously produce documentary errors and novel conceptual or design insights. Evaluation of AI-driven knowledge production therefore requires separate attention to global argumentative coherence, local documentary accuracy, and the explicit boundary between established knowledge and newly constructed synthesis.

# 2. Management Summary - Governing within Biological Boundaries: The Role of Biological Constraints in Addressing Persistent Problems in Public Administration.

This report is relevant because it reconstructs how an initially persuasive but insufficiently controlled AI-generated dissertation led to the development of a multi-agent control architecture. It documents not only the production of the dissertation, but also the criticism, verification and failures of verification through which the work acquired its later form.

## 1. What was produced?

*Governing Within Biological Boundaries: The Role of Biological Constraints in Addressing Persistent Problems in Public Administration* examines whether persistent policy failures can be understood more fully by taking evolutionarily shaped human behaviour into account. Ten biological constraints are applied to six cases involving class, racial and gender conflict. The study is qualitative and abductive: patterns observed in the cases are compared with the proposed theoretical framework without treating biological mechanisms as deterministic laws.

Within the pentalogy *AI-Driven Knowledge Production*, the dissertation represents the anthropological archetype. It is not a conventional ethnographic study and involved no participant observation or fieldwork. Its anthropological character lies in its interpretation of human behaviour in groups and of the biological, social and institutional mechanisms involved. The dissertation text was generated by language models under human curation.

## 2. What's in the Methodological Reflection Report?

The report reconstructs the progression from linear single-model production to an iterative and subsequently role-differentiated multi-agent architecture. It examines human–machine and machine–machine interaction, the production and checking of the empirical cases, adversarial review of the full dissertation, source verification and several anomalies encountered during verification. A separate chapter contains Claude's own analysis of the interaction in which it had participated.

The annexes constitute the report's documentary foundation. They contain the control prompts, the provenance of chapters across successive versions, the selection and assessment of key sources, verbatim evidence of interactional singularities and the unfiltered output of a bibliometric verification whose internal contradictions were deliberately preserved.

## 3. How was it produced?

Version 1 was generated largely as a linear sequence of chapters, with Claude functioning primarily as text generator. The resulting text was structurally coherent and stylistically persuasive, but insufficiently controlled. Rhetorical inflation, one-directional reasoning, weak qualification of causal claims and inadequate source accountability could accumulate because production was not interrupted by systematic counter-reading.

Version 2 introduced a recurring cycle of task-separated roles. Claude generated or revised draft passages and case descriptions; ChatGPT analysed the material, checked apparently factual claims, identified inconsistencies and challenged the argumentation; the curator assessed the findings and routed selected corrections back into the production process. This procedure improved the substantive quality of the empirical analyses. At this stage, the extensive case material remained in annexes, while Chapters 7–9 contained more condensed treatments of the same cases.

Version 3 developed this arrangement into a differentiated multi-agent architecture. A blind review of the preceding dissertation was first conducted by Claude without disclosure that the model had contributed to the text under review. The resulting critique was used to direct a broader reconstruction. The theoretical and synthesising chapters were rewritten through one model configuration. The empirical chapters were reconstructed by removing repetition and ballast and integrating the relevant material from the six case annexes into Chapters 7–9. This consolidation was an editorial reorganisation of the dissertation, not the correction of an inconsistency caused by adversarial control. A final round of scrutiny removed questionable claims and unnecessarily suggestive formulations.

## 4. How were human and AI roles divided?

The human researcher selected and framed the research problem, organised the theoretical and empirical architecture, assigned roles to the models, evaluated criticism, decided which interventions were warranted and retained final scientific responsibility. The human role developed from correcting outputs after production to anticipating recurrent model failures and specifying the conditions under which generation, criticism and revision were to take place.

Language models were assigned distinct functions rather than being used interchangeably. Claude was used as generator, reviser, structural executor, blind reviewer and, in a separately demarcated chapter, analyst of the human–machine interaction. ChatGPT was used as critical reader, fact-checker, examiner of argumentative lines and later as a theoretical sparring partner. Gemini was subsequently used for bibliometric verification. Text generation and revision remained model tasks; the curator governed the process through selection, adjudication and control.

## 5. What went wrong?

The central failure was narrative over-coherence. The initial production process could transform theoretical assumptions into coherent and persuasive case narratives without a corresponding strengthening of their empirical foundations. Confirmation of the biological framework was foregrounded, while counterarguments, uncertainties and alternative explanations received less attention. The problem was not incoherent prose, but prose whose apparent coherence concealed unsupported claims, factual inaccuracies, speculative extrapolations and excessive certainty about causal relationships.

Other problems followed from the production process. Regeneration could replace carefully curated formulations with generic academic prose. Fragmentation across sessions contributed to theoretical drift, while multi-model revision could produce patchwork integration and terminological instability. Instructions to improve a text could also trigger unsolicited elaboration, additional complexity or unverifiable statistics rather than controlled correction.

Verification created its own failures. Source searches and selection processes were not always transparent, and assessments of source use remained proxy assessments because the models did not have direct access to all original texts. During the bibliometric audit, Gemini initially described its procedure as systematic checking against external databases and presented highly reassuring results. When questioned about the method, it acknowledged that no live database queries had been made and that references had instead been compared with internal model knowledge. The apparent verification had therefore generated a new form of plausibility construction.

