colleague.ai

# An Introduction to AI:

# AI is Maths, Not Magic

A beginner's explanation of AI, large language models, and AI agents for K–12 educators.


In partnership with:

AmplifyLearn.AI

NSF National Science Foundation

U.S. Department of Education

IES Institute of Education Sciences

# AI is Maths, Not Magic


Before we discuss how to use AI or evaluate AI tools, we need to understand what AI actually is. There are three concepts every educator should understand: AI, large language models, and AI agents. Each builds on the one before.

## Part A: The Basics of AI

Artificial intelligence is software that learns patterns from data and uses those patterns to make predictions or decisions. The reason AI feels mysterious — or threatening, or astonishingly powerful — is not because of what it fundamentally is. It's because of the scale at which modern AI does this, and the sophistication of what it can predict.

Here is the distinction that matters most: traditional software follows rules a human explicitly programmed: “if this, then that.” AI software learns rules from examples. You don't tell it the rule; you show it enough examples that it figures out the pattern for itself. This is called machine learning, and it is the foundation of almost every AI system you will encounter.

Spam filters learn the patterns of unwanted email — certain phrases, sender behaviours, formatting patterns — and route messages away from your inbox automatically. Every time you mark something as spam, you are providing another training example.

NETFLIX Spotify

Netflix and Spotify recommendations predict what you want to watch or listen to next by identifying patterns in what you and behaviourally similar users have selected. The system doesn't know you enjoy thrillers — it knows that users with your pattern of choices often select certain other titles. Both have millions of users, billions upon billions of data points used by its models - what you clicked on, scrolled over, watched, partially watcher, rated, loved, etc.

Face ID on your phone recognises your face by learning the precise geometry of your facial features from thousands of reference images captured during setup. The phone doesn't "know" it's you — it detects a match between the incoming image and the stored learned pattern.

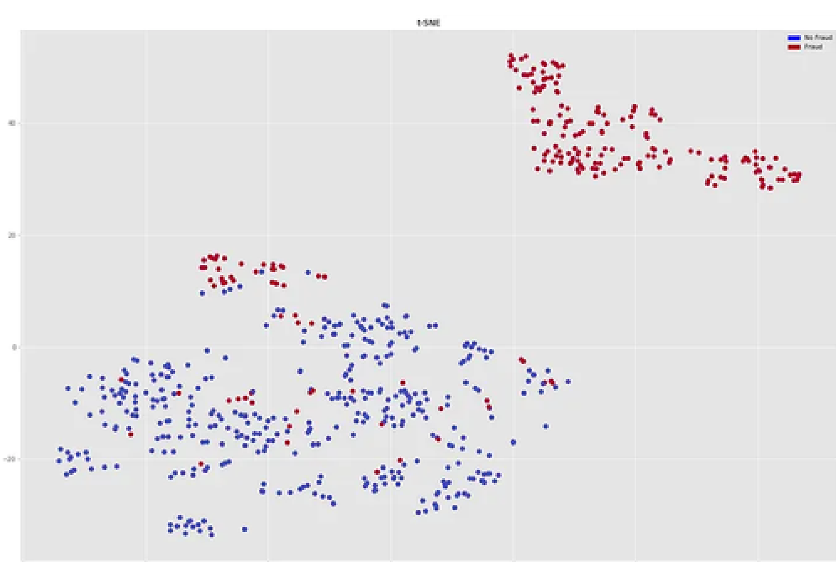
received revived re-sieve

Autocorrect and predictive text suggest the next word as you type by recognising statistical patterns in how words follow one another across billions of examples of written text. The phone isn't magic — it predicts what typically comes next in that context.

## Fraud Detection

Fraud detection on your bank account flags unusual purchases by learning what your spending pattern looks like and alerting when an incoming transaction deviates significantly from it.

Here a scatter plot illustrates a basic way to identify fraudulent transactions (red) from normal account holder transactions (blue):



This distinction explains both AI's strengths and its limitations. AI is extraordinarily good at spotting patterns in large, complex datasets — things humans would struggle to do at scale and speed. It is not, however, doing anything that a cognitive scientist would recognise as reasoning or thinking. It is not feeling. It does not know what is true. It is applied mathematics, running at a scale that was previously impossible.

None of these examples require the software to think, understand, or feel. They require it to find and apply patterns.

