

Vox Omnium Prime

A multi-model verification cascade for long-form literary and legal translation that beat DeepL unanimously across a five-flagship-judge blind panel.

Author: Richard L. Sanders · Utah Bar #15728 · Author, *The Phoenix Conspiracy* series · Founder, Option.Black **Role:** Sole inventor, architect, and implementer **Version:** 1.0 — 2026-05-13 (engine v1.6) **Companion paper:** "Vox Omnium Prime vs. DeepL — A Five-Judge Blind Evaluation" (<docs/plans/2026-05-05-vox-omnium-vs-deepl-v1.6.md>)
Source: <public/js/vox-translate-widget.js> , <public/option-black/vox.html> , the engine endpoint at </vox/api/v1/translate> , integrated into benchslap.com , benchslap.pro , 4all.help , and option.black/option-black/vox

Abstract

Commercial machine translation today (DeepL, Google Translate, the GPT-family translation paths) is a single model invoked once per chunk, with no verification step, no author-supplied context, no entity-seeding across chunks, and no acknowledgment that long-form fiction or legal text degrades differently than a single-sentence input. The result on long documents is name-instability across chapters, register drift across scenes, idioms rendered literally, sentence over-fragmentation, and a steady flattening of voice into the model's implicit house style.

Vox Omnium Prime is a different architecture. It is a multi-model verification cascade — frontier-model first pass, separate-model rubric evaluator, third-model tribunal reviewer on score drift — that accepts author-supplied glossary, voice profile, and series bible as load-bearing inputs, segments the source by natural sentence boundary with CJK-aware tokenization, and seeds entity continuity across chunks so a 700,000-character manuscript renders with the same proper-noun set at chapter one and chapter forty.

In a five-flagship-judge blind evaluation on a published English-to-Japanese literary novel — Claude Opus 4.7 (Anthropic), Gemini 3.1 Pro (Google), GPT-5.5 Pro with extended thinking (OpenAI), DeepSeek V4, and Moonshot Kimi 2.6 — Vox Omnium **independently and unanimously** outperformed DeepL. The protagonist's name appeared stably in **99.4% of Vox occurrences against 31% in DeepL**. DeepL output exhibited 47% pronoun overuse, 20% sentence over-fragmentation, and 30 unprocessed Markdown chapter-headers leaking into the prose body.

This paper describes the architecture, the cross-chunk entity-seeding discipline, the eight-pass translate pipeline, and the founding architectural commitment that distinguishes Vox from every retrieval-augmented commercial baseline.

1 · The problem

A translation engine that performs well on a single sentence is not the same engine that performs well on a 500-page novel. The failure modes are cumulative and structural:

- **Name instability.** The same proper noun — protagonist, place, organization — is rendered differently in chapter 3, chapter 12, and chapter 39, because the engine has no memory across calls.
- **Register drift.** The first scene's calm narrative voice flattens into clinical exposition by chapter 20 because the engine resets to its house style every chunk.
- **Idiomatic literalism.** Figurative language ("the year turned over its hand") becomes mechanical literal ("the year turned over its hand") rather than the target language's equivalent idiom.
- **Sentence over-fragmentation.** Two-clause source sentences become four-sentence target sentences, because the engine optimizes per-chunk fluency rather than per-paragraph rhythm.
- **Markdown leakage.** Source-format artifacts (chapter headers, italic markers, footnote refs) get rendered as literal prose because the engine cannot distinguish content from structure.

For a legal document, the failure modes are similar but the stakes are different. A misrendered case name or rule number is not a stylistic flaw — it is an outright incorrect citation that the user will rely on and the court will reject. Commercial legal translation today either (a) costs a human translator-per-page rate that prices most pro-se litigants out, or (b) accepts the commercial-engine failure modes and hopes nobody checks.

