Case Study C4-002
Mathematical Integrity in Patent NMT
Contextual Pollution (The "Triple Prime" Contagion)

Cédric Stéphany — Technical Translation & AI Alignment Specialist

Case Study Metadata

Dataset ID:	C4-002
Category:	Mathematical Integrity — Constraint 4
Focus:	Token Contagion in Formulas
Model:	Generic NMT
Domain:	Semiconductor Physics / Power Diodes

1 The Context: Invariant Variables

In semiconductor patent claims, mathematical formulas define the physical boundaries of the invention. Two rules are absolute:

1. **Variable Invariance:** A variable named $Lp2$ in English must be $Lp2$ in French. It cannot become $lp2$ (case sensitivity matters).
2. **Operator Integrity:** Mathematical operators $(\cdot, \times, +)$ must be preserved exactly.

Key Concept

The "Contextual Pollution" Effect:
Generic NMT models are sensitive to local context. If a specific symbol (like the triple prime $'''$) appears frequently in the surrounding text (e.g., reference numerals $10'''$, $45'''$), the model may "hallucinate" this symbol into unrelated strings, such as mathematical formulas, corrupting the data.

2 The Glitch: The "Triple Prime" Infection

In Claim 9, the generic model allowed the reference numeral style to infect the mathematical formula, rendering the claim mathematically undefined.

2.1 Forensic Evidence (Claim 9)

2.2 Why This Matters

- **Undefined Range:** The expression $0,3'''$ is mathematically meaningless. It looks like "Zero comma three triple-prime." The multiplication operator $(\cdot)$ has been deleted.
- **Variable Drift:** The variable $Lp2$ was lowercased to $lp2$. In physics, L usually denotes Inductance or a specific Length, while l might denote mean free path. Changing the case changes the variable.

Source Formula (English)	NMT Output (Hallucination)	Golden Rewrite (Correct)
"...range from **0.3·Lp2**..."	× "...plage de **0,3$'''$ lp2**..." (Operator replaced by Symbol)	"...plage de **0,3·Lp2**..." (Math Preserved)

Table 1: Symbol Contagion in Mathematical Range

- **Invalidation Risk:** A patent claim with a nonsensical mathematical range ("0.3 triple prime") cannot be construed by a court, leading to invalidity for indefiniteness.

3 Alignment Methodology

3.1 The "Math-Freeze" Protocol

To prevent text-based tokens from polluting math zones, we treat formulas as **Translatable Objects**.

Alignment Methodology

Regex Protection Layers:

1. **Formula Detection:** Identify patterns containing mathematical operators ($=, <, >, \cdot, \times$) and alphanumeric variables.
2. **Variable Locking:** Enforce case-sensitivity. `Lp2` must match `Lp2`. `IF source.case != target.case THEN REJECT`.
3. **Symbol Quarantine:** Explicitly forbid the insertion of text-reference symbols ($'$, $''$, $'''$) inside detected float values or mathematical ranges.
4. **Placeholder Strategy:** Replace the formula with a placeholder token `[[MATH_TAG_01]]` during translation, then restore the exact string post-processing to ensure zero alteration.

4 Key Insights

Key Concept

What This Case Study Demonstrates:

1. **AI is Suggestible:** The model saw $'''$ fifty times in the text, so it guessed that 0.3 should also have $'''$. It mimicked the "style" of the paragraph at the expense of the math.
2. **Math is Not Language:** You cannot "translate" a formula. You must "transfer" it. Any attempt by an LLM to "interpret" the math usually leads to formatting errors.
3. **Case Sensitivity is Structural:** In English, we can be sloppy with capitalization. In code and physics variables, we cannot.

Portfolio: Patent Translation AI Alignment Framework
Author: Cédric Stéphany
Specialization: Technical Translation (FR↔EN) — Patents, Telecommunications, Semiconductors
Contact: cedric@tmcwx.com
Last Updated: January 19, 2026