## Part B: The Basics of Large Language Models (LLMs)

A Large Language Model — an LLM — is a specific and very powerful type of AI trained on a vast amount of written text. We are talking about billions of documents: millions of books, millions of websites, billions of academic articles, news archives, research papers, forum threads, etc. The task this model is trained to do is deceptively simple: given a sequence of words, predict the word most likely to come next. This is called next-token prediction - not that more “magic” than autocorrection!

ChatGPT
Claude
Ollama
Gemini

**Search History Training Data**

why do cats vomit
why do cats vomit
why do cats vomit
why do cats pee in the garden
why do cats cough
why do cats vomit
why do cats vomit
why do cats smell bad
why do cats meow

Predict the next word for the following search phrase:

“why do cats _____”

**60%: “vomit”**
10%: “pee in the garden”
10%: “cough”
10%: “smell bad”
10%: “meow:

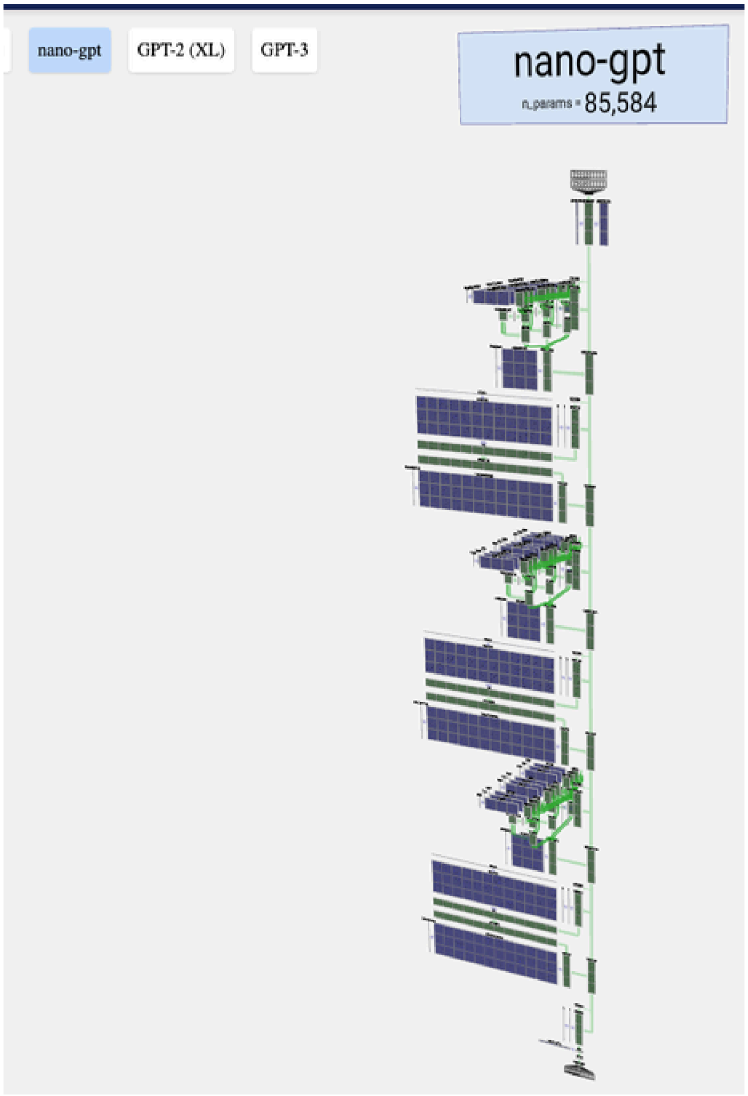
What makes LLMs remarkable is not that the task is complex — it isn't. It is that when you do this at enormous scale, with sufficient training data and computing power, something extraordinary emerges. The model produces text that is fluent, coherent, and often impressively accurate. It can explain concepts, draft emails, answer questions, summarise documents, generate lesson plans - all by prediction.

Here are the major models on the market in the recent months and years:

| **Provider** | **Flagship Models** | Known For |
|---|---|---|
| Google DeepMind | Gemini 3.1, 3.5 Flash | Reasoning benchmarks, video & audio models, Workspace |
| OpenAI | GPT-5.5, GPT-5.4, etc. | Multimodal (image, text, video, and audio in one), adoption and first to market |
| Anthropic | Opus 4.8, Sonnet 5, Fable 5 / Mythos 5, Haiku | Coding, long-form writing, safety focus |
| xAI | Grok 4, Grok 4.3 | Real-time X data access, huge context windows |
| Meta | Llama | Open-source ecosystem |
| DeepSeek | V4, R1 | Open model that can be customized |

The most important thing to understand about an LLM: it is not thinking. It is predicting. It has learned, from billions of examples, what fluent text in response to a given prompt looks like — and it produces text that matches that learned pattern. **Plausible** and **accurate** are not the same thing.