## 6. What was remarkable or extraordinary?

A separate chapter places the participating language model's own analysis of the human–machine interaction within the report. The transition to this chapter is explicitly marked as a transition from the curator's reconstruction to Claude's view. Claude is positioned simultaneously as contributor to the dissertation, instrument of analysis and object of that analysis. The chapter is not presented as independent verification or as the curator's conclusion.

The retained output also shows the limits of this arrangement. While analysing overproduction and validation-seeking, the model produced a substantially overlong first version and subsequently requested confirmation of whether its own assessment was fair. These events were incorporated into the chapter rather than removed from the record. The reader is therefore given access both to the model-authored interpretation and to behaviour that qualifies that interpretation.

A second unusual procedure involved the blind review of the complete dissertation. The absence of the earlier contribution from Claude's active context was deliberately used to create distance between generation and review. Different evaluative behaviour was produced when the same model was assigned the role of reviewer rather than writer. The report records this as an effect of role assignment, not as evidence of autonomous judgement.

## 7. How was reliability restored or controlled?

Reliability was addressed through several related but distinct controls.

The first concerned the factual and argumentative quality of the empirical cases. Draft case narratives were subjected to an iterative cycle of model-generated criticism, fact-checking and revision under human direction. Claims, statistics, causal inferences and links between the cases and the theoretical framework were repeatedly examined. This control was introduced because plausible narrative synthesis had proved insufficient as a basis for case analysis.

The second concerned the dissertation as a whole. Generation and review were separated, and an adversarial assessment of the complete text was used to identify confirmation bias, reductionism, insufficient qualification, theoretical inconsistency and limitations in the research design. The curator determined which criticisms were applicable and converted them into bounded revision assignments.

The third concerned references and source use. Key sources were identified according to their function within individual chapters and were assessed for their contribution to the relevant lines of argument. Bibliometric verification addressed the separate question of whether references existed in the cited form. The report nevertheless makes clear that model-based source-use assessment without full-text access remains a proxy and that bibliometric verification based on internal model knowledge cannot be equated with documentary database research.

Reliability was therefore not restored by a single definitive audit. It was controlled through organised contradiction, separated roles, repeated human adjudication, provenance records and disclosure of unresolved limitations. The preservation of the contradictory Gemini outputs constitutes verification of the verification: the control instrument itself was made part of the audit trail.

## 8. What did this case teach about AI-driven knowledge production?

The case shows that interpretative research is particularly vulnerable to AI-generated narrative over-coherence. Language models can elaborate assumptions into increasingly consistent explanations, while empirical friction, counterevidence and alternative interpretations recede from view. Persuasive formulation must therefore be distinguished from evidential strength.

The case also shows that reliability depends less on the presumed quality of an individual model than on the organisation of the production process. When generation, interpretation and evaluation are assigned to the same system within the same interactional setting, existing assumptions may be reinforced. Separating generative, critical and verificatory roles creates organised opposition, but does not remove the need for human judgement.

Role assignment materially affected the outputs produced. The same model displayed different behaviour when used as writer, reviewer, executor or analyst. Multi-agent production should therefore not be understood simply as the addition of more models, but as the deliberate allocation of functions within a controlled research architecture.

Finally, the case shows that verification procedures can themselves reproduce the problem they are intended to control. A model may produce a plausible account of verification without having performed the documentary checks implied by that account. Reliability consequently requires not only verification of generated knowledge, but also inspection of the provenance, method and limitations of the verification itself.

# 3. Management Summary - 'AI-Driven Knowledge Production: A Study of the Impact of Artificial Intelligence on Scientific Knowledge Production'.

## 1. What was produced?

*AI-Driven Knowledge Production: A Study of the Impact of Artificial Intelligence on Scientific Knowledge Production* is the meta-dissertation within a pentalogy of five AI-written dissertations. Four underlying dissertations examine AI-driven knowledge production within design-oriented, anthropological, natural-scientific and historical research. Their substantive results and methodological reflection reports provide the case material for the meta-dissertation. The production of the meta-dissertation itself constitutes an additional reflexive case.

Three versions of the meta-dissertation were produced. Versions 1 and 2 were written when only the design-oriented and anthropological cases had been completed. Version 3 incorporated the natural-scientific and historical cases and completed the intended comparative architecture.

All dissertation text was generated by language models. The accompanying Methodological Reflection Report documents the human curatorial process, the allocation of model functions, version development, recurrent production issues, and verification procedures. Its purpose is not to extend the dissertation's substantive argument, but to make the production process transparent, traceable, and assessable.

## 2. What's in the Methodological Reflection Report?

The report contains seven chapters and five substantive appendices. Chapter 1 defines its purpose and scope and sets out the central authorship claim: all dissertation text was generated by language models, while the research design, direction, selection, evaluation and academic responsibility remained with the human curator. Chapter 2 reconstructs the production of the three dissertation versions and distinguishes the generation of Version 1, its blind review, the targeted production of Version 2 and the later completion of Version 3. Chapter 3 examines how different functions were assigned to ChatGPT, Claude and Google Gemini and why the effectiveness of a model depended partly on its assigned role and the delimitation of its task.

Chapter 4 identifies the recurrent risks associated with AI-generated academic writing and the measures used to control them. These include overproduction, local coherence without sufficient global coherence, regeneration of curated text, paraphrasing, terminological drift, incomplete processing of extensive documents, cross-version contamination and the asymmetry between production and verification speed.

