Comparative Evaluation of Prompting Strategies for Structured Information Extraction : Analysis of the Accuracy-Latency Trade-off on GPT-4o-mini.

Empirical study of Zero-Shot, Chain-of-Thought, and Evidence-Based approaches applied to a corpus of heterogeneous and noisy job description documents.

The integration of Large Language Models (LLMs) into industrial document processing workflows requires a rigorous trade-off between accuracy, cost, and latency. This study evaluates the performance of six prompting strategies (Zero-Shot, Few-Shot, Chain-of-Thought, Role Prompting, Structured, Evidence-Based) on a Structured Information Extraction (SIE) task applied to a corpus of 100 heterogeneous and noisy HR job description documents. Using the gpt-4o-mini model, we demonstrate that increasing prompt complexity is not correlated with improved accuracy for this type of task. The Zero-Shot strategy achieves optimal performance (Semantic Accuracy: 93.8% | Latency: 1.36s), significantly outperforming contextual approaches such as Few-Shot (86.0% | 3.77s) and Chain-of-Thought (93.2% | 3.35s), while dividing inference costs by a factor of >2. Furthermore, we evaluate a novel approach, Evidence-Based Extraction, which, despite displaying lower recall (91.0%), offers the highest guarantee of textual faithfulness, positioning it as a preferred solution for compliance audits requiring "zero-hallucination."

AUTHOR : Jean-François de Greling 01/2026

SUMMARY

1. INTRODUCTION

1.1 Industrial and Scientific Context

The automation of administrative workflows has long faced a major technological barrier: the unstructured variability of natural language. Historically, Intelligent Document Processing (IDP) relied on deterministic approaches, combining Optical Character Recognition (OCR) and rigid syntactic rules (Regular Expressions, geometric templates). While effective on standardized forms, these systems collapse when confronted with the fluidity and unpredictability of unconstrained human communication.

The emergence of Large Language Models (LLMs), based on the Transformer architecture, has caused a fundamental disruption in this field. Unlike previous discriminative approaches, generative LLMs introduce a capability for contextual semantic understanding. They no longer seek to locate information by its position (X,Y coordinates) or by an adjacent keyword, but by the global meaning of the document. This transition from structure-based extraction to meaning-based extraction paves the way for robust automation of so-called "dirty" or "chaotic" data.

However, applying these models to the Human Resources (HR) domain presents a particularly arduous generalization challenge, characterized by three technical bottlenecks:

1. **Radical Format Heterogeneity:** Unlike invoices or purchase orders, HR documents (job descriptions, CVs, cover letters) follow no ISO standard. Relevant information may be buried in a dense narrative paragraph, scattered in a minimalist bulleted list, or drowned in a legal footer. This lack of a predictable schema renders traditional templating methods inoperable.
2. **Multilingual Complexity and "Code-Switching":** In a globalized labor market, HR corpora are intrinsically polyglot. It is common to observe, within the same document, an alternation between the local language (e.g., French) and English technical terminology (e.g., "Business Developer", "Data Scientist"). The extraction model must therefore possess sufficient linguistic plasticity to navigate between these registers without loss of precision.
3. **Inference of Implicit Data:** HR extraction often requires transforming qualitative information into structured quantitative data. For example, a mention such as "Confirmed expertise in project management" must be translated by the model into a normalized category (e.g., Senior), or an implicit duration (e.g., > 5 years), without drifting into generative hallucination.

This context poses the central problem of our study: facing this complexity, how should the model be steered? Should we guide the LLM with complex instructions and examples (Few-Shot, Chain-of-Thought) to compensate for data ambiguity, or is the intrinsic capacity of the model ("Zero-Shot") now sufficient to process this informational chaos?

1.2 Problem Definition

The integration of LLMs into industrial production pipelines is not reducible to a simple maximization of the accuracy metric. It poses a multi-objective optimization problem, often modeled as an "inference trilemma" imposing a constant trade-off between three critical variables:

1. **Semantic Precision:** The model's capacity to extract information without hallucination or omission, critical in contractual or HR contexts.
2. **End-to-End Latency:** The processing time per document. In a recruitment workflow processing thousands of applications, high latency creates unacceptable bottlenecks.
3. **Computational Cost:** Directly correlated to the number of tokens generated and consumed.

Currently, the dominant literature in Prompt Engineering tends to address the first point (Precision) at the expense of the other two. The commonly accepted heuristic hypothesis—which we term here the "Complexity Dogma"—posits that an increase in prompt sophistication is indispensable for processing noisy data.

This hypothesis relies on the empirical success of techniques such as Chain-of-Thought (Wei et al., 2022) or Few-Shot Prompting (Brown et al., 2020) in mathematical or logical reasoning tasks. By extension, the scientific community has often presumed that this positive correlation between reasoning length (intermediate tokens) and result quality applied uniformly to all NLP tasks, including Information Extraction (IE).

Our study formally challenges this generalization. We posit the existence of a complexity threshold beyond which the addition of reasoning instructions or contextual examples becomes counter-productive for shallow extraction tasks. This phenomenon, which we designate as "Contextual Overload," suggests that for recent, highly optimized models (such as GPT-4o-mini), prompt complexity not only induces a linear penalty on latency and cost but also risks introducing cognitive noise that degrades final precision.

Consequently, the central research question is no longer "How to complicate the prompt to maximize performance?", but rather: "**What is the maximum degree of simplicity we can achieve without sacrificing extraction rigor?**"

2. STATE OF THE ART

This section positions our study relative to recent advances in Prompt Engineering and evaluation methodologies for generative systems.

2.1 Prompt Engineering for Information Extraction

The advent of LLMs has shifted the performance burden from weight fine-tuning toward the optimization of input instructions (In-Context Learning). The foundational work of Brown et al. (2020) on GPT-3 established the Few-Shot Prompting paradigm, demonstrating that providing a few example pairs (input/output) within the context allows the model to infer the task and the expected output format without parameter updates. Although this method has become the industry standard for format adaptation, it presents a major drawback: the consumption of the context window and the linear increase in inference costs. Conversely, Zero-Shot Prompting (Kojima et al., 2022) relies solely on the generalization capacity of the pre-trained model, an approach that is economically more viable but historically deemed less stable for complex structures.