Here is a visualisation of one of ChatGPT's earlier models, and how it's built. You can see layers upon complex layers of model configurations and setup:




A model will produce a very confident-sounding paragraph about something entirely wrong, if that wrong information matches the patterns it learned. Imagine that all the data I used to train a word prediction model was and sounded like Shakespeare. The model would produce outputs that sounded like 16th-century English, not today's English. Understanding this should shape every interaction you have with an LLM-based tool.

One more concept worth keeping in mind: the model is not the tool you use. GPT-4, Claude, and Gemini are models — the underlying mathematical systems. ChatGPT, Colleague AI, and similar platforms are tools built on top of those models. The same underlying model can power very different tools, with very different outputs depending on how the tool has been configured, fine-tuned, and designed. **Knowing which model a tool is "powered by" tells you almost nothing useful about its quality for educational purposes.**

## Examples of LLMs in action:

**Asking an AI to explain a concept** and receiving a clear, well-structured paragraph is not evidence that the model understands the concept. It predicts what a good explanation looks like, based on the millions of examples it encountered during training.

**An AI translating a newsletter into another languag**e applies learned patterns about how words and structures correspond between languages. For common language, this works well. For technical, legal, or culturally specific content, errors are more frequent.

**An AI summarising a long document** is predicting which sentences best represent the whole — based on learned patterns about what summaries look like, not an understanding of the document's meaning.

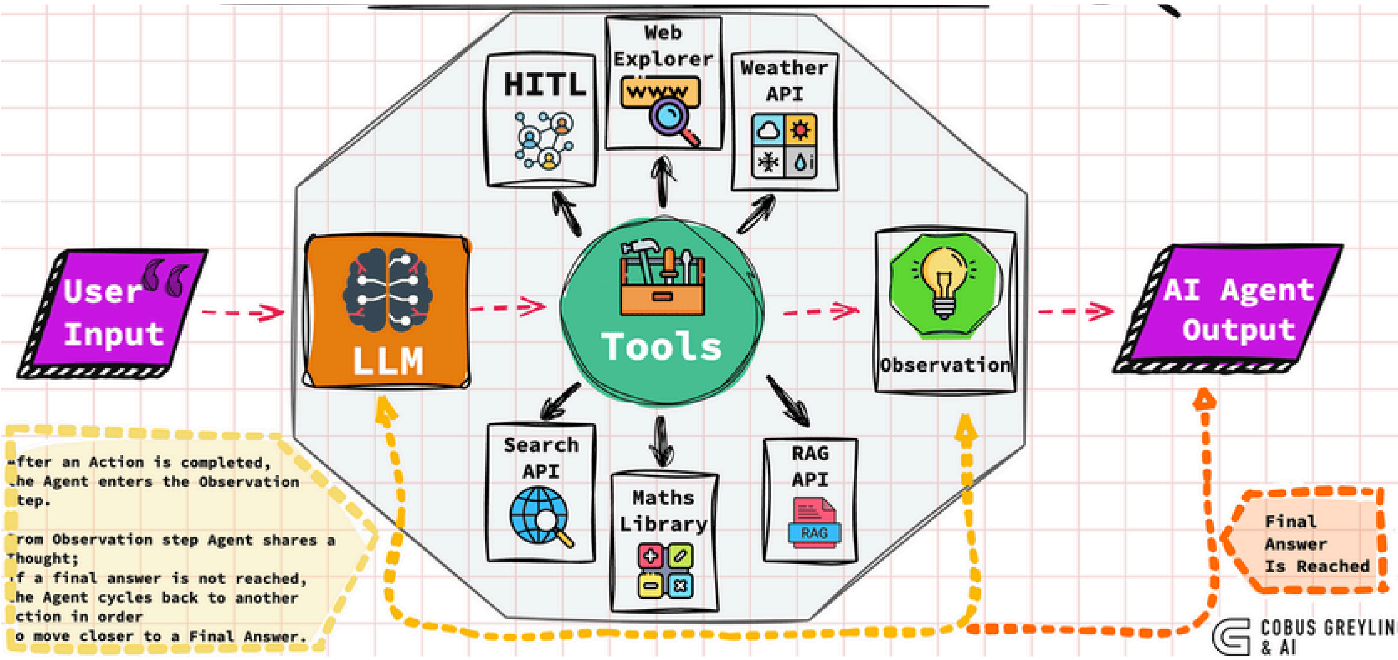
**A website chatbot answering a user's question** generates a plausible response from the text it has been trained on. If the correct answer has changed but the data hasn't been updated, the chatbot will give the wrong answer.