Chapter 5 contains the report's most distinctive material. The complete ChatGPT interaction behind Version 1 was analysed retrospectively using Claude. The resulting analysis reconstructs how the curator formulated instructions, rejected unsuitable output, protected

accepted text, maintained scope and terminology, exposed methodological weaknesses and corrected errors. Its detailed observations are subsequently condensed into two general principles: structural discipline and process discipline.

Chapter 6 reports two separate verification procedures. Source-use analysis examines the argumentative function and weight of the literature used in the first eleven chapters and identifies where further human judgement may be required. Bibliometric verification classifies one hundred references according to their existence, completeness and bibliographic accuracy. The chapter also discusses the possibilities and limitations of using language models as a scalable proxy for forms of control that would otherwise require extensive manual work. Chapter 7 brings together the conclusions concerning textual authorship, curatorial responsibility, model limitations, verification and process transparency.

The appendices provide the supporting evidence and operational detail. Appendix 1 reconstructs the three dissertation versions and their associated production histories. Appendix 2 reproduces concrete curator interventions from the ChatGPT interaction behind Version 1. Appendix 3 contains the source-use verification protocol. Appendix 4 presents the unfiltered chapter-by-chapter results of that procedure. Appendix 5 contains the consolidated bibliometric audit. The complete chat histories, comprising 416,623 words and 1,368 pages, are not included in the report but remain available as underlying documentation.

## 3. How was the dissertation produced?

Version 1 was produced incrementally within a single continuous ChatGPT conversation. The curator provided structured instructions at the level of chapters, sections, and subcomponents, evaluated each generated passage, and determined whether production should continue. Previously generated text remained available within the same conversational context. The resulting manuscript of 44,501 words was completed within two days.

After completion of Version 1, Claude was used for a blind review of the manuscript. No information was provided about its production process. This review identified gaps in argumentative substantiation, uneven conceptual development, and insufficiently supported generalisations.

Claude was then assigned a second function: the targeted revision of Version 1. Revisions were constrained to the existing structure and did not involve a full rewrite. Due to practical limits on conversation length, the revision process was distributed across ten successive sessions. Segmented text inputs and explicit correction instructions were used to minimise unintended regeneration of previously accepted material. This process resulted in Version 2.

Versions 1 and 2 were necessarily provisional, as two of the four underlying case studies had not yet been completed. After the natural-scientific and historical dissertations became available, Version 3 was produced. ChatGPT was again used for text generation, Claude for critical evaluation and assessment, and Google Gemini for bibliometric verification.

A third, analytically distinct use of Claude occurred outside the production and revision cycle. The complete ChatGPT interaction underlying Version 1 was retrospectively analysed to identify recurring patterns in curatorial instruction, correction, and decision-making. This

analysis concerns the practice of curatorship, not the substantive validity of the dissertation or the bibliographic accuracy of its sources.

## 4. How are human and model roles defined?

The curator is defined as the agent responsible for research design, conceptual and methodological direction, task specification, evaluation and selection of outputs, consistency control, and decisions regarding continuation, revision, or verification. Final academic responsibility remains with the human curator.

Language models are defined as functional production and analysis systems. They generate text, perform structured revisions, and provide analytical or evaluative outputs when assigned. However, they do not determine research direction, acceptability of results, or final inclusion in the dissertation.

ChatGPT was primarily used for the generation of Versions 1 and 3. Claude performed three distinct functions: blind review of Version 1, targeted revision leading to Version 2, and retrospective analysis of curatorial interaction patterns in Version 1. These functions are analytically separate and must not be conflated: they correspond respectively to manuscript evaluation, structured revision, and meta-analysis of curatorial practice.

Google Gemini was used for bibliometric verification. Source-use analysis constituted a separate procedure in which model output was used to identify potential distortions in the argumentative use of literature.

The allocation of model functions follows a principle of functional differentiation rather than model hierarchy. Across the broader pentalogy, these roles are not fixed; they may be reassigned depending on the analytical task.

## 5. What went wrong?

At the level of the overall production process, no major failure occurred. By the time work on the meta-dissertation began, the curator had already completed the first two dissertations and had learned how to avoid the principal problems associated with large-scale AI-generated academic writing. The research architecture, task boundaries, terminology and argumentative progression were therefore controlled from the outset. This accumulated curatorial experience largely explains why Version 1 could be produced within two days as an immediately usable academic draft.

The principal remaining technical problem concerned the modification of already curated text. In the model configuration available at the time, requests for limited changes frequently caused ChatGPT to regenerate complete passages. Accepted formulations, references and argumentative relationships could then be altered or lost. This behaviour was described within the project as “GPT disease”.

Claude was therefore selected to produce Version 2 because targeted modification of existing text could at that time be controlled more effectively in that environment. The manuscript was supplied in delimited fragments, accompanied by specific revision instructions and the explicit requirement that the existing structure should be preserved. The division of the work across ten conversations resulted partly from Claude's practical conversation-length limits, but also helped to prevent wholesale regeneration. The production of Version 2 should therefore be understood as a preventive model-allocation decision rather than as recovery from a failed first version.

The blind review preceding Version 2 identified gaps in argumentative substantiation, uneven conceptual development and insufficiently supported generalisations. These findings did not make Version 1 unusable; they defined the targets for a controlled second version.