Concurrently, to solve tasks requiring multi-step reasoning, Wei et al. (2022) introduced Chain-of-Thought (CoT). This technique encourages the model to generate intermediate reasoning steps before producing the final response. While CoT has proven its indisputable superiority on logical, mathematical, or commonsense reasoning tasks, its efficacy on pure Information Extraction (IE) tasks remains subject to debate. Recent studies suggest that for "shallow" tasks (requiring simple identification rather than complex deduction), the computational overhead of CoT—induced by the generation of reasoning tokens—does not necessarily translate into a gain in precision, and may even introduce noise via "reasoning hallucination." Our study aims to quantify this specific limitation within the HR context.

2.2 LLM Evaluation: Beyond Exact Match

The automatic evaluation of generative systems poses a metrological challenge. Traditional metrics derived from Question-Answering (QA) research, such as Exact Match (EM) or the ROUGE score, penalize any syntactic deviation from the Ground Truth.

However, as highlighted by recent work on semantic evaluation, Exact Match is too rigid a metric for modern LLMs. These models tend to correctly rephrase information rather than servilely copying it (e.g., transforming "3 to 5 years" into "3-5 years"). The exclusive use of EM leads to a massive underestimation of the actual performance of the models, generating False Negatives that reflect formatting variations rather than comprehension errors.

To mitigate this bias, it is necessary to adopt so-called "Smart" or semantic metrics, integrating normalization steps (currency conversion, numerical entity extraction, handling of negation synonyms). It is this hybrid approach—coupling syntactic rigor with semantic validation—that we formalize in our experimental protocol.

3. METHODOLOGY

3.1 The "HR Chaos" Dataset: Constitution and Annotation Protocol

To guarantee an impartial and robust evaluation of language models, we constituted the "HR Chaos Dataset" (N=100), a hybrid corpus of job descriptions designed to maximize the entropy of input data.

3.1.1 Hybrid Composition Strategy

The corpus adopts a balanced composite structure (50/50) aimed at reconciling real-world representativeness and coverage of edge cases:

- 50% Real-World Data: Job descriptions collected via web scraping on public job platforms. These documents guarantee the ecological validity of the study, reflecting natural inconsistencies, industry jargon, and unpredictable structures of the current market.
- 50% Synthetic Data (Adversarial Synthetic Data): Job descriptions specifically generated to simulate "pathological cases" identified as problematic for classic parsers (e.g., ambiguous salary formats, fragmented layouts, dense narrative descriptions). This Data Augmentation allows testing the model's reasoning limits.

3.1.2 Linguistic Characteristics and Out-of-Distribution (OOD)

The corpus is predominantly Anglophone (>99%), corresponding to the standard vehicular language of large models. However, we introduced an "Out-of-Distribution" (OOD) sample: a job offer entirely written in French. This controlled insertion aims to test the cross-lingual robustness of the model and its ability to maintain the JSON extraction schema (whose keys are in English) even when facing an unexpected shift in source language.

3.1.3 Ground Truth Annotation Protocol

The establishment of the Ground Truth followed a strictly extractive and non-inferential annotation protocol, to minimize human interpretation bias. The applied golden rules are as follows:

- Principle of Literalness: Only information explicitly present in the text was annotated.
- "Strict Null Policy":
    - If information is not explicitly mentioned, the field is labeled null.
    - Salary: If no amount or range is quantified (e.g., vague mention "Competitive Salary"), the field is labeled null.
    - Experience: If no duration (e.g., "3 years") or quantified level (e.g., "Junior") is mentioned, the field is labeled null. We prohibited subjective inference (e.g., deducing "5 years" from the term "Senior").

This rigor, validated by an automated algorithmic audit (*script audit_dataset.py*), guarantees that the benchmark measures the reading capacity of the model and not its capacity to "guess" missing data.

3.2 Extraction Task Formalization and Target Schema

We model the task as a Structured Information Extraction (SIE) process, the objective of which is to project an unstructured document D toward a structured representation S compliant with a predefined JSON ontology.

Given the highly degraded nature of the input corpus ("Chaos Dataset"), characterized by incomplete or informal text segments, we adopted a Fully Nullable Schema. Each field is allowed to take the value *null* if the information is not explicitly present, thus prioritizing precision (avoiding hallucinations) over recall.

The output schema is defined by the following four entities:

1. role (Type: String | Nullable)

- Semantic Definition: The functional title of the position to be filled.
- Missing Data Handling: Although this field semantically constitutes the central object of the offer, our corpus includes pathological cases (e.g., informal email correspondence, text fragments) where this information is absent. In these configurations, the value *null* is imposed. Any attempt to infer a generic title (e.g., "Employee") is considered an error.
- Normalization: In the presence of a title, the model must isolate the exact denomination (e.g., "Senior Backend Engineer") while excluding adjacent advertising segments.

2. company_name (Type: String | Nullable)

- Semantic Definition: The legal entity recruiting.
- Anonymity Constraint: For "Blind Ads" published by recruitment agencies without mention of the final client, the field must strictly return *null*. Extraction of the agency name instead of the final employer is counted as a Named Entity Recognition (NER) error.

3. years_experience (Type: String | Nullable - Normalized)

- Semantic Definition: The professional experience constraint explicitly formulated in the body of the offer.
- Extended Inclusion Criteria: Unlike purely numerical approaches, our Ground Truth captures the expression of need under three distinct forms, reflecting the variability of HR language:
    - Quantified Durations: Any mention containing a temporal metric (e.g., "3 years", "10+ ans").
    - Explicit Negations: Any mention stipulating the absence of prerequisites (e.g., "No previous experience necessary").
    - Defined Qualitative Statuses: Phrases describing a specific career stage or an alternative to a degree (e.g., "First professional experience").

- Exclusion of Title Inferences: A strict distinction was made between prerequisites (body of text) and titles (position title). Seniority terms present only in the position title (e.g., "Senior Manager") without corroboration in the text were annotated as null. This rule aims to prohibit subjective inference (e.g., not arbitrarily deducing "5 years" from the single word "Senior") and to test the model's capacity to locate the "Requirements" section.

