**4 Phase Interest model:**

Many students progress through school and early adulthood without developing meaningful interest in any academic or career domain. When a learner does not enter any phase of the interest‑development continuum Curiosity Activation, Engagement Sustainment, Personal Relevance Formation, or Passion‑Driven Mastery, they remain in a pre‑interest state where no subject feels engaging, relevant, or motivating. Over time, this lack of identified interest can lead to persistent academic disengagement, difficulty choosing educational pathways, low intrinsic motivation, and limited skill development. As students transition into higher education and adulthood, the absence of clear interests contributes to career uncertainty, reduced resilience, lower life satisfaction, and a tendency to drift rather than make intentional choices. This problem highlights the need for early detection of interest‑development gaps and structured support systems that help learners explore, discover, and cultivate meaningful interests across all stages of life.

**Why use a ML model?**

Using a machine‑learning model gives your interest‑detection system the ability to learn from real student behavior rather than relying only on fixed rules. As more responses are collected, the model begins to recognize subtle patterns such as mixed interests, emerging preferences, or phase‑specific engagement signals that a simple scoring system would miss. This makes predictions more accurate, more personalized, and more adaptive over time. ML also allows the test to evolve with different age groups, cultures, and learning styles, ultimately creating a smarter, more responsive system that grows better the more it is used.

**4 Phase of Interest (4PF):**

1. Curiosity Activation - Short bursts of attention sparked by novelty, surprise, or external cues.
2. Engagement Sustainment - Continued involvement supported by meaningful tasks, scaffolding, or relevance.
3. Personal Relevance Formation - Students begin to connect content to personal goals, values, and identity.
4. Passion-Driven Mastery - Deep, enduring interest with self-regulation, persistence, and proactive exploration.

**Age Groups:**

1. Primary (6–12 years):
   1. Curiosity Activation: Bright visuals, games, surprises, hands-on experiments.
   2. Engagement Sustainment: Structured tasks with immediate feedback, short challenges, group activities.
   3. Personal Relevance Formation: Linking topics to everyday life (pets, sports, family), encouraging storytelling.
   4. Passion-Driven Mastery: Early signs of “favorite subjects,” self-initiated projects (drawing, building, coding basics).
2. Secondary (13–18 years):
   1. Curiosity Activation: Real-world problems, debates, multimedia, social relevance.
   2. Engagement Sustainment: Longer projects, peer collaboration, gamified progress tracking.
   3. Personal Relevance Formation: Career exploration, elective choices, connecting learning to identity and values.
   4. Passion-Driven Mastery: Deep dives into chosen domains (science fairs, coding clubs, arts portfolios), self-directed study.
3. Higher Education (19–24 years):
   1. Curiosity Activation: Exposure to cutting-edge research, internships, interdisciplinary courses.
   2. Engagement Sustainment: Seminars, labs, collaborative research, mentorship.
   3. Personal Relevance Formation: Aligning studies with career goals, social impact, personal philosophy.
   4. Passion-Driven Mastery: Specialization, independent research, entrepreneurship, leadership in student organizations.
4. Lifelong Learners (25+):
   1. Curiosity Activation: New hobbies, workplace challenges, emerging technologies.
   2. Engagement Sustainment: Online courses, communities of practice, structured learning paths.
   3. Personal Relevance Formation: Skill-building for career advancement, personal enrichment, social contribution.
   4. Passion-Driven Mastery: Becoming mentors, thought leaders, or experts; integrating learning into lifestyle.

**Interest Domains:**

1. STEM (Science, Technology, Engineering, Mathematics)
2. Arts & Creative Expression
3. Business, Economics & Entrepreneurship
4. Health, Medicine & Life Sciences

## **Question Dataset Structure:**

The **question dataset** serves as the raw material for the interest-detection system. Its design ensures clarity, balance, and age-appropriateness. Each question is tagged with metadata and structured to capture interest phases and domains.

**Key Components:**

1. **Metadata**
   * Each question is tagged with:
     + **Age group** (Primary, Secondary, Higher Education, Lifelong Learner)
     + **Interest phase** (Curiosity Activation, Engagement Sustainment, Personal Relevance Formation, Passion-Driven Mastery)
2. **Options and Answer Scoring**
   * Each question has **four options** labeled with unique codes (Opt\_1–Opt\_4)
   * Each option is mapped to **one or more domains** (STEM, Arts, Business, Health)
   * Scores reflect the **phase intensity** (e.g., higher for Passion-Driven Mastery)
3. **Age‑Appropriate Language**
   * Questions and options are tailored to cognitive level:
     + **Primary (6–12):** simple, concrete, activity-based
     + **Secondary (13–18):** identity and real-world relevance
     + **Higher Education (19–24):** abstract, career-oriented, research-driven
4. **Distractor Options**
   * Neutral or unrelated options are included to detect:
     + Low interest
     + Random answering
5. **Multi-Domain Signals**
   * Options may **naturally overlap domains** if they genuinely belong to more than one
   * No forced overlaps are introduced artificially
6. **Balanced Coverage**
   * Across the dataset:
     + All domains appear **roughly equally**
     + All phases are represented to avoid bias
7. **Sample Size**
   * Designed to accommodate **large, diverse response sets** per age group
8. **Synthetic + Real Data**
   * Initially generated synthetically
   * Later augmented with **real student responses** for improved accuracy

## **ML Training Dataset Structure:**

Once students respond to the questions, their **answers are transformed into a structured dataset** for ML training. This dataset captures **multi-dimensional signals** that indicate phase and domain interest.

**Key Components:**

1. **Response Records**
   * Each record corresponds to a student’s response to a question
   * Includes:
     + Question ID
     + Selected option code(s)
     + Age group
2. **Labels**
   * **Multi-output labels** for ML:
     + **Interest phase** (single label per question)
     + **Interest domain(s)** (multi-label possible if the option overlaps)
3. **Feature Encoding**
   * Student responses can be encoded for ML models:
     + One-hot or ordinal for option selection
     + Additional contextual features (age, gender, prior responses)
4. **Normalized Scores**
   * Option scores can be normalized across phases to standardize input to the model
5. **Balanced and Diverse**
   * Ensures **all domains and phases** are represented
   * Helps prevent model bias toward a particular domain or phase
6. **Multi-Domain Signals**
   * Preserves **natural overlaps** from the question dataset
   * Enables the model to learn **mixed interests**
7. **Dataset Size**
   * Large enough to allow **generalization** across students, age groups, and cultural contexts
8. **Synthetic + Real Responses**
   * Training starts with synthetic responses
   * Updated with **real student responses** for higher model fidelity

Dataset Resource: Real world uni and school data, Manual + LLM to generate and MCQs and refine them