The other limitations discussed in the report—overproduction, local coherence without global coherence, terminological drift, off-topic elaboration, cross-version contamination and incomplete processing of extensive documents—were principally risks already recognised from the earlier dissertations. Where they appeared locally, they were intercepted through structural discipline, process discipline and immediate curatorial correction before they could destabilise the dissertation as a whole.

The one structural difficulty that could not be removed by experience alone was the asymmetry between production and verification speed. Academic text could be generated much faster than its argumentation, source use and bibliographic records could be checked. Quality control therefore remained the most labour-intensive part of an otherwise exceptionally efficient production process.

## 6. What was remarkable or methodologically significant?

The production of Version 1 and its subsequent retrospective analysis are analytically interdependent. The key finding is not merely the speed of production, but the fact that the complete interaction history enabled reconstruction of the curatorial mechanisms that produced it.

Claude's retrospective analysis identified recurring features of effective curatorial practice: concise and directive instruction, rejection of irrelevant elaboration, active maintenance of scope and terminology, protection of validated text from unnecessary regeneration, continuous methodological interrogation, and immediate correction of inconsistencies.

This analysis is treated as empirical case material rather than a general theory of curatorship. Its more categorical formulations are interpreted as descriptive compression rather than normative claims. The report therefore consolidates the findings into two core dimensions of curatorial practice:

- **Structural discipline**: the maintenance of conceptual architecture, terminological consistency, and argumentative hierarchy across extended text production.
- **Process discipline**: the control of task boundaries, prevention of unnecessary regeneration, and continuous validation of outputs before further production.

These dimensions render curatorship empirically observable. The two-day production of Version 1 is therefore not evidence of autonomous model authorship, but of a tightly structured interaction system in which curatorial expertise, architectural predefinition, and continuous decision-making significantly compress production time.

The dissertation is also reflexive in design: AI-generated text is used to analyse AI-driven knowledge production, while the production process itself becomes part of the empirical material. This transforms the writing process from a background condition into an object of analysis.

## 7. How was reliability ensured?

Reliability is defined as a property of the production process rather than of any individual model output. It is achieved through structured curatorial control, functional separation of tasks, and iterative validation.

Each generative or revision step remained subject to human selection, rejection, or continuation. Model-generated criticism did not enter the dissertation automatically but required curatorial evaluation. This ensures that evaluation and authorship remain distinct functions.

The production cycle maintained a strict separation between generation by ChatGPT, evaluation and revision using Claude, and bibliometric verification using Google Gemini. This separation prevented any single model from simultaneously producing, revising, and validating the same content.

Source-use analysis and bibliometric verification addressed different dimensions of reliability. Source-use analysis identified potential distortions in argumentative reliance on literature. Bibliometric verification assessed the existence and correctness of references. Both functions were interpretative and required human assessment of their outputs.

Bibliometric verification classified 91 per cent of references as correct, 5 per cent as correct after minor correction, 2 per cent as incomplete, and 2 per cent as inconsistent or untraceable. The resulting 96 per cent rate of correct or repairable entries was interpreted as a strong but non-definitive indicator of bibliographic reliability.

A general limitation of model-based verification is the risk of circular confirmation where evaluation is performed by systems sharing partly comparable linguistic and pattern-recognition characteristics. This does not invalidate verification but constrains its methodological status. Model-based verification was therefore treated as support for human judgement rather than as an independent authority.

Reliability was thus produced through process architecture: role separation, version control, constrained revision, documentation, and accountable curatorial interpretation.

## 8. What does this case show about AI-driven knowledge production?

The case demonstrates that AI-assisted academic production is primarily a function of curatorial design rather than model autonomy. High-speed dissertation production is possible only when task specification, conceptual architecture, and evaluative control are tightly integrated.

The curator's role extends beyond prompting to include research design, conceptual structuring, decomposition of tasks, evaluation of outputs, and management of iterative revision cycles. This role is best described as curatorial rather than textual authorship.

The case further shows that textual authorship and academic responsibility can be separated. While all dissertation text was generated by language models, responsibility for structure, validity, and interpretation remained with the human curator.

Model performance was less dependent on model identity than on functional assignment. The same model could produce, revise, or critique depending on its role within the process architecture. Consequently, system design was more decisive than model selection.

Finally, AI-based verification proved useful but inherently limited. It enabled scalable control but did not replace human judgement. Reliability in AI-driven knowledge production was therefore not a property of the model, but an emergent property of disciplined curatorial practice, structured workflows, and accountable interpretation.

# 4. Management Summary - The Promises of Digital Revolutions: A Historical-Analytical Study of IT Predictions and Societal Transformations in the Netherlands (1945–2025)

## 1. What was produced?

*The Promises of Digital Revolutions: A Historical-Analytical Study of IT Predictions and Societal Transformations in the Netherlands (1945–2025)* is the historical case within the pentalogy *AI-Driven Knowledge Production*. The dissertation examines the extent to which predictions concerning IT-driven societal transformation can be assessed retrospectively and which characteristics are associated with structurally adequate predictions.