4. salary_range (Type: String | Nullable - Normalized)

- Semantic Definition: The proposed remuneration, including amount, currency, and periodicity.
- Precision-First Principle: Facing ambiguous or purely marketing mentions (e.g., "Competitive salary based on profile", "Attractive remuneration"), the model must return *null*. Only explicit numerical values are accepted to guarantee the integrity of subsequent statistical analyses.

3.3 Prompting Strategies and Experimental Conditions

To isolate the impact of prompt engineering on extraction performance, we defined a comparative protocol testing six distinct experimental conditions. Each condition represents a specific interaction paradigm documented in the literature.

All experiments were conducted on the gpt-4o-mini-2024-07-18 model with a temperature fixed at tau=0 to minimize stochastic variability.

3.3.1 Zero-Shot (Baseline: Instruction Following)

This strategy constitutes our experimental baseline. It relies exclusively on the model's generalization capabilities acquired during the pre-training phase and refined by Reinforcement Learning from Human Feedback (RLHF), as described by Ouyang et al. (2022).

- Technical Mechanism: The prompt is reduced to its simplest functional expression: a task definition (T) and the output format (S). No external data or examples are injected into the context. The model must perform a direct projection P(y|x, T) where x is the document and y is the structured output.
- Theoretical Justification: According to Kojima et al. (2022), modern LLMs are natural "Zero-Shot Reasoners." This condition aims to evaluate the model's intrinsic performance without the *Formatting Bias* that could be induced by misaligned examples.

3.3.2 Few-Shot Prompting (In-Context Learning - ICL)

Consistent with the paradigm established by Brown et al. (2020) with GPT-3, this approach exploits In-Context Learning. Here, we adopt a K=1 (One-Shot) configuration.

- Technical Mechanism: We prefix the target instruction with a demonstrative tuple (x_{demo}, y_{demo}). This pair serves as a contextual anchor. The model does not learn in the sense of weight updates (Gradient Descent) but uses the attention mechanism to identify and reproduce the syntactic schema of the provided example.

- Hypothesis: Adding an example is intended to calibrate the output space, thereby reducing JSON syntax errors and guiding the model on the expected granularity (e.g., date formatting). However, it introduces a risk of "blind mimicry" if the example differs linguistically from the target document.

3.3.3 Role Prompting (Persona Conditioning / Steerability)

This strategy relies on the work of Reynolds & McDonell (2021) regarding the "steerability" of generative models. We inject a system instruction ("System Message") defining a persona: "You are an expert HR Data Analyst specializing in resume parsing."

- Technical Mechanism: Role assignment acts as a conditioning of the latent space. By defining an expert persona, we attempt to shift the probability distribution of generated tokens toward a semantic subspace associated with professional rigor and HR vocabulary.
- Hypothesis: This conditioning should theoretically reduce naive hallucinations by simulating the behavior of a qualified human agent, fostering a more critical reading of the document.

3.3.4 Structured Prompting (Schema-Constrained)

This approach abandons the narrative aspect to focus on a strict programmatic definition of constraints, inspired by Type Definition Languages.

- Technical Mechanism: The prompt integrates a verbose and atomic description of each expected field, including data types (String, Null), value constraints (e.g., "Must be a number"), and missing value handling rules. The prompt acts as a formal pseudo-grammar injected into the natural context.
- Objective: Minimize response entropy. By making validation rules explicit *a priori*, we seek to prevent "structure drifts" (e.g., missing or misnamed JSON keys) frequently observed in free generations.

3.3.5 Chain-of-Thought (CoT: Reasoning Injection)

We implement the famous method by Wei et al. (2022), which posits that generating intermediate reasoning steps improves the resolution of complex problems.

- Technical Mechanism: The prompt is modified to include the injunction "Let's think step by step" and requires the production of a "reasoning" field before the final result. The inference process is decomposed into x -> z -> y, where z represents the chain of thought (identification of key sections, ambiguity analysis).
- Critical Hypothesis: While CoT is proven for logical reasoning, its utility for Information Extraction (IE) is controversial. We test here whether "thinking time" (reasoning tokens) improves precision or simply introduces unnecessary latency and discursive noise.

3.3.6 Evidence-Based Extraction (Contribution: Grounding Strategy)

Facing the endemic problem of hallucinations documented by Ji et al. (2023), we introduce a strategy focused on textual Faithfulness, inspired by Extractive QA tasks.

- Technical Mechanism: This strategy imposes a Strict Verbatim constraint. Unlike other methods allowing rephrasing (e.g., transforming "3 ans" into "3 years"), Evidence-Based forbids any abstraction. The model receives the instruction to act as a substring extractor: if the information is not literally present, it must return *null*.
- Objective: To evaluate a "High-Precision / Low-Recall" safety scenario. This approach aims to eliminate subjective inferences (e.g., deducing a salary from a seniority level) to guarantee that every extracted datum is factually auditable in the source document.

(The complete templates of prompts used for each condition are available in Annex A).

3.4 Experimental Infrastructure and Model Specifications

All inferences were performed via the OpenAI API, using a standardized Python execution environment to guarantee fairness in latency measurements.

3.4.1 Model Choice

Our experiments rely on the gpt-4o-mini-2024-07-18 checkpoint. This choice is motivated by the desire to evaluate the industrial viability of "distilled" and economically efficient models, as opposed to massive models (Large-Scale Models like GPT-4o) whose operational costs are often prohibitive for processing high-volume document flows.

3.4.2 Decoding Hyperparameters

To ensure strict reproducibility of results and neutralize the stochastic variability inherent in generative models, we fixed the sampling temperature at tau=0.

This configuration forces a Greedy Decoding strategy, where the model systematically selects the token with the highest log-probability. This deterministic approach is critical for factual extraction tasks, where model "creativity" is considered a hallucination risk.

3.4.3 Output Constraints (Constrained Decoding)

In order to decouple the evaluation of semantic performance (text comprehension) from syntactic performance (format compliance), we enabled the API's native "JSON Mode." This functionality acts as a strong constraint on output logits, guaranteeing that the generated string is syntactically valid (RFC 8259), thereby eliminating noise related to parsing errors (JSONDecodeError).

