ROMENTER AI

MODULE 01 // EXECUTIVE BRIEFING

The Evolution of NLP

Prepared By :

Ben Jerome

Founder of Romenter.ai

www.romenter.ai

For decades, the construction industry has relied on structured data—rows and columns in Excel, P6 schedules, and financial ledgers—to manage risk. However, the vast majority of project intelligence lies in unstructured data: daily logs, emails, variation orders, and contracts. Natural Language Processing (NLP) is the branch of Artificial Intelligence dedicated to unlocking this text-based data.

This briefing outlines the history of NLP, tracing its journey from rigid, fragile systems to the adaptive, "intelligent" models we see today (like GPT). To make these concepts tangible, we utilize analogies from the built environment. We move from the digital equivalent of a bricklayer following a single, rigid drawing to a Master Architect who understands the design intent of the entire development.

Crucially, we conclude with why this evolution is a game-changer for Forensic Auditing. We have moved from simple keyword searches (finding the smoking gun only if you know exactly what the gun looks like) to semantic understanding (detecting the intent to defraud before the gun is even fired).

Phase I: The Foundation – Rules-Based Systems (1950s – 1980s) The Analogy: The Rigid Bricklayer.

In the early days of NLP, computer scientists approached language the way a strictly supervised bricklayer approaches a wall. The bricklayer is given a very specific set of instructions: "If you see a corner, place a quoin. If you see a gap, fill it with mortar. " They do not look at the building as a whole; they only execute "If X, then Y" commands.

How It Worked: These were known as Symbolic NLP or Rules-Based Systems. Linguists and computer scientists manually hand-coded thousands of grammatical rules. They tried to map out the English language into a giant decision tree.

If a sentence has the word "bank, " check the neighboring words.

If the neighbor is "river, " classify "bank" as geography.

If the neighbor is "money, " classify "bank" as finance.

ROMENTER AI

DATA SOVEREIGNTY: Public LLMs (like ChatGPT) learn from your uploads. Romenter isolates your data. When you upload drawings to the Insight Engine, they are vectorized locally within your secure partition. Your intellectual property never trains the public model.

The Limitation: The "Variation Order" Problem Imagine managing a project where every single potential variation had to be written into the contract before breaking ground. If a site condition changed that wasn't explicitly defined in the rulebook, work would stop or, worse, proceed incorrectly.

Rules-based systems were incredibly brittle. They couldn't handle the nuance of human communication. In construction, a "light beam" might refer to a structural steel member that isn't heavy, or it might refer to a laser level. A rules-based system struggles to differentiate without an explicit rule for every possible context. As language (like a construction project) is infinite in its variability, you simply cannot write enough rules to cover every scenario.

Relevance to Auditing: In the past, auditing software relied on this logic. You could search for specific "bad" words like "bribe" or "error. " But if a contractor wrote, "Let's discuss the facilitation fee for the permit, " a rules-based system would miss it entirely because "facilitation fee" wasn't in its hard-coded dictionary of bad words.

Phase II: The Statistical Turn (1990s – 2010s) The Analogy: The Experienced Estimator.

By the 1990s, computing power increased, and we moved from manually writing rules to Statistical NLP. This is comparable to an experienced Estimator who stops looking at the physics of every bolt and instead relies on historical data.

The Estimator might say, "I don't need to calculate the exact labor hours for this drywall. I know from the last 50 projects that 1,000 sq ft of drywall usually costs $X. " They are using probability based on past occurrences.

How It Worked: Instead of teaching the computer grammar rules, researchers fed it massive amounts of text and let it calculate probabilities. The machine learned that "concrete" is followed by "pour" 80% of the time, and by "eat" 0.001% of the time. This was the era of "Bag of Words" models—the computer threw all the words into a bag to count them, often ignoring the order.

The Limitation: Lack of "Structural Integrity" While better than rules, this approach lacked structural understanding. The Estimator knows the price of the wall but doesn't necessarily know why the wall is there. In statistical NLP, the sentence " The contractor sued the client" and "The client sued the contractor" look mathematically very similar—they contain the exact same words. A "Bag of Words" model struggles to distinguish who is the plaintiff and who is the defendant.

ROMENTER AI

DATA SOVEREIGNTY: Public LLMs (like ChatGPT) learn from your uploads. Romenter isolates your data. When you upload drawings to the Insight Engine, they are vectorized locally within your secure partition. Your intellectual property never trains the public model.

Relevance to Auditing: Statistical auditing tools could flag anomalies based on frequency. They might notice that a specific vendor appears in invoices 500% more often than the average. This is useful, but it generates high false positives. It detects volume, not intent. It sees the "what, " but misses the "why."

Phase III: Neural Networks & RNNs (2010s – 2017) The Analogy: The Critical Path Scheduler (Gantt Chart).

Around 2010, Deep Learning re-energized the field. We developed Recurrent Neural Networks (RNNs).

Think of an RNN like a Critical Path Method (CPM) schedule. In a CPM schedule, Task B cannot happen until Task A is complete. You process the project linearly. You pour the foundation, then you frame, then you roof. The state of the roof depends entirely on the sequence of events that came before it.

How It Worked: RNNs read text sequentially, one word at a time, from left to right. When the model read the fifth word in a sentence, it retained a "memory" of the previous four. This allowed for better context. It understood that "The bank of the river" was different from "The bank of England" because it remembered the word "river" or "England" appearing nearby.

The Limitation: The "Long-Lead Item" Memory Loss The fatal flaw of RNNs was "vanishing gradients, " which is technical jargon for short-term memory loss.