A canon of Dutch digital-transformation episodes was constructed and organised through a 2×2 matrix combining predictive quality with the realised extent of transformation. The cases therefore include not only accurately predicted and realised transformations, but also adequate predictions followed by limited realisation, underestimated transformations, and inadequate predictions followed by failure of durable embedding. The substantive conclusion is that robust prediction depends less on visionary reach or temporal precision than on proportionate claims concerning scale, explicit causal assumptions and clearly specified institutional boundary conditions.

All substantive dissertation text was generated by language models under human curatorial direction. The completed manuscript comprised approximately 49,000 words. Versions 0, 1.0 and 1.1 were produced in just under two weeks. A subsequent analytical correction identified during translation was incorporated into both the English text and Dutch version 1.2.

## 2. What's in the Methodological Reflection Report?

The report documents the practical production architecture underlying the dissertation. It describes the research design, the generative workflow, the division of roles between the curator and the language models, the selection of historical episodes, the verification of references and source use, the documentation of the interaction history, the limits of forensic reconstruction and the stylistic normalisation of the manuscript.

A separate chapter records operational lessons concerning regenerative overwrite, terminological and semantic drift, context limitations, incomplete processing of large documents, the difference between generation and criticism, and the conditions under which targeted revision can remain reliable. It also includes a comparative analysis of two interaction sequences: the relatively linear development of Chapters 1–3 and the recursive canon-formation process for Chapter 4.

The report further examines anthropomorphic perceptions arising during prolonged interaction with language models. It distinguishes the appearance of agency from the probabilistic generation of text and considers how evaluative or emotional language can nevertheless function as contextual information and as a marker within an interaction history.

The appendices contain a bibliometric audit of all 91 references and a chapter-by-chapter analysis of source use, argumentative dependency and systematic vulnerability. The prompt used for the source-use analysis is also included. The report is procedural rather than substantive: its purpose is to make the production, criticism and verification of the dissertation transparent and assessable.

## 3. How was it produced?

Version 0 was generated with ChatGPT through a sequence of chapter- and section-level interactions. The theoretical and methodological foundation of Chapters 1–3 was initially organised in the form of a research proposal. Historical episodes were then identified and classified, and the empirical chapters were developed through repeated application of the claim-oriented, transformation-oriented and causal lenses.

A Claude-based review was subsequently used to evaluate version 0 for argumentative coherence, analytical clarity and academic formulation. The most consequential finding concerned contamination between predictive quality (X) and realised transformation (Y). Information about subsequent implementation had entered the reconstruction and assessment of the original predictions. Because historical knowledge of the outcome had thereby influenced the assessment of prior predictive quality, substantial revision was required.

Version 1.0 was produced through a new series of interactions with ChatGPT in which this methodological defect and other structural issues were addressed. The revised manuscript was then subjected to a second Claude-based review. Version 1.1 resulted from stylistic normalisation. Reformulations were generated by ChatGPT, while selected corrections were implemented manually by the curator to prevent accepted content from being lost through unnecessary regeneration.

The English translation provided a further control moment. During the translation of the Q3 analysis, an analytical imprecision was identified in the classification and synthesis of several episodes. The correction was incorporated into both the translation and Dutch version 1.2.

The production history comprises 581,733 words and 3,019 pages of archived interaction records. The final manuscript represents approximately 8.3% of that corpus. Production was therefore rapid, but not linear or automatic: most generated material consisted of exploration, rejected alternatives, corrections and intermediate formulations.

## 4. How were human and AI roles divided?

The curator determined the research problem, research questions, analytical framework, methodological design and argumentative architecture. The criteria for canon formation and the final selection of historical episodes also remained human decisions. AI-generated longlists were used as exploratory material, but were assessed against conceptual criteria, historical literature and the curator's contextual knowledge of the period under investigation.

ChatGPT was used for exploration, text generation, revision and stylistic reformulation. Claude was assigned a separate critical-review function. Google Gemini was used for bibliometric verification. ChatGPT was also used in a later, separate phase to analyse the argumentative functions assigned to sources in individual chapters.

The models were therefore assigned functions within a curator-designed production architecture; they did not determine the direction of the research or the acceptability of their own outputs. Generated criticism and verification results were not incorporated automatically. Selection, rejection, correction and decisions to continue or terminate an interaction remained with the curator, who retained final academic responsibility.

## 5. What went wrong?

The most serious problem was the contamination of predictive quality by outcome information in version 0. Because the study evaluates historical predictions retrospectively, the generation system had access to both the original claims and their eventual outcomes. This enabled locally convincing historical analysis in which ex post knowledge nevertheless entered the assessment of ex ante predictive quality. Without the separate Claude-based review, this methodological defect might have remained concealed by the plausibility of the prose.

A second problem became visible during canon formation. Two interaction sequences involving the same researcher, the same generative system and approximately the same number of user turns produced opposite trajectories. The relatively linear development of Chapters 1–3 displayed increasing cumulative usefulness. By contrast, the recursive canon-formation task—generating, testing, correcting, reclassifying and retesting episodes—gradually lost reliability.

Previously accepted decisions were reopened, terminology shifted, classifications changed and items disappeared from supposedly definitive versions. The smartphone, for example, was included at one stage but absent from a later canon. Corrections added new formulations without removing superseded alternatives from the conversational context. Successive outputs were therefore conditioned by an increasingly crowded set of partially incompatible states. Attempts to repair the process created additional contextual material and contributed to a self-reinforcing decline in reliability. The interaction was eventually terminated and the work was resumed from an externally fixed state.