4. EVALUATION FRAMEWORK

Evaluating LLM performance on structured extraction tasks poses a metrological challenge: how to distinguish a factual error (hallucination) from a benign stylistic variation (rephrasing)? To address this issue, we developed a Hybrid Evaluation Engine (*evaluate.py*) that simultaneously calculates two accuracy metrics for each prediction, coupled with an operational efficiency measure.

4.1 Accuracy Metrics

To decouple syntactic rigor from semantic validity, we apply a double validation on each extracted field (y_{pred}) against the ground truth (y_{gold}).

4.1.1 Strict Match Accuracy (EM - Exact Match Relaxed)

This first metric measures the model's ability to strictly respect the source formulation or the expected format.

- Algorithmic Definition: After basic normalization (lowercase, removal of superfluous whitespace), the score is binary (0 or 1).
- Success Condition: The score is validated if strict textual inclusion is detected $y\_gold \subseteq y\_pred \; or \; y\_pred \subseteq y\_gold$, or if both values are *null* (correct handling of missing information).
- Interpretation: A high score indicates that the model acts as a faithful "copyist," respecting the principle of literalness.

4.1.2 Smart Semantic Accuracy (Semantic & Numerical)

This metric, coded specifically for this study, aims to correct False Negatives generated by Exact Match when the model correctly rephrases information. It integrates three levels of algorithmic intelligence:

- Strict Match Inheritance: If the Strict Match is validated, the Smart Match is automatically validated.
- Numerical Extraction and Comparison (Numerical Set Intersection): For quantitative fields (salary_range, years_experience), we apply an advanced regex extraction function extract_numbers().
    - *Unit Normalization:* The function handles the automatic conversion of multiplicative suffixes (e.g., detecting that "80k" is equivalent to "80000.0").
    - *Intersection Logic:* The score is validated (1.0) if the intersection between the set of numbers extracted from the prediction and that of the ground truth is not empty ($S\_pred \cap S\_gold \neq \emptyset$). This allows validating "at least 5 years" against "5 years" without penalty.
    - Semantic Handling of Negations: To mitigate the lexical variability of prerequisite absence, we defined a negation lexicon (e.g., *no, none, without, 0 year, aucun*). If y_{gold} and y_{pred} both contain a token from this lexicon, semantic equivalence is validated (e.g., "No experience" == "None").

4.2 Efficiency Metric

Beyond accuracy, the industrial viability of the system is evaluated by its latency.

End-to-End Latency: Measure of the total execution time of the inference (in seconds), including prompt pre-processing, token generation by the OpenAI API, and output JSON parsing.

This metric is averaged per strategy to quantify the "temporal cost" of prompt complexity (notably for Chain-of-Thought).

5. EXPERIMENTAL RESULTS

This section presents a quantitative and qualitative analysis of the performance of the six prompting strategies on the test corpus (HR Chaos Dataset, N=100). The evaluation relies on benchmarking the predictions generated by the gpt-4o-mini-2024-07-18 model against the Ground Truth, along three metric axes: semantic accuracy (Smart Match), syntactic fidelity (Strict Match), and inference latency.

5.1 Global Performance and Strategy Hierarchy

Table 1 synthesizes the average results obtained. Strategies are ordered by decreasing performance on the critical semantic accuracy metric (Smart Match).

Rank	**Strategy**	**Smart Match (Semantic)**	**Strict Match (Rigid)**	**ΔSem−Rig**	**Mean Latency (s)**	**Slowdown Factor**
1	Zero-Shot (Baseline)	93.8%	92.2%	+1.6 pp	1.36 s	1.0x (Ref)
2	Role Prompting	93.8%	92.5%	+1.3 pp	1.42 s	1.04x
3	Chain-of-Thought	93.2%	91.5%	+1.7 pp	3.35 s	2.46x
4	Evidence-Based	91.0%	89.8%	+1.2 pp	1.39 s	1.02x
5	Structured Prompting	87.8%	86.5%	+1.3 pp	1.59 s	1.17x
6	Few-Shot (1-Shot)	86.0%	84.2%	+1.8 pp	3.77 s	2.77x

[Table 1: Comparative Benchmark of Extraction Strategies (N=100)]

Note: $\Delta Sem - Rig$ represents the gap in percentage points (pp) between semantic and strict accuracy.

5.2 Detailed Analysis by Prompting Paradigm

Data analysis allows for the isolation of four distinct dynamics depending on the nature of the provided instruction.

5.2.1 Direct Inference (Zero-Shot and Role Prompting)

Direct inference approaches constitute the semantic performance ceiling of this study.

- Performance: The Zero-Shot strategy achieves a semantic score of 93.8%, statistically identical to that of Role Prompting.
- Impact of Persona: Assigning a role ("HR Expert") yields no significant semantic gain compared to the neutral instruction. However, it induces a slight additional latency (+0.06 s compared to the baseline), relegating Role Prompting to third position in terms of speed.
- Efficiency: The Zero-Shot method remains the fastest in the benchmark (1.36 s), confirming that the total absence of additional context favors maximum velocity generation.

5.2.2 Extractive Inference (Evidence-Based)

The Evidence-Based method distinguishes itself by its efficiency and robustness profile.

- Execution Speed: With a mean latency of 1.39 s, it ranks second, surpassing Role Prompting. This negligible gap with the baseline (+0.03 s) demonstrates that adding a verbatim constraint imposes almost no computational penalty.
- Textual Faithfulness (Grounding): This strategy displays the lowest Delta differential in the benchmark (1.2 pp). Although its global recall is lower (91.0%), this low Smart/Strict gap proves a quasi-total adherence to the source text. Errors observed are predominantly precautionary omissions (False Negatives) rather than hallucinations, making it a secure, high-velocity alternative.

5.2.3 Reasoned Inference (Chain-of-Thought)

The Chain-of-Thought (CoT) strategy positions itself in third place regarding precision (93.2%), failing to outperform the Zero-Shot baseline.

- Computational Overhead: CoT induces a major latency penalty, with a mean response time of 3.35 s, representing a slowdown factor of x 2.46.
- Failure Analysis: The significant gap between Smart and Strict (Delta = 1.7 pp) suggests that the reasoning phase encourages the model to interpret and rephrase data rather than extracting them faithfully, introducing light semantic noise without net benefit.

