Concerns regarding translation solution:

# AI reliability (hallucinations/confabulations)

* #1 clarification: Our solution is **not** a generalized conversational model.
  + Our model institutes a number of checks and validations to maintain contextual, syntactical, legal, and ethical integrity
    - CROSS-VALIDATION: There are two models in the CasaLingua architecture, each performing their own tasks as well as the ability to ‘sanity-check’ the outputs of the other.
      * In conjunction with this are comprehensive auditing elements, Human-in-the-loop observability and correction, and adherence to ethical and bias concerns through the use of proven guardrails and published guidelines.
    - LEGAL: The model is trained on a corpus of language defined by and specific to each jurisdiction being served
      * Comprehensive human readable auditing of every input/output to the pipeline.
      * The corpus of language can easily be adapted to fit the inevitable changes in the legal landscapes that occur.
    - HALLUCINATIONS/CONFABULATIONS
      * GPT-type models seek to confidently answer questions by offering up the most likely words or phrases applicable to a question.
      * Our model does not implement a pipeline allowing for ‘best-guess’ or ‘likely’ answers to questions. The model performs accessible translation services within a fixed-scope subject and does not scrape unvetted sources of data.
      * If an interaction fails in some way, the reason for this is clearly audited and can be fielded and corrected by “Human-in-the-loop”.
      * Corrections to failures are reincorporated into the corpus of data the models base their knowledge upon

# Hardware/Platform Requirements

* - **In-Use API Computational Cost**
  + Significantly lower than training phase (See below)
  + Current reasonable estimates indicate API computational and space requirements to the existing Bloom Housing infrastructure would be approximately 10x current, non-AI module, operations.
  + We will provide accurate usage metrics when we have more information on currently used Bloom Housing platform/infrastructure.
* **EXAMPLE TRAINING PHASE REQUIREMENTS**
  + Training phase is a one time process for a defined corpus of data and will be done on developer hardware rather than deployed platform. Below are some example requirements for the **TRAINING** phase of each model listed

| *Model* (Parameters, if applicable) | SIZE | Pros | Cons | Deployment Notes |
| --- | --- | --- | --- | --- |
| *LLaMA2* (13B) | 24GB | High performance, open weights, best for legal text w/ fine-tuning. | Needs a strong GPU (or Jetson + quantization) | Can run on local server with GGML/GGUF |
| *Mistral* (7B) | 16GB | Lightweight, fast, good at summarization and simplification | Slightly less coherent than GPT-4 for complex text passages | Ideal balance for Jetson / edge device usage |
| *Mixtral* (8x7B [MoE]) | 48GB | Top-tier reasoning with modular routing | Large memory footprint required | Best for server setups only |
| *Phi-2* / *TinyLlama* | < 8GB | Can run on Jetson Nano appliance with quantization | Needs extensive prompt engineering and rules | Suitable for low-resource devices |

| **Model** | **Parameters** | **GPU\_RAM\_Needed** | **GPU\_Hours\_1\_Epoch** | **GPU\_Hours\_2\_Epochs** | **Estimated\_Cost\_USD** |
| --- | --- | --- | --- | --- | --- |
| TinyLlama (1.1B) | 1.1B | 12–16 GB | 6–10 | 12–20 | $24–40 |
| Mistral-7B | 7B | 24–48 GB | 35–60 | 70–120 | $140–240 |
| Mixtral (12.9B) | 12.9B | 48–80 GB | 70–100 | 140–200 | $280–400 |

**Deployment Utilities:**  
Use GGUF or llama.cpp or Ollama to containerize and deploy the selected model offline.

* - Reliability/Affordability over sub-contracted translators/legal aides