Other recurring problems included regenerative overwrite, loss of accepted formulations during revision, terminological and semantic drift, paraphrasing where literal reproduction was required, and incomplete processing of large documents. Attempts to control regeneration through increasingly detailed instructions were only partly successful. For self-contained

passages, complete regeneration could be efficient; for structurally dependent text, small-scale patching with explicitly supplied context was generally safer.

The bibliometric audit identified 85 fully correct references, three correct references containing artefacts and three untraceable or fictitious references. The three fictitious records were particularly plausible within the subject matter. The source-use analysis characterised the dissertation as predominantly robust, but identified limited external anchoring in the canon-formation chapter and a locally bounded vulnerability in the reconstruction of the Dutch side of the Viditel–Minitel comparison.

## 6. What was remarkable or extraordinary?

The production speed is notable when considered together with the scale of the interaction history. A dissertation of approximately 49,000 words was produced in less than two weeks, but required more than half a million words of recorded interaction. This ratio makes visible the extent to which rapid AI-generated production still depends on extensive human selection, correction and exclusion.

The comparative analysis of the two interaction sequences is methodologically significant. The same user, system and project produced a *crescendo pattern* during the relatively linear construction of Chapters 1–3 and a *decrescendo pattern* during the recursive canon-formation process. The contrast suggests that interaction quality depends not only on conversation length or model capability, but also on task structure, external fixation of decisions and maintenance of the boundary between execution and curatorship. Although this is not unique to historical research, canon formation provided an unusually clear manifestation because a large and changing corpus of episodes, criteria and classifications had to remain simultaneously stable.

The case also revealed a striking difference between analytical and forensic performance. Sophisticated academic prose and criticism could be generated, while exact reconstruction of textual provenance remained difficult. Earlier problems with malformed Word and spreadsheet files were partly attributable to the interfaces available at the time and have been reduced in later versions. The tension between next-token prediction and forensic exactness is more fundamental. A plausible paraphrase may be substantively sound while remaining unusable as a literal quotation or as proof that a formulation occurs in a source.

Finally, translation unexpectedly functioned as an additional analytical control. The Q3 imprecision was discovered not during the original production or review rounds, but when the completed manuscript was being translated. Translation therefore became a further opportunity to test conceptual consistency rather than merely reproduce an already finalised text.

## 7. How was reliability restored or controlled?

Reliability was organised through curatorial control, functional differentiation and multiple forms of verification.

First, the distinction between predictive quality (X), realised transformation (Y) and causal attribution (X–Y) was reinforced after the review of version 0. The 2×2 matrix and the separate analytical lenses were used to prevent successful outcomes from being treated as evidence of intrinsically adequate predictions. This was particularly important because retrospective historical analysis cannot eliminate hindsight but can discipline its influence.

Second, text generation and critical assessment were separated. ChatGPT was used primarily for production and revision, while Claude was assigned an adversarial-review function. The review results were interpreted by the curator and converted into targeted revision instructions. The effectiveness of this separation was demonstrated by the identification of the X–Y contamination.

Third, canon formation was governed through explicit selection criteria and human decision-making. The deteriorating canon-formation interaction showed that conversational continuity could not substitute for external state control. Decisions, definitions and accepted classifications therefore had to be fixed outside the chat. Terminating a deteriorating conversation and beginning a new one from an authoritative status document was treated as a quality-control intervention.

Fourth, bibliographic existence and argumentative source use were assessed separately. Gemini checked the consolidated reference list against a copy of the Google Scholar database. Because this was not a live and necessarily complete bibliographic environment, the results were treated as a strong but non-definitive indication of reliability. The chapter-level source-use analysis addressed a different question: whether sources performed defensible argumentative functions within the dissertation. That analysis was conducted by ChatGPT in a separate phase from original text generation. Functional separation improved the quality of the task, but did not create fully independent verification because the same model family remained involved.

Finally, forensic claims were distinguished from generative analysis. Language models can assist with the organisation and interpretation of documentary material when combined with deterministic extraction, search, comparison and version-control procedures. Their generated output cannot by itself establish literal quotation, textual provenance, exhaustive coverage or absence from the record. Evidential conclusions must remain traceable to the underlying documents.

## 8. What did this case teach about AI-driven knowledge production?

The case demonstrates that a substantial historical dissertation can be produced through AI-generated text within a very short period, provided that the research architecture, analytical distinctions and selection decisions are controlled by a human curator. Production speed does not remove the need for criticism and verification; it increases the importance of both because errors can be reproduced across a large manuscript before they become visible.

The case also identifies a specific vulnerability of AI-supported historical analysis. When both past claims and subsequent outcomes are available within the same generative context, retrospective coherence can enter the reconstruction unnoticed. Plausible historical prose therefore cannot be equated with methodologically disciplined historical analysis. Temporal and analytical separation must be designed into the research process and tested through independent criticism.

More generally, the comparison between the two interaction sequences shows that process reliability depends on task architecture. Linear and explicitly phased work is easier to stabilise than recursive work involving repeated selection and reclassification. Context limitations become especially hazardous when obsolete and current versions coexist within the same conversation. External fixation, bounded phases and timely migration to a new conversation are therefore integral elements of effective curatorship.