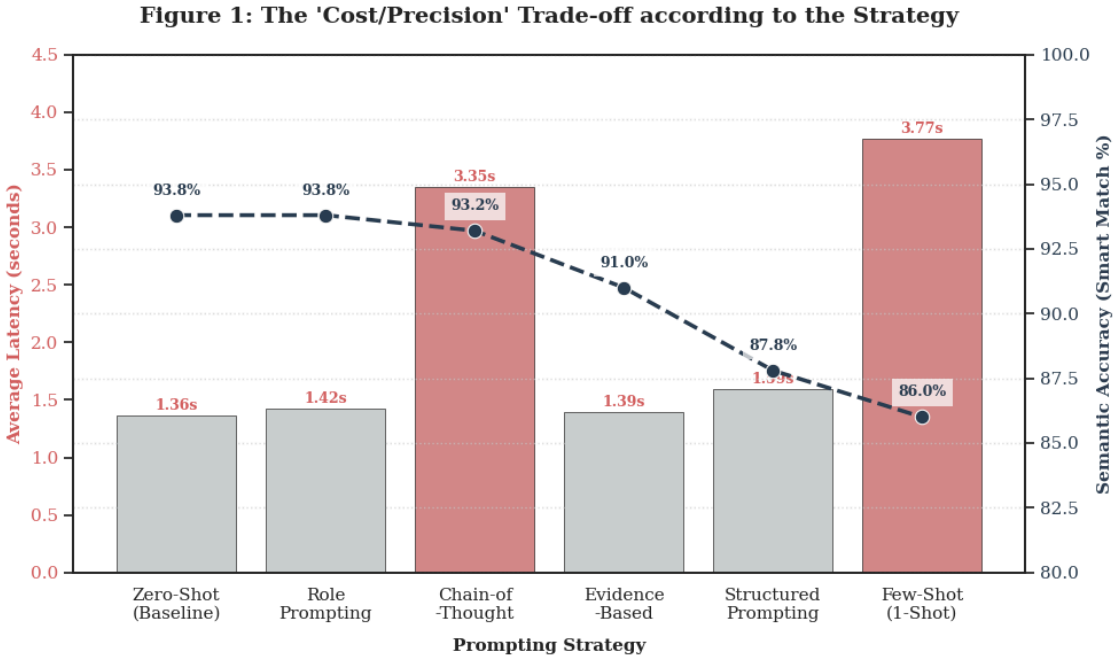
5.2.4 Constrained and Contextual Inference (Structured and Few-Shot)

Methods imposing strong structural constraints or contextual examples significantly underperform.

- Structured Prompting (87.8%): Strict JSON schema definition degrades performance (-6.0 pp) and slows inference (1.59 s) compared to direct methods, as the model struggles to reconcile schema rigidity with data entropy.
- Few-Shot Prompting (86.0%): Adding a single example (K=1) results in the lowest performance of the comparison. Besides prohibitive latency (3.77 s, i.e., x 2.77), this method suffers from contextual alignment bias, increasing the semantic error rate.

5.3 Latency and Efficiency Analysis

Figure 1 (derived from Table 1 data) illustrates the non-linear relationship between prompt complexity and processing time.



[Figure 1: The Cost/Precision Trade-off according to the Strategy]

High-Performance Cluster (Latency < 1.5s): This group is led by Zero-Shot (1.36s), immediately followed by Evidence-Based (1.39s). Both methods maximize processing throughput, making the Evidence-Based approach particularly attractive for production environments requiring both speed and traceability (verbatim).

Overload Cluster (Latency > 3.0s): The CoT and Few-Shot methods consistently exceed 3.3 seconds. For CoT, time is consumed by the generation of reasoning tokens. For Few-Shot, the latency is attributable to the processing of the input context window (Time-to-First-Token).

5.4 Synthesis of Observations

Cross-analysis of metrics establishes three empirical findings:

1. Dominance of Minimalism: The Zero-Shot strategy is the global optimum (Max Precision / Min Latency).
2. Viability of Evidence-Based: As the second fastest strategy, Evidence-Based represents the best compromise for applications requiring strict extraction (low hallucination risk) without sacrificing processing throughput.
3. Counter-Productivity of Complexity: Adding constraints (Structured), examples (Few-Shot), or reasoning (CoT) degrades the performance/cost ratio on this specific model.

6. DISCUSSION AND INTERPRETATION OF RESULTS

Beyond raw performance metrics, this section proposes a qualitative analysis of the observed cognitive and operational mechanisms. It aims to explain the behavioral discrepancies between strategies and to identify the trade-offs inherent to their industrial deployment.

6.1 The Evidence-Based Paradox: "Metric Error" or Textual Faithfulness?

In-depth analysis of error logs (via debug_errors.py) reveals that the statistical performance of the Evidence-Based strategy (91.0%) is artificially lowered by the rigidity of our quantitative evaluation schema.

- Excessive Diligence Phenomenon: The "False Negatives" recorded for this method are mostly not reading errors, but typology conflicts.
- Case Study (Job_005): The Ground Truth expected *null* for experience (absence of quantified duration). The Evidence-Based method, constrained to "extract evidence," correctly identified and extracted *"PhD in experimental Physics"*.
- Interpretation: The model succeeded in its Semantic Recall task (finding the skill information) but failed the Schema Filtering task (keeping only numbers).
- Performance Re-evaluation: If the objective were to audit the content of a job offer without strict numerical formatting constraints, this method would likely be the most effective. Unlike Zero-Shot, which takes the autonomous decision to mask qualitative information (by returning *null*), Evidence-Based surfaces raw data to the user. It acts as a High-Fidelity Scanner, ensuring that no nuance (degree, status, certification) is lost, even if this requires post-hoc normalization.

