Product Requirements Document:

1. Problem Statement

International students reading long English-language texts (academic, news, general web content) default to **whole-text machine translation** as a coping mechanism. This produces two compounding failures:

1. **Comprehension loss**: automated translation flattens nuance, connotation, and register — students miss what the author actually means, not just individual words.
2. **Skill stagnation**: translation-as-crutch removes the friction required for language acquisition. Students never build independent reading fluency because they never have to.

Core: Reduce translation to only the words that are actually unknown → forcing engagement with the original English text.

---

2. Target User

Attribute	Description
Primary persona	International student in a US academic setting, non-native English speaker
Language pair (v1)	Mandarin (L1) → English (L2)
Proficiency range	CEFR B1–C1 (assessment determines exact placement)
Context of use	Reading academic texts, news articles, general web content in English
Current behavior	Copy-pastes full paragraphs into Google Translate / DeepL

---

3. Goals & Success Metrics

Goal	Metric
Reduce reliance on whole-text translation	% of flagged words revealed vs. full-page translate-tool usage (self-reported or measurable if scoped later)
Build vocabulary retention	Words saved to dictionary → words later recognized without click (would require re-test)

Increase reading confidence/speed on English text	User-reported CEFR level progression across repeat assessments
Demonstrate product viability	Complete working demo on live web pages

Note: retention and progression metrics require a review/re-assessment loop (so users can re-test if they have left the application for a bit or feel like their recommendations/text translations aren't accurate to their skill level

---

4. Scope

In Scope (for the MVP)

- CEFR-based self-assessment (user-set for now, quiz-based later)
- Live content-script scanning of any webpage
- Word-level difficulty detection against a CEFR-tiered frequency list
- Inline gloss replacement (English → Mandarin) for words above user's level
- Click-to-reveal: original English word + definition + nuance note
- Personal dictionary (saved words persist across sessions)
- Single language pair: Mandarin ↔ English

Out of Scope

- Multi-language support (architecture should not preclude it, but no other language ships in v1)
- Flashcard review scheduling (spaced repetition)
- Pop-culture article recommendation engine
- Formal CEFR placement test (currently manual dropdown selection)
- Cross-device sync of dictionary (currently local to browser profile)
- Mobile/non-Chrome browser support

---

5. Functional Requirements

5.1 Assessment (Not yet built)

ID	Requirement
F1.1	User can select or take a test to determine CEFR level (A1–C2)
F1.2	Level is retestable at any time, retained via chrome.storage.sync
F1.3	System stores level history to support future progression tracking

5.2 Content Scanning & Overlay

ID	Requirement	Status
F2.1	Scan visible text nodes on any webpage on load	❌ Not built
F2.2	Flag words whose difficulty tier exceeds user's CEFR level	❌ Not built
F2.3	Replace flagged word inline with native-language (Mandarin) gloss	❌ Not built
F2.4	Click reveals original English word + definition + nuance note	❌ Not built
F2.5	Re-scan page content dynamically loaded after initial load (SPA/infinite scroll)	❌ Not built
F2.6	Match inflected word forms (plurals, verb tenses) to base dictionary entries	❌ Not built

5.3 Personal Dictionary

ID	Requirement	Status
F3.1	Save a revealed word to personal dictionary via one click	❌ Not built
F3.2	View saved words in popup UI	❌ Not built
F3.3	Prevent duplicate entries for the same word	❌ Not built
F3.4	Export dictionary or sync across devices	❌ Not built

5.4 Flashcards (Not yet built)

ID	Requirement
F4.1	Generate flashcards from personal dictionary entries
F4.2	Apply spaced-repetition scheduling (SM-2 algorithm) to review order
F4.3	Track review count and mastery per word

5.5 Pop-Culture Context Finder

ID	Requirement
F5.1	Given a saved word, query a news/content API for recent articles containing it

F5.2 Surface 2–3 article suggestions per word in the dictionary view

F5.3 Prioritize sources relevant to US cultural context (per original goal of cultural familiarity, not just linguistic exposure)

---

6. Technical Requirements

6.1 Standards & Data Sources

Component	Standard/Source	Rationale
User proficiency labeling	CEFR (A1–C2)	Familiar framework to international students; simpler UX than a numeric score
Word difficulty scoring	wordfreq (Python, MIT license) mapped to CEFR tiers	Free, no licensing cost (vs. COCA, which requires paid license for full corpus)
Translation/definition/nuance	CC-CEDICT (open, offline)	Structured lexical entries with sense distinctions, unlike sentence-level MT (Google Translate) which flattens nuance

---

8. Milestones (Proposed)

Phase	Deliverable
Phase 0	Working content-script demo: scan, flag, gloss, reveal, save. Single language pair, curated dataset
Phase 1	Lemmatization + full wordfreq dataset integration (removes the two biggest fidelity gaps)
Phase 2	CEFR placement assessment (replaces manual dropdown)
Phase 3	Flashcard generation with SM-2 spaced repetition
Phase 4	Pop-culture article finder (NewsAPI integration)
Phase 5	Multi-language expansion beyond Mandarin

Comment: I know this doesn't include the website element yet, but I wanted to get your POV on the extension first, then we can iterate