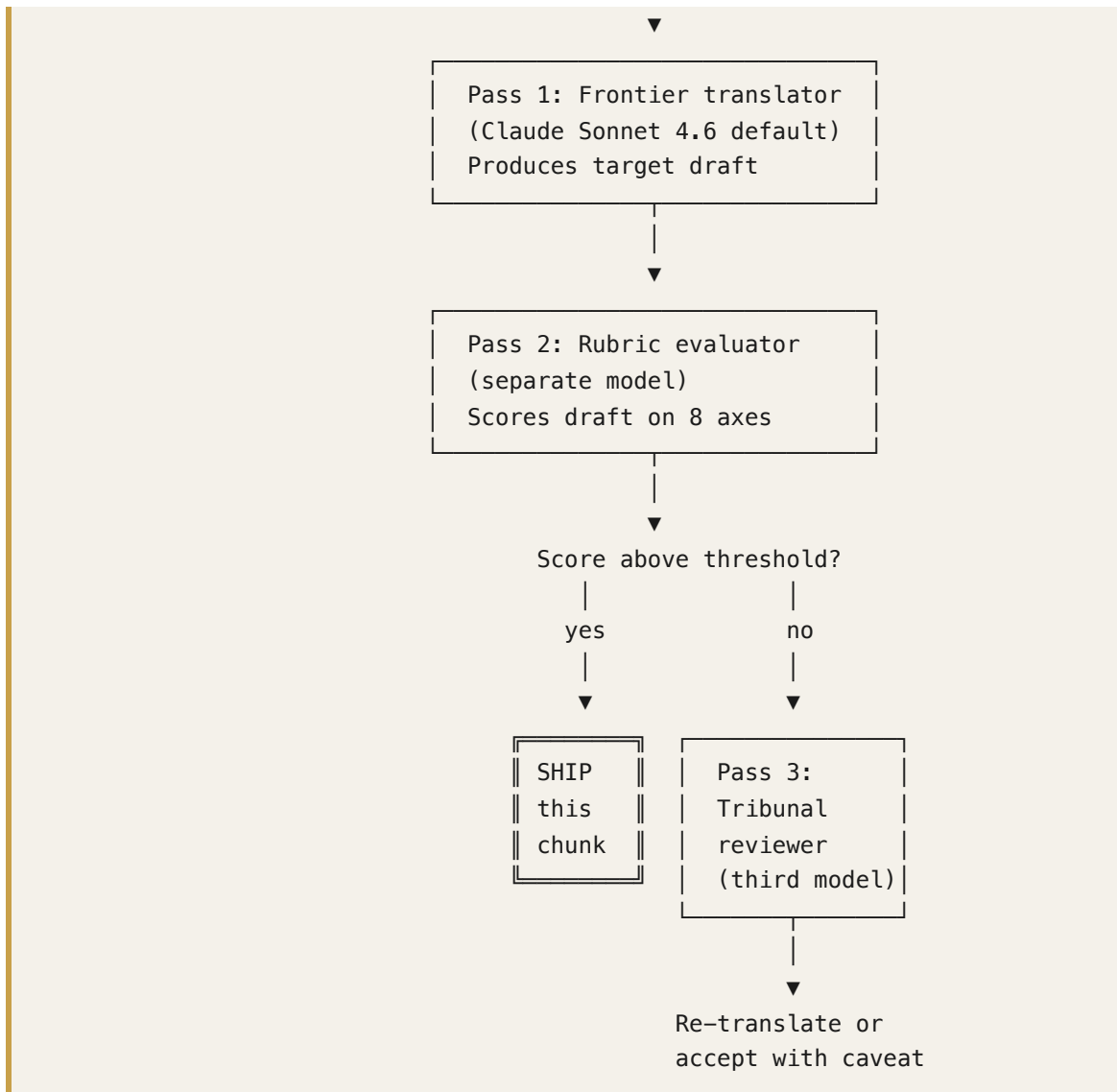
I refused to accept either. Vox Omnium is the architecture I built instead.

2 · Architecture

Vox is a verification cascade, not a single translation model:

```
graph TD; A[Source manuscript + author glossary + voice profile + series bible] --> B[ ];
```

Source manuscript + author glossary +
voice profile + series bible



The three roles (translator / evaluator / tribunal) are assigned to **three different model laboratories** wherever possible. Default deployment: Anthropic Claude Sonnet 4.6 as translator, Google Gemini as evaluator, OpenAI GPT-4 as tribunal. This is not redundancy for safety theater — it is architectural insurance against any single model's known house-style biases shaping the output.