"Failures" of the Evidence-based strategy obtained via debug_errors.py displayed in the terminal:

Plaintext

```
--------------------------------------------------------------------------------------------------

JOB      | FIELD           | TRUTH (Gold)              | AI PREDICTION

--------------------------------------------------------------------------------------------------

job_003   | years_experience  | None                    | Experience as a professional w..

job_005   | years_experience  | None                    | PhD in experimental Physics, M..

job_007   | years_experience  | None                    | Final-year Master's student (B..

job_011   | years_experience  | None                    | Previous customer support expe..

job_011   | salary_range      | None                  | Attractive salary

job_014   | salary_range      | 80000,00€ à 100000,00€      | €80k and €100k OTE
```

job_016 | years_experience | None | Entry Level

job_019 | years_experience | None | Previous customer support expe..

job_023 | years_experience | None | No experience necessary

job_024 | years_experience | None | Professional

6.2 The Futility of Reasoning for Extraction (CoT Analysis)

The relative failure of the Chain-of-Thought (CoT) strategy, which does not outperform the Zero-Shot baseline despite prohibitive computational cost (+146% latency), corroborates a structural hypothesis regarding the nature of NLP tasks.

- Task Taxonomy: CoT is beneficial for Sequential Reasoning tasks (Mathematics, Symbolic Logic). However, information extraction is a Local Pattern Recognition task. The model does not need to "reflect" to copy a string of characters.
- Discursive Drift: Qualitative analysis shows that the thinking step introduces noise. By "chatting" before extracting (*"The text mentions a bonus scheme which implies..."*), the model dilutes its attention and increases the probability of hallucination via auto-suggestion.
- Operational Conclusion: In a high-volume industrial context, CoT represents unjustified "technical debt," consuming 2.5x more resources for a statistically identical result.

6.3 Limits of Contextual Approaches (Few-Shot & Structured)

Strategies relying on strong context injection (Few-Shot, Structured) demonstrated significant limits when facing the entropy of the "Chaos" dataset.

- Few-Shot Alignment Bias: With minimal performance of 86.0%, Few-Shot ($K=1$) suffers from a phenomenon of blind mimicry. The model tends to reproduce the characteristics of the provided example (language, date format, currency) rather than adapting to the target document. This lack of flexibility is critical for multilingual or heterogeneous corpora. Furthermore, its extreme latency (3.77s) due to prompt size disqualifies it for real-time use.
- Structured Prompting Rigidity: Imposing a strict JSON schema (via complex type definitions) paradoxically degraded performance (87.8%). Facing "dirty" data that does not fit into boxes (e.g., mixed "Fixed + Variable" salaries), the model constrained by structure often failed to produce valid output or truncated information, whereas Zero-Shot navigated ambiguity with greater flexibility.

6.4 The Victory of Minimalism (Zero-Shot & Role Prompting)

The dominance of direct approaches (Zero-Shot and Role Prompting at ~93.8%) validates the maturity of current models (gpt-4o-mini) regarding Instruction Following.

- Role Play Saturation: The absence of significant gain via Role Prompting indicates that the model possesses, within its pre-trained weights, a sufficient representation of the data extraction concept. Activating an "HR Expert" persona has become redundant.
- Maximum Efficiency: These methods lie on the optimal Pareto frontier: they offer the best accuracy for the lowest cost (Latency ~1.4s).

6.5 Synthesis and Architectural Recommendations

At the conclusion of this comparative analysis, we propose a decision matrix for HR extraction system engineering:

1. For Pure Performance (Mass Production): Prioritize Zero-Shot. It is the fastest, least expensive, and most robust solution to format variations.
2. For Security and Audit (Human-in-the-loop): Prioritize Evidence-Based. Although its metric score is lower (due to quantitative bias), its capacity to extract all relevant information (including qualitative) without hallucination makes it the best assistant for a human validator.
3. To Avoid: Chain-of-Thought and Few-Shot should be discarded for this type of task, as their cost-benefit ratio is strictly negative.

7. CONCLUSION AND RECOMMENDATIONS

This study aimed to isolate the optimal prompting strategy for structured information extraction from heterogeneous HR documents, such as job offers, using the gpt-4o-mini model. At the conclusion of this rigorous experimental protocol, we are able to formulate definitive conclusions for the industrialization of such systems.

7.1 Scientific Synthesis: The Triumph of "Occam's Razor"

Our results formally invalidate the heuristic hypothesis that prompt complexity positively correlates with extraction accuracy. Instead, we observe a phenomenon of Contextual Overload:

- Simplicity > Complexity: The minimalist Zero-Shot strategy dominates all metrics, offering the best semantic accuracy rate (93.8%) and the lowest latency (1.36 s).
- The Reasoning Dead-End: The Chain-of-Thought approach, while indispensable for logical tasks, proves counter-productive for extraction (Latency Delta +146%, stagnant accuracy). It introduces useless discursive noise.
- The Failure of Mimicry: The Few-Shot approach on high-entropy data ("Chaos Data") induces severe formative alignment biases, degrading overall performance (-7.8 pp vs Baseline).

In summary, for modern Instruction-Tuned models, intrinsic generalization capacity is sufficient to handle complex extraction tasks without artificial guidance.

7.2 Industrial Implications

From an integration perspective or for developing proprietary RAG solutions, these discoveries dictate a revision of standard AI pipeline architectures.

7.2.1 Architecture Recommendation: The "Lean Zero-Shot" Pipeline

We recommend abandoning complex structure prompts (multi-shots, verbose schemas) in favor of direct and streamlined instructions for information extraction.

- Target Configuration: gpt-4o-mini Model + Zero-Shot Prompt + Temperature 0.
- Benefit: This configuration guarantees maximum system stability and facilitates prompt maintenance (no examples to update).

7.2.2 Resource and Cost Management (Token Economy)

Abandoning Chain-of-Thought and Few-Shot translates into an immediate and massive economic gain.

- Inference Cost Reduction (OPEX): By eliminating reasoning tokens (CoT) and input examples (Few-Shot), we reduce the volume of tokens consumed per document by approximately 60%.

- Latency Gain (UX): Moving from a mean processing time of ~3.5s (CoT/Few-Shot) to ~1.4s (Zero-Shot) allows for envisaging real-time applications without degrading user experience.

7.2.3 The Role of Evidence-Based: The Safety Audit

Although we recommend Zero-Shot for mass production, the Evidence-Based strategy must be retained as a specialized audit tool.