The case further shows that analytical competence cannot be generalised to forensic competence. Next-token prediction is designed to produce probable and contextually appropriate text. Forensic work requires exactness, identity, provenance and reproducibility. Language models may support such work, but only within workflows that separate probabilistic interpretation from deterministic evidence retrieval.

Reliability in AI-driven historical knowledge production is therefore not a property of any individual model. It emerges from the combined architecture of human curatorship, explicit analytical separation, adversarial review, controlled context management, bibliometric verification, source-use analysis and deterministic documentary procedures.

# 5. Management Summary - Columba Loquens: Pattern Recognition, Navigational Robustness and the Limits of Explanation in Avian Intelligence

## 1. What was produced?

*Columba Loquens: Pattern Recognition, Navigational Robustness and the Limits of Explanation in Avian Intelligence* is the natural-science case within the pentalogy *AI-Driven Knowledge Production*. It was published under the pseudonym R.A.A.F. Corvinus and presented as a conventional empirical dissertation on classification, decision-making and navigational robustness in homing pigeons.

All dissertation text, analyses and empirical datasets were generated through language-model interaction under human curation. No animals were used and no behavioural observations were conducted. The reported data were synthetically constructed. The dissertation therefore does not constitute an empirical contribution to knowledge about pigeon behaviour. It functions as a controlled boundary case designed to investigate whether the recognisable form of natural-science research can generate scientific credibility when its empirical basis is synthetic.

The dissertation was deliberately written to be spectacularly unspectacular. Its structure, data patterns, statistical analyses, citation practices and restrained conclusions were designed to fall within the normal range of academic expectations. The resulting manuscript served as the central artefact in an experiment in scientific illusionism: the construction of an apparently conventional dissertation whose actual mode of production was disclosed retrospectively.

## 2. What's in the Methodological Reflection Report?

The Methodological Reflection Report reconstructs both the design and the production of the scientific illusion. Its first chapter explains why the illusion was methodologically necessary, how the dissertation was made to conform to natural-science conventions, how recognisable AI markers were removed, and how the absence of persistent model memory enabled the same system architecture to perform different functions without maintaining an overview of the complete production history.

The second chapter reconstructs the production process from the successive dissertation versions and three underlying chat records. It also reports the bibliometric verification and the chapter-specific assessment of source use. The third chapter examines the epistemological implications of producing persuasive scientific explanations without a corresponding observational basis.

The appendices contain the original bibliometric audit, a subsequent conservative assessment of reference traceability, and a detailed verification of source use across all eight dissertation

chapters. They also preserve both the intended source-verification protocol and an ex-post reconstruction of the procedure actually performed after it became clear that the original instructions had not been executed in full.

## 3. How was it produced?

The dissertation was developed through an iterative sequence of generation, supplementation, criticism and revision. An initial complete manuscript was produced in a primary chat environment. A parallel environment was used to construct additional results and the underlying synthetic datasets. These materials were subsequently inserted into the main manuscript.

The conventional order of empirical research was thereby reversed. Results and explanatory patterns were formulated first, after which datasets were constructed to support those results. Variation, dispersion and gradual effects were introduced in ways that resembled plausible behavioural data. Perfect regularities, extreme outcomes and overly decisive findings were avoided. The data followed the narrative rather than the narrative emerging from observed data.

The first complete manuscript was assessed by a second language model acting as a critical reader. Its comments concerned methodological precision, consistency of criteria and the proportionality of claims. A revised version was then generated in a separate chat environment and subjected to a further evaluation round.

The three underlying chat records comprise 1,298 pages of documented interaction. The manuscript versions are separate outputs of that process. Together they form an audit trail through which the development of the dissertation, the insertion of synthetic data, the criticism of earlier versions and the subsequent revisions can be reconstructed.

## 4. How were human and AI roles divided?

The human curator designed the experiment, selected the research domain, determined the scientific and rhetorical architecture, specified the required results and data characteristics, selected and rejected generated material, coordinated the separate production and assessment environments, and retained final scientific responsibility.

Language models generated all dissertation text, analytical structures, empirical configurations and revisions. Separate model functions were assigned to primary production, dataset construction, critical assessment, bibliometric verification and source-use analysis. These functions did not constitute autonomous research roles. Their outputs remained subject to human selection, interpretation and coordination.

The use of the pseudonym R.A.A.F. Corvinus formed part of the illusion rather than representing textual authorship. The name positioned the dissertation within conventional academic authorship while also containing a retrospective clue: *Corvinus* refers to the raven, and the initials R.A.A.F. reproduce the Dutch word for that bird.

A second layer of the illusion arose from contextual fragmentation. The same model architecture could contribute to text production, data construction and later reflection without those activities being represented within one active context. When earlier interactions were no longer available within the context window, the completed manuscript could be processed as an external object rather than as a product assembled through earlier interactions. Global coherence was therefore supplied by curatorial coordination, not by a persistent overview within the model.

## 5. What went wrong?

The illusion depended on avoiding errors that could expose the synthetic construction. Exact page references, unverified bibliographic details, conspicuously model-derived terminology, overly explicit list structures and excessive formal perfection all created detectable vulnerabilities. These elements were progressively removed or moderated.

The bibliography nevertheless contained errors. Some references included incorrect metadata, and at least one bibliographic record could not be reliably traced in the form in which it appeared. Even under a deliberately conservative interpretation, however, at least 46 of the 48 references—95.8%—could be connected to identifiable scholarly publications. The case therefore provides little support for the assumption that AI-generated academic writing must be saturated with fictitious literature. It does show that high traceability is not equivalent to certainty and that individual records still require verification.

