

## Step 1: Data Collection & Extraction

### Raw CHAT Corpus Files

CHILDES Database • Multi-language • Ages 0-9

### 1. Format Conversion & Age Classification

- Parse CHAT files → Extract participants & metadata
- Convert to JSON format • Classify by child age (<3, ≥3)
- Total processed: 5,000+ raw dialogues

### 2. Fragment Extraction & Standardization

- Question-centered extraction (adult questions as anchors)
- 15-line context windows • CHI/ADU role standardization
- Complete interaction turns preservation

### 3. Length Filtering & Validation

- Minimum 15 lines required • Q&A cycle completeness check
- Remove truncated fragments • Ensure sufficient content

**Filter Rate: 45% removed**

### 4. Deduplication & Diversity

- Similarity threshold: 85% • Hash-based exact matching
- Retain higher quality versions • Ensure topic diversity

**Deduplication Rate: 15% removed**

### 5. Semantic Quality Screening

- Information entropy calculation • Unique vocabulary count
- Semantic richness assessment • Interaction complexity

**Final Output: High-quality excerpt collection**

## Step 2: Data Annotation

### Annotation Schema

- Strategy type: 15+ categories
- Schema goal: Assimilation, Transformation, Restructuring
- Cognitive alignment level classification:
  - Full alignment • Partial alignment
  - Disalignment • Unknown

**Powered by: Gemini-2.0-flash API**

```
{
  "id": "Dialogue unique identifier",
  "metadata": {
    "domain": "Dialogue domain",
    "topic": "Dialogue topic",
    "child_age": "Child age"
  },
  "dialogue": [
    {
      "turn": "Turn number",
      "speaker": "CHI or ADU",
      "utterance": "Dialogue content",
      "annotation": {
        // Only required for ADU utterances
        "strategy_tags": ["Strategy subtype-type"],
        "schema_goal": "A or T or R",
        // Only required for CHI utterances
        "alignment_level": "Full/Partial/Disalign/Unknown"
      }
    },
    // More turns...
  ],
}
```

## Step 3: Synthesis & Generation

### Dual-Path Generation Strategy

- Path 1 (75%):** Real data augmentation  
**Path 2 (25%):** Theory-guided generation

### Structured Prompt Engineering

1. Role & task definition (LLM as educator)
2. Core constraints (word limits: Adults≤100, Childs≤50)
3. Theory injection (A/T/R concepts, strategies)
4. Alignment requirements (10-20% disalign)
5. Output format (strict JSON schema)
6. Quality checklist (coherence, educational value)
7. Final validation (all requirements met)

### Data Generation Methods

- Authentic fragment extension (preserve naturalness)
- From-scratch generation (theory as seed)
- Extract & modify excerpts (maintain coherence)
- A-T-R cognitive guidance chain formation

### Post-processing & Optimization

- Semantic similarity check (prevent redundancy)
- Label normalization & standardization
- Active strategy identification & supplementation
- Iterative refinement based on quality metrics

## Final Dataset Output

### ExploraTutor Dataset Statistics

#### Core Metrics:

- Total Dialogues: 2,045 high-quality samples
- Q&A Pairs: 17,682 interaction turns
- Age Coverage: 0-9 years (full range)
- Avg Dialogue Length: 8.6 turns

#### Strategy Distribution:

- 15+ distinct strategy types identified
- Average 3.2 strategies per dialogue
- Balanced distribution across categories

#### Schema Goal Coverage:

- Assimilation (A): 35%
- Transformation (T): 40%
- Restructuring (R): 25%

#### Alignment Distribution:

- Full Alignment: 43%
- Partial Alignment: 32%
- Disalignment: 15%
- Unknown: 10%

#### Quality Assurance:

- Overall Quality Score: 90±5 points
- Expert Agreement:  $\kappa = 0.82$  (substantial)
- Safety Compliance: 100%

## Step 4: Three-Layer Quality Control System

### Layer 1: Automated Filtering

**Pass Rate: 55% | Processing: 3,700+ samples**

#### 5 Evaluation Dimensions:

1. Lexical Adaptation (vocabulary appropriateness)
2. Strategy Diversity (≥3 types per dialogue)
3. Schema Goal Coverage (balanced A/T/R)
4. Alignment Distribution (10-20% disalignment)
5. Dialogue Coherence (response relevance)

**Safety Screening:** Sensitive word filtering

**Scoring:** Multi-dimensional weighted assessment

**Threshold: Total score ≥80, No dimension <60%**

### Layer 2: Expert Evaluation

**Pass Rate: 86% | Cohen's K = 0.82**

#### Double-blind Review System:

- 2 Child Psychology/Education PhD experts
- Independent 5-point Likert scale rating
- High inter-rater reliability ( $\kappa=0.82$ )

#### 5 Assessment Dimensions:

1. Age Appropriateness
2. Dialogue Naturalness
3. Teaching Effectiveness
4. Schema Development
5. Conflict Resolution Support

**Criteria: Score ≥3, Expert difference ≤2**

### Layer 3: Final Quality Assurance

**Selection Rate: 10% | Quality Score: 90±5**

#### Value Alignment Check:

- Positive, healthy, upward content
- No bias or inappropriate guidance

#### Unqualified Data Processing:

- Expert feedback → Iterative optimization
- Return to synthesis path for refinement
- Direct rejection if unimprovable

#### Quality Monitoring:

- Continuous assessment • Early warning system

## Model Fine-tuning

- Qwen2.5-7B-Instruct
- Deepseek-llm-7b-Base
- ChatGLM-6B-Base

LLaMA Factory • LoRA method • SFT approach

## Evaluation Metrics

#### Automatic Evaluation:

- Language Adaptation • Topic Relevance
- Strategy Diversity • Schema Coverage

#### Human Evaluation:

- Schema Development Support • Dialogue Naturalness
- Cognitive Conflict Resolution • Strategy Quality
- Information Accuracy • Socio-Emotional Support