Use Case: For sensitive missions requiring "zero-tolerance" for hallucination (e.g., GDPR Compliance Audit, Contract Clause Verification), this module can be activated to guarantee that every extracted datum is strictly present in the source document, acting as a Safety Net.

8. BIBLIOGRAPHY

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). *Language Models are Few-Shot Learners.* Advances in Neural Information Processing Systems (NeurIPS), 33, 1877-1901. (Fundamental reference for the Few-Shot strategy)

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., ... & Fung, P. (2023). *Survey of Hallucination in Natural Language Generation.* ACM Computing Surveys, 55(12), 1-38. (Used to justify the necessity of the "Evidence-Based" approach and the management of hallucination risks)

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). *Large Language Models are Zero-Shot Reasoners.* Advances in Neural Information Processing Systems (NeurIPS), 35, 22199-22213. (Reference for the Zero-Shot baseline and emergent capabilities without examples)

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., ... & Lowe, R. (2022). *Training language models to follow instructions with human feedback.* Advances in Neural Information Processing Systems (NeurIPS), 35, 27730-27744. (Key reference explaining why the model follows Zero-Shot instructions so effectively thanks to RLHF/Instruct Tuning)

Reynolds, L., & McDonell, K. (2021). *Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm.* In Proceedings of the CHI Conference on Human Factors in Computing Systems (Extended Abstracts). (Theoretical reference for the Role Prompting / Persona strategy)

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E., Le, Q., & Zhou, D. (2022). *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models.* Advances in Neural Information Processing Systems (NeurIPS), 35, 24824-24837. (Fundamental reference for the Chain-of-Thought strategy and the critique of its usage outside reasoning tasks)

OpenAI. (2024). *GPT-4o-mini System Card.* OpenAI Technical Reports. Retrieved from https://openai.com (Technical specification of the model used for inference)

9. APPENDICES

This section provides the technical artifacts necessary for the full reproduction of the experiments presented in this study.

Annex A: Prompt Templates (Experimental Conditions)

Below are the exact instructions (System Prompts & User Prompts) injected into the gpt-4o-mini model for each condition. Dynamic variables are denoted within curly braces {}.

A.1 Zero-Shot (Baseline)

Extract information from the job description below into JSON format.

Target Schema: {{JSON_SCHEMA}}

{{FORMATTING_RULES}}

Job Description:

"""

{{INPUT_TEXT}}

"""

A.2 Few-Shot (One-Shot Learning)

Extract information from the job description into JSON.

Example Input:

"""

We are hiring a Python Dev at Google. Salary is $120k. No exp needed.

"""

Example Output:

{"role": "Python Dev", "company_name": "Google", "years_experience": "No exp needed", "salary_range": "$120k"}

Now do the same for this text:

Job Description:

"""

{{INPUT_TEXT}}

"""

A.3 Role Prompting (Persona)

As an expert, analyze the text below and extract these fields into JSON format.

Target Schema: {{JSON_SCHEMA}}

{{FORMATTING_RULES}}

Job Description:

"""

{{INPUT_TEXT}}

"""

A.4 Structured Prompting (JSON Definition)

Task: Information Extraction

Source Type: Job Description

Output Type: JSON

Definitions:

- role: The job title.

- company_name: The hiring entity.

- years_experience: Minimum experience requirement (substring).

- salary_range: Compensation details.

Input Text:

{{INPUT_TEXT}}

Response (JSON):

A.5 Chain-of-Thought (CoT)

Analyze the job description step-by-step before extracting data.

1. First, identify where the Role is mentioned.

2. Second, look for the Company Name.

3. Third, find specific mentions of Experience.

4. Fourth, look for Salary digits.

Output a JSON with two keys:

"reasoning": "your step-by-step analysis here",

"data": {{JSON_SCHEMA}}

{{FORMATTING_RULES}}

Job Description:

"""

{{INPUT_TEXT}}

"""

A.6 Evidence-Based (Verbatim Constraint)

Task: Data Extraction

constraint: Strict Verbatim Copying

Extract specific substrings from the text to populate the JSON below.

Field Definitions:

- role: The exact job title found in the header or text.

- company_name: The exact organization name.

- years_experience: The exact phrase specifying time (e.g. "3+ years") or level (e.g. "No experience").

- salary_range: The exact figures or range found (e.g. "£30k - £40k").

Critical Rules:

1. COPY-PASTE: Values must be exact substrings from the input text.

2. NULL: If the information is not explicitly stated, use null.

3. PREFERENCE: If multiple values exist, choose the most specific one (e.g. prefer "No experience necessary" over "Entry Level").

Input Text:

"""

{{INPUT_TEXT}}

"""

Output (JSON):

Annex B: Characterization of the "Chaos" Dataset

The test corpus consists of documents exhibiting high structural entropy. Below are anonymized examples representative of the difficulties encountered.

job_012.txt :

```
Henner, an international and independent group, creates, manages,
and markets innovative solutions in personal insurance. We are
1,850 employees present in 20 countries worldwide. Every day, we
support 64,000 client companies and 2.3 million beneficiaries.



Our expertise in social protection is summed up in one word:
Caring. Caring is much more than an idea. It's a promise that
Henner makes to its clients and partners to carry out its social
protection consulting, brokerage, and management services with a
constant focus on providing exceptional care and attention to the
quality of its support.









THE POSITION



Our Commercial Department is recruiting its future "Sales Manager
(F/M)" on a permanent contract, to support our growth objectives.
The position is based in Neuilly sur Seine (92) and reports to the
Department Head.



You will be responsible for monitoring, controlling, and
collecting remuneration flows from insurers in France and
internationally.



Your main activities include :



Leading and developing a portfolio of strategic key accounts in
personal insurance (life and health)
```

Structuring and negotiating complex insurance and reinsurance arrangements, taking into account clients' specific needs

Managing and optimizing profit-sharing accounts, with a high level of technical expertise

Managing, supporting, and developing the skills of a dedicated team

Building strong and lasting relationships with insurers, reinsurers, and key client partners

Leveraging your professional network to generate new business opportunities and foster growth.

WHY JOIN US ?

By joining our team, you become part of a committed collective : kindness is the foundation of all our relationships.