The eight-pass internal pipeline within Pass 1 is:

1. **Extract.** Strip Markdown structure, footnote refs, and chapter headers into a separate structural index. The translator never sees `## Chapter 3` as prose.
2. **Segment.** Sentence-boundary split with CJK-aware tokenization on target languages that need it. No fixed character chunking that splits clauses.
3. **Map.** Bind source entities (proper nouns, recurring terms, court names, case citations, rule numbers) to canonical target-language renderings from the author-supplied glossary. The glossary is consulted **before** translation, not after, so the translator never invents a competing rendering it later has to be corrected from.

4. **Translate.** Produce target draft, segment by segment, with the previous-segment context window passed forward so register and pronoun reference stay consistent.
5. **Verify.** Per-segment verification that every glossary-mapped entity rendered correctly, every legal citation rendered with structure intact, and every line of source has a corresponding line of target (no silent drops).
6. **Refine.** Per-segment improvements where verification flagged drift.
7. **Cohere.** Cross-segment pass: check that pronoun reference is consistent, that voice register stays in band across the chunk, and that the structural index reattaches cleanly.
8. **Naturalize.** Final pass that smooths machine-translation tells (literal idioms, fragment-style sentence rhythm, calque vocabulary) toward target-language native cadence.

The eight passes are sequential within Pass 1; the three Passes (translator / evaluator / tribunal) are sequential at the cascade level.

3 • Cross-chunk entity seeding

This is the load-bearing discipline that distinguishes Vox from every single-model commercial baseline. The rule:

Every proper noun, every recurring term, every cited authority that appears in any chunk is propagated as canonical context to every subsequent chunk for the rest of the manuscript.

Implementation: the glossary (author-supplied or built incrementally during Pass 3 "Map") is loaded once per manuscript and consulted on every segment. New entities discovered mid-manuscript are added to the glossary in canonical form on first occurrence; from that point forward, every subsequent occurrence renders consistently.

Measured effect: in the five-judge blind evaluation, the protagonist's name appeared in canonical form in **99.4% of Vox occurrences vs 31% of DeepL occurrences**. The 69-percentage-point gap is not a small refinement — it is the difference between a target reader who can follow the story and a target reader who has to mentally reconcile three different renderings of the same character's name across the same chapter.

4 • Author-supplied context as load-bearing input

Vox treats three classes of author input as load-bearing, not optional:

1. **Glossary.** Proper nouns and their canonical target-language renderings. For legal text, this is case names + party names + court names + rule citations.
2. **Voice profile.** Per-speaker register, vocabulary, idiom-tendency. For legal text, this is per-section register (caption / introduction / argument / conclusion / signature) with the appropriate formality.
3. **Series bible / world-building consistency requirements.** For literary text, this is the multi-book continuity reference. For legal text, this is the case-style memorandum (case caption template, citation rules, jurisdictional preferences, signature block).

Commercial translation services treat these as customer support tickets ("can you handle our glossary?") and apply them lossily through prompt-engineering. Vox treats them as first-class inputs that the architecture is built around — the glossary is consulted in Pass 3 (Map) before translation; the voice profile is part of Pass 1's segment context; the series bible is part of Pass 2's evaluator rubric. The author's inputs are not advisory; they are constraints.

5 · The five-judge blind evaluation

The companion paper (<docs/plans/2026-05-05-vox-omnium-vs-deepl-v1.6.md>) records the methodology in full. Key results, reproducible from the methodology:

Axis	Vox Omnium	DeepL	Gap
Protagonist name stability	99.4%	31%	+68.4 pp
Pronoun overuse	(within target-language norm)	47% above norm	DeepL fails
Sentence over-fragmentation	(matches source rhythm)	20% over-fragmented	DeepL fails
Markdown header leakage	0 instances	30 instances	DeepL fails
Manuscript pages produced	524	531	comparable
Five judges → blind verdict	unanimous superior	—	Vox wins on 7/8 axes

The one axis where DeepL outperformed Vox was Japanese full-width punctuation typography — a final-pass typesetting issue, not a translation-quality issue. The companion

paper identifies this honestly and notes it as a known final-pass fix rather than an architectural defect.