Imagine a project spanning five years. By the time you reach Year 5 (the end of a long paragraph), the scheduler has forgotten the specific design constraints agreed upon in Year 1 (the start of the paragraph).

If a variation order describes a complex dispute starting with "In reference to the initial agreement regarding the steelwork... " and then writes 200 words of legal jargon before concluding " ...the contractor is liable, " an RNN might forget the subject was "steelwork" by the time it reaches "liable. " It processes the sequence but loses the connection between distant but related events.

ROMENTER AI

DATA SOVEREIGNTY: Public LLMs (like ChatGPT) learn from your uploads. Romenter isolates your data. When you upload drawings to the Insight Engine, they are vectorized locally within your secure partition. Your intellectual property never trains the public model.

Phase IV: The Transformer Revolution (2017 – Present) The Analogy: Building Information Modeling (BIM) & The Master Architect.

In 2017, Google researchers published a paper titled "Attention Is All You Need, " introducing the Transformer architecture (the "T" in GPT).

This is the shift from 2D CAD drawings to BIM (Building Information Modeling). In a BIM model, every element is aware of every other element simultaneously. If you move an HVAC duct in the North Wing, the model instantly calculates clashes with the structural beams, even if they were designed by different teams. The system doesn't read the building from left-to-right; it looks at the whole building at once.

How It Worked: The Mechanism of "Attention" Transformers abandoned the linear, word-byword reading of RNNs. Instead, they ingest the entire sentence (or paragraph) simultaneously. They use a mechanism called Self-Attention.

Imagine the sentence: "The concrete cracked because it was cured too quickly. " When the model processes the word "it, " an RNN has to guess what "it" refers to based on the last few words. A Transformer, however, assigns "attention scores. " It looks at every other word in the sentence to see which has the strongest relationship to "it."

Does "it" refer to "cracked"? (Weak link)

Does "it" refer to "quickly"? (Weak link)

Does "it" refer to "concrete"? (Strong link)

The model understands the relationships and dependencies between words, regardless of how far apart they are. Just as a Master Architect knows that the load on the foundation is directly related to the change in roofing material 50 stories up, the Transformer understands that the word "fraud" in paragraph 4 is linked to the "shell company" mentioned in paragraph 1.

Generative Pre-trained Transformers (GPT): These models are "pre-trained" on essentially the entire internet. They have read every building code, every legal contract template, and every project management handbook available online. They aren't just matching words; they have built a conceptual map of how human language functions.

ROMENTER AI

DATA SOVEREIGNTY: Public LLMs (like ChatGPT) learn from your uploads. Romenter isolates your data. When you upload drawings to the Insight Engine, they are vectorized locally within your secure partition. Your intellectual property never trains the public model.

The "So What?": Implications for Forensic Auditing Why does a Construction Project Director need to care about the shift from Rules-Based to Transformers? Because it fundamentally changes how we protect capital.

1. From "Keyword Search" to "Concept Detection" The Old Way (Rules/Statistical): To find corruption, an auditor would search for keywords: "gift, " "bribe, " "kickback. "

The Transformer Way: You can ask the AI, "Find me all communications where the tone suggests an unethical pressure to approve a payment. " The model can identify an email that says, "Bob, it would be a shame if your golf club membership renewal got lost in the mail next week. Let's get that invoice signed. " There is no "bribe" keyword here. A rules-based system sees a conversation about golf. A Transformer sees coercion.

2. Resolving the "Claim Avalanche" The Problem: When a project goes to litigation, contractors often dump millions of pages of unstructured PDF correspondence to overwhelm the client's legal team.

The Transformer Solution: Transformers can digest these millions of documents, summarising the narrative thread of a specific delay event. They can map the "DNA" of the claim—linking the initial Request for Information (RFI) to the site meeting minutes, to the WhatsApp message between site engineers, and finally to the formal Variation Order. It constructs a timeline of intent, not just a list of files.

3. Anomaly Detection in Specifications The Application: Transformers can compare the "Design Intent" (the Architect's narrative) against the "Technical Specs" (the engineering data). The Audit: If the narrative says "High-grade, weather-resistant cladding for marine environments, " but the spec sheet lists a cheaper, non-marine grade aluminum, a Transformer can flag this semantic contradiction. A keyword search would miss it because both "cladding" and "aluminum" are valid construction terms. The contradiction is only visible if you understand the meaning of "marine environment. "

ROMENTER AI

DATA SOVEREIGNTY: Public LLMs (like ChatGPT) learn from your uploads. Romenter isolates your data. When you upload drawings to the Insight Engine, they are vectorized locally within your secure partition. Your intellectual property never trains the public model.

Conclusion:

The transition of Natural Language Processing has mirrored the maturation of the construction industry itself.

We started with Rules-Based systems (The Bricklayer): Effective for simple, repetitive tasks but fragile when conditions changed.

We moved to Statistical/RNNs (The Estimator/Scheduler): capable of prediction and sequence, but prone to losing the "big picture."

We have arrived at Transformers (The Architect/BIM): Systems that understand context, relationships, and intent across vast datasets.

For Project leaders, this is not just IT trivia. It is a shift in risk management. We now have the capability to audit the meaning of our project data, not just the metrics. In an industry where a single misunderstood clause or an overlooked correspondence can cost millions, having an AI that understands the "Design Intent" of our language is the ultimate risk mitigation tool.

ROMENTER AI

DATA SOVEREIGNTY: Public LLMs (like ChatGPT) learn from your uploads. Romenter isolates your data. When you upload drawings to the Insight Engine, they are vectorized locally within your secure partition. Your intellectual property never trains the public model.

ROMENTER AI

THEORY INTO PRACTICE
LAUNCH INSIGHT ENGINE