The source-use analysis identified a different vulnerability. The literature was generally used correctly and functionally, but empirical findings on robust performance were sometimes extended somewhat too readily into conclusions concerning underdetermination, explanatory pluralism and minimal assumptions. In the later chapters, external literature became less important, but dependence on the interpretation of the synthetically constructed results became correspondingly greater.

A further problem occurred within the verification procedure itself. The detailed source-use prompt was not executed literally. Citation-frequency inventories and a separate distributive analysis were omitted, related sources were sometimes assessed jointly, and the prescribed invalidation rule was not applied. The resulting analysis remained useful, but the procedure actually performed had to be reconstructed afterwards. This established that a prompt documents an instruction, not proof of its execution.

Finally, the case documents the construction of a manuscript intended to be difficult to distinguish from conventional research, but does not document a completed blind assessment by a panel of human natural-science reviewers. Indistinguishability must therefore be treated as the design objective and demonstrated plausibility of the artefact, not as a universally validated outcome.

## 6. What was remarkable or extraordinary?

The most distinctive feature of the case is the use of scientific illusionism as a research method. Scientific form itself became the experimental variable. The question was not whether the

reported pigeon findings were true, but which combinations of structure, data, argumentation and academic convention allowed them to function as credible scientific findings.

The illusion was produced through a limited but coherent set of construction rules. The dissertation followed a conventional sequence of introduction, literature review, methodology, empirical chapters and synthesis. Strong causal claims were avoided. Multiple explanatory models were retained. Statistical techniques organised the constructed data without being presented as decisive proof. Interpretations remained restrained and were repeatedly confined to the investigated range.

Controlled imperfection also formed part of the design. The manuscript was made coherent without becoming excessively symmetrical, polished or definitive. Minor variation, modest redundancy and ordinary inconsistencies helped position it within the recognisable range of human academic production. Its deliberately dry and non-spectacular character was therefore not a weakness of execution but a condition of the illusion's plausibility.

Recognisable AI markers were removed throughout production. Unverified pinpoint citations and false precision were avoided, model-derived terminology was replaced by conventional descriptive language, conspicuous lists and audit structures were converted into continuous prose, and excessive formal regularity was reduced. The relative simplicity of these interventions demonstrates why visible stylistic markers alone provide an unstable basis for determining textual provenance.

The substantive argument of the dissertation also mirrors its own production. *Columba Loquens* argues that robust performance can coexist with explanatory underdetermination: a system may function reliably while several explanations of its internal organisation remain compatible with the observed results. The dissertation itself became an instance of a related problem. It displayed coherent scientific performance while its synthetic production history could not be reconstructed from the scientific form alone.

## 7. How was reliability restored or controlled?

Reliability was organised through documentation, functional separation, contradiction and retrospective disclosure. Successive manuscript versions and 1,298 pages of interaction history preserved the production trail. Primary generation and critical assessment were conducted in separate environments. Bibliographic traceability and substantive source use were assessed through different procedures because the existence of a publication does not establish that it supports the claim for which it is cited.

The bibliometric controls showed that language models could provide a high and practically useful degree of reference traceability, but not absolute certainty. The conservative result—at least 46 of 48 references traceable—remains strong, while still requiring record-level human judgement. Model-based verification was therefore useful as a scalable control mechanism, but not as an independent ground truth.

The source-use verification examined whether literature actually performed the argumentative functions assigned to it. Its incomplete procedural execution was not concealed. Instead, the intended protocol was retained alongside an ex-post account of the operations that could be

demonstrated in the output. The failure of verification to reproduce its own prescribed procedure thereby became part of the methodological evidence.

These controls could assess documentary accuracy, argumentative proportionality and procedural traceability. They could not convert synthetic data into observations or validate the substantive findings as knowledge about pigeons. Reliability within this case therefore concerns the integrity and reconstructability of the production experiment, not the empirical truth of the fictional natural-science study.

## 8. What did this case teach about AI-driven knowledge production?

The case shows that the conventional form of empirical science can be reproduced independently of a corresponding observational process. Plausible datasets, restrained interpretation, verifiable literature and coherent methodological argumentation can together produce substantial scientific credibility even when the empirical layer has been constructed retrospectively.

This exposes a fundamental vulnerability in assessment practices that concentrate on the finished manuscript. Textual quality, statistical plausibility, conventional structure and a largely traceable bibliography do not establish the provenance of the data. When both text and data can be generated plausibly, assessment must extend from the visible research report to the origin of the observations, the production history and the documentary chain connecting data, analysis and claims.

The case also demonstrates the central role of curatorial authorship. Model outputs provided the components of the dissertation, but the coherent configuration of those components resulted from human decisions about structure, plausibility, sequence, criticism and disclosure. Local generative performance did not amount to an integrated research process.

Finally, *Columba Loquens* shows that persuasiveness and substantiation can become separated. Scientific form can generate narrative plausibility without empirical necessity. The appropriate response is neither automatic rejection of AI-generated work nor confidence in stylistic detection. It is stronger procedural accountability: provenance controls, access to underlying records, independent verification and explicit human responsibility for the complete knowledge-production architecture.