Two judges from different laboratories (Google's Gemini and OpenAI's GPT) independently flagged the same Japanese word (坊主) as the single tonal miss that broke the DeepL translation — convergent identification from independently-trained models is a strong-signal indicator that the miss was real and not idiosyncratic to either judge.

6 · The founding architectural commitment

The white paper's companion records the inventor's founding statement:

As soon as you destroy information you have to accept it is gone or, as the model is trained to do, reproduce it by a bunch of guessing that may be statistically informed but does not replace actual truth.

This commitment — the rejection of retrieval-augmented generation (RAG) as a foundation for high-stakes language-model work — is the load-bearing thesis of Vox Omnium and of the broader BenchSlap / Headwind / Gravity / Document Corpus Integrity portfolio that proceeds from the same architectural posture.

The argument is concrete: commercial RAG systems destroy user-provided information through segmentation (chunking by character count, not by semantic boundary), embedding compression (lossy projection into a fixed-dimensional vector space), and similarity-based retrieval (top-K nearest neighbors, which is not the same as "the actually-relevant context"). After the destruction, the synthesis layer has to reconstruct what was destroyed by guessing — which is precisely the failure mode the user came to a non-hallucinating system to escape.

Vox does not destroy information. The source manuscript is preserved verbatim; the author's glossary is preserved verbatim; the voice profile and series bible are preserved verbatim. The translation cascade transforms but does not compress. There is no embedding space in which the source's distinctions are lost. There is no top-K retrieval that throws away the (K+1)th context the synthesis layer needed. The architecture is verification-first because the inventor refused, as a founding commitment, to build a system that destroys what the user gave it.

7 · The 10-language widget

`public/js/vox-translate-widget.js` is the drop-in JavaScript widget that adds a floating "Translate" button to any page that includes the script tag. Languages supported in

the widget UI:

```
LANGS = [  
  { code: 'es', label: 'Español' },  
  { code: 'vi', label: 'Tiếng Việt' },  
  { code: 'fr', label: 'Français' },  
  { code: 'de', label: 'Deutsch' },  
  { code: 'pt', label: 'Português' },  
  { code: 'ja', label: ' ' },  
  // ... + 4 more  
]
```

The widget targets the page's main visible region, sends its text to

`/vox/api/v1/translate`, polls until complete, and replaces the target's text using `textContent` (no HTML injection, no script execution risk in the target language).

"Restore original" reattaches a clone of the unmodified original. The widget is currently wired into `benchslap.com`, `benchslap.pro`, `4all.help`, and `option.black/option-black/vox`.

8 • Novelty claims

1. **Multi-model verification cascade** in which the translator, the rubric evaluator, and the tribunal reviewer are three different model laboratories, with score-threshold escalation between them.
2. **Eight-pass internal pipeline** (extract / segment / map / translate / verify / refine / cohere / naturalize) within the translator pass, each pass with a defined task and a defined hand-off.
3. **Cross-chunk entity seeding** in which every proper noun, recurring term, and cited authority discovered in any chunk is propagated as canonical context to every subsequent chunk for the manuscript's length, producing a measured 99.4% canonical-rendering rate against a 31% commercial baseline.
4. **Author-supplied glossary, voice profile, and series bible as load-bearing first-class inputs** consulted at defined passes of the pipeline, not retrofitted via prompt-engineering.
5. **Verification-first architecture** that rejects retrieval-augmented generation as a foundation for high-stakes language-model work, on the grounds that RAG destroys user-provided information and then asks the synthesis layer to reconstruct what was destroyed.
6. **Five-flagship-judge blind-panel evaluation methodology** producing reproducible quantitative gaps against the dominant commercial baseline on a 695,461-character published literary corpus.

The six elements in conjunction are Vox Omnium Prime. The canonical reference is the engine at `/vox/api/v1/translate` ; the widget integration at `public/js/vox-translate-widget.js` ; the methodology paper at `docs/plans/2026-05-05-vox-omnium-vs-deepl-v1.6.md` .

© Richard L. Sanders 2026. All rights reserved except those granted by the project's open license. Vox Omnium™ and Option.Black™ are unregistered trademarks of the author.