You can count on a strong culture of mutual support and knowledge sharing to help you develop your potential. You will be trained on our tools as well as the basics of insurance. Whether or not you come from the insurance sector, at Henner we will offer you a career path tailored to you, with close and ongoing support.

You will benefit from all the social advantages granted to our teams : vacation bonus, profit-sharing bonus, company savings and retirement plan with attractive employer contribution, 75% reimbursement of transportation costs, fully covered individual or family health insurance by Henner, childcare allowance, medical teleconsultations, and access to our company restaurant.

You are also guaranteed respect for your work-life balance thanks to our company agreements and a flexible work organization that allows up to 2 days of remote work per week, or even 3 days for people with disabilities, caregivers, or pregnant employees.

HENNERFR

```
THE PROFILE



You have 10 to 15 years of experience in personal insurance, with
a solid background in life and health insurance.



Your expertise is built on a proven mastery of complex insurance
and reinsurance structures, as well as profit-sharing account
management.



You have demonstrated strong managerial abilities through leading
both commercial and technical teams, combining high standards,
support, and strategic vision.



Your active and influential professional network within the
insurance and reinsurance ecosystem is a real asset to generate
new business opportunities and support the Group's growth.



You are fluent in english, both written and spoken, and
comfortable working in an international environment.



Your strategic mindset goes hand in hand with strong operational
skills and excellent interpersonal abilities, enhancing your
credibility with key partners and clients.



In short, you are a recognized leader (rigorous, inspiring, and
committed), someone others can rely on, who thrives in a
collaborative and positive work environment.
```

job_056.txt :

```
Role: Foster Coordinator (Animal Welfare)

Organization: Save The Cats Foundation

Location: Manchester, UK
```

```
Do you love animals? Do you want to make a tangible difference in
the lives of abandoned kittens?

Save The Cats is a non-profit organization dedicated to rescuing
stray cats and finding them forever homes. We are growing fast and
need a dedicated person to manage our foster network.



The Role:

As a Foster Coordinator, you will be the main point of contact for
our 50+ foster families. You will coordinate vet visits, ensure
fosters have enough food and litter, and update our adoption
database. You need to be organized, empathetic, and good on the
phone.



Requirements:

We are looking for someone with a big heart. No specific
professional experience is required, though organizational skills
are key. You must be able to commit to roughly 10 hours a week,
mostly on weekends or evenings.



Compensation:

Please note that this is a volunteer position (Unpaid). However,
it is an incredible opportunity to gain experience in the
non-profit sector and event management. All travel expenses and
phone bills related to the role will be reimbursed.
```

job_093.txt :

```
--- Job Description ---

SilverLake Capital is a premier private equity firm managing over
$10B in assets. We are seeking a highly organized and discreet
Executive Assistant to support our Chief Financial Officer. This
is a high-pressure environment requiring 24/7 availability during
deal closings.
```

```
--- Responsibilities ---

* Manage an extremely active calendar of appointments.

* Arranging complex and detailed travel plans, itineraries, and
agendas.

* Prepare expense reports and reconcile credit card statements.

* Draft confidential correspondence and board meeting materials.

* Gatekeep and manage incoming calls/emails.



--- Qualifications ---

* Bachelor's degree required.

* 10+ years of experience supporting C-Level executives in Finance
or Law.

* Expert proficiency in Microsoft Office (Outlook, Excel,
PowerPoint).

* Impeccable written and verbal communication skills.

* Ability to work overtime as needed.



--- Compensation ---

* Base Salary: $100,000 - $130,000 per annum.

* Discretionary Year-End Bonus (typically 20-30% of base).

* 100% paid health insurance premiums.

* Catered lunch daily.
```

Annex C: Hybrid Validation Algorithm

To guarantee fair evaluation beyond simple string matching, we formalized the numerical validation logic used in our script evaluate.py.

```
ALGORITHM Smart_Scoring_Validation

INPUTS:

   gold_text  : Ground Truth (String)

   pred_text  : AI Prediction (String)

   field_name : Field type (e.g., "salary_range", "years_experience")

OUTPUT:

   Score (1.0 for Success, 0.0 for Failure)

CONSTANTS:

   NEGATION_KEYWORDS = ["no ", "none", "not ", "0 year", "without", "aucun"]


// --- MAIN FUNCTION ---

FUNCTION Calculate_Smart_Score(gold_text, pred_text, field_name):

   // 1. Normalization (Lowercase, removal of extra whitespace)

   t_gold <- NORMALIZE(gold_text)

   t_pred <- NORMALIZE(pred_text)

   // 2. Strict Verification (Textual Inclusion)

   // If one is included in the other, it is immediately valid.

   IF (t_gold IS IN t_pred) OR (t_pred IS IN t_gold) THEN:

      RETURN 1.0

   // 3. Numerical Validation (Only for Salary and Experience)

   // We extract numbers and handle units (80k = 80000).

   IF field_name IS "salary_range" OR "years_experience" THEN:
```

```
        nums_gold <- EXTRACT_NUMBERS(t_gold) // Returns a set {80000}
        nums_pred <- EXTRACT_NUMBERS(t_pred) // Returns a set {80000, 100000}
        // If the intersection of sets is not empty (at least one common number)
        IF INTERSECTION(nums_gold, nums_pred) IS NOT EMPTY THEN:
            RETURN 1.0
    // 4. Semantic Validation of Negations
    // We check if both texts express an absence (e.g., "None" vs "No experience")
    gold_is_negative <- CONTAINS_KEYWORD(t_gold, NEGATION_KEYWORDS)
    pred_is_negative <- CONTAINS_KEYWORD(t_pred, NEGATION_KEYWORDS)
    IF (gold_is_negative IS TRUE) AND (pred_is_negative IS TRUE) THEN:
        RETURN 1.0
    // 5. Default Failure
    RETURN 0.0
END FUNCTION
// --- UTILITY FUNCTIONS ---
FUNCTION EXTRACT_NUMBERS(text):
    // Detects numbers and converts 'k' suffixes (thousands)
    // Ex: "80k" becomes the number 80000.0
    Find all patterns "Digit" + "Optional Suffix (k)"
    For each match:
        IF suffix == "k" THEN value = value * 1000
        Add value to result set
    RETURN set_of_numbers
```