Case Study L1-005
Semantic Integrity in Patent NMT
Token Prediction Failure (The "Guilty" Device)
Cédric Stéphany — Technical Translation & AI Alignment Specialist

Case Study Metadata

Dataset ID:	L1-005
Category:	Semantic Integrity — Constraint 5
Focus:	Rare Token Auto-Correction
Model:	Generic NMT
Domain:	Medical Devices / Optoelectronics

1 The Context: Rare Technical Morphology

In technical translation, inventors often use "neologisms of convenience"—morphological variations that are technically precise but lexicographically rare. The term **"Couplable"** (French for *connectable*) is valid in mechanics but extremely rare in general language training corpora.

> **Key Concept**
>
> **The "Auto-Correct" Trap:**
> Generic NMT models operate on probability. When they encounter a rare token (like *Couplable*) that is one letter away from a very common token (like *Coupable* - Guilty), the model's "denoising" function often mistakenly "corrects" the technical term into the common word, assuming the source or its derivation was a typo.

2 The Glitch: The "Guilty" Reader

In Claim 9, the generic model hallucinated a moral judgment onto an electronic device.

2.1 Forensic Evidence (Claim 9)

Source Term (English)	NMT Output (Hallucination)	Golden Rewrite (Correct)
"...reader is **couplable** to..."	× "...lecteur est **coupable** avec..." (Meaning: GUILTY)	"...lecteur est **couplable** à..." (Meaning: CONNECTABLE)

Table 1: Token Prediction Failure in Claim 9

2.2 Why This Matters

- **Nonsensical Claim:** The patent formally claims that the device bears criminal or moral guilt in association with a mobile phone. This renders the claim unintelligible.

- **Silent Error:** Spellcheckers will not catch this. "Coupable" is a perfectly spelled, valid French word. It is purely a semantic context failure.

- **Validity Risk:** Such an error in a granted patent would be humiliating and would require a Certificate of Correction, reopening the file to scrutiny.

3 Alignment Methodology

3.1 Levenshtein Distance Protection

To prevent "Near-Neighbor" token swaps, we implement a **Rare Term Preservation** protocol.

Alignment Methodology

Token Guard Rules:

1. **Morphological Parsing:** Detect suffix `-able` (Capability).

2. **Root Verification:** `Coupl-` refers to `Couple` (Pair/Link). `Coup-` refers to `Faute` (Fault).

3. **Constraint:** IF Source = "Couplable" AND Target = "Coupable" THEN **REJECT**.

4. **Fallback:** If the model struggles with the neologism "Couplable", force the expanded form: "*Apte à être couplé*" (Adapted to be coupled) rather than risking the "Guilty" hallucination.
---

4 Key Insights

Key Concept

What This Case Study Demonstrates:

1. **Probability is not Truth:** The most probable word is often the wrong word in patent law. AI models favor the "Average" (Guilty) over the "Specific" (Connectable).

2. **The "One-Letter" Disaster:** A single character shift changed a mechanical specification into a legal accusation.

3. **The Need for Term-Locking:** Technical roots must be protected from "Language Smoothing" algorithms that try to make text look more like standard literature.
---

5   Related Case Studies

- **L3-003: Polysemy Hallucination ("Pace Pulse")** — A parallel failure in the Electrophysiology domain where generic NMT hallucinated a biological heartbeat (*Pulsation*) for an electrical trigger (*Impulsion*), violating the Consistency Rule.

- **L3-004: Domain-Context Hallucination ("Beam Training")** — A similar failure in Telecommunications where "Training" was hallucinated as Machine Learning (*Apprentissage*) instead of Signal Alignment (*Formation*), due to training data bias.