# Part C: The Basics of AI Agents

If LLMs represent a significant leap from traditional software, AI agents represent the next step forward from LLMs.

An LLM can draft an email, but it can't actually send one. It doesn't have access to an emailing tool.

Now imagine if you gave an LLM access to tools: Gmail, Microsoft Outlook, Google Drive, Microsoft Teams, Slack, searching the internet on a web browser, Canvas, Google Classroom, Sigma, your personal calendar, your school's LMS, etc. This is essentially what an agent is - and **it's far more powerful than just an LLM now that it can do things.**




An LLM responds to a prompt with a piece of text. An AI agent takes a goal and figures out what to do about it — carrying out a sequence of actions, using tools like web search, databases, calendars, or file systems, checking its progress, adjusting its approach, and continuing until the goal is achieved. Where an LLM answers your question, an agent carries out your request.

## Examples of AI Agents:

A scheduling agent reads your calendar, identifies open meeting slots, drafts an email offering those options, sends it, monitors for a reply, and adds the confirmed meeting to your calendar.

A research agent takes a topic, searches multiple websites, reads and compares the results, evaluates relevance, and produces a cited summary — without you directing each individual step.

A grading agent opens student submissions one by one, applies your rubric, writes individualised feedback, assigns scores, and populates your gradebook, using a variety of tools.

AtlantiCare REGIONAL MEDICAL CENTER

AtlantiCare, a regional healthcare system in New Jersey, deployed an AI documentation agent that listens to patient consultations and generates structured clinical notes directly into the electronic health record. Adoption was strong and resulted in a 42% reduction in documentation time, and 66 min. saved per clinician per day.

General Mills

General Mills deployed an AI-driven supply chain optimization system that assesses 5,000+ daily shipments and has produced over $20 million in savings since fiscal 2024. The system autonomously evaluates shipment routing, timing, and vendor performance, only flagging exceptions for human review rather than requiring approval on every decision

JPMorganChase

JPMorgan runs agents that generate investment banking presentations in 30 seconds compared to the hours junior analysts previously spent, draft M&A memos, automate trade settlement, and detect fraud in real time. They have hundreds of AI agents across their organization that help to reduce workload and speed up tasks.

The practical implication: with an LLM, you write a prompt and receive a response. With an agent, you state a goal and the system determines the steps. This is a significant advancement in what AI can do. It means AI can increasingly handle tasks that previously required a human to manage each step of the process.

This doesn't mean agents are autonomous or infallible. They make errors, take wrong turns, and require oversight. But the trajectory is clear: AI is moving from text generation toward task completion. **Educators who understand this difference will be better positioned to evaluate, adopt, and appropriately oversee these tools as they become more prevalent in schools.**

# Where The "Intelligence" Comes From


If AI learns from data, then the data used to create patterns and predictions matters enormously. Not just how much of it there is, but what it contains, where it came from, and what it does and doesn't represent.

Training data is the body of examples a model learns from. For an LLM, this typically includes billions of web pages, digitised books, academic articles, and other written material. The model has learned the patterns of language from this data — which means it has also absorbed the errors, biases, gaps, and inaccuracies in that data. The principle of garbage in, garbage out predates AI by decades, and it applies here with particular force. If the training data is unrepresentative, biased, or contains errors, the model's outputs will reflect that. What if you trained a tool for ages 7-8 on only adult conversations and texts?

Think of a machine learning model like a recipe that a cook is still perfecting. The "parameters" are like the amounts of each ingredient—how much salt, how long to bake, how much sugar—that the cook adjusts until the dish tastes right. In a machine learning model, these parameters are numbers the computer tunes, little by little, based on examples it's shown (like past data), until its "recipe" for making predictions gets as accurate as possible. A simple model may have a few of these adjustable numbers. But large, powerful models—like the ones behind chatbots or image generators—can have billions of parameters, each one a tiny dial that gets nudged during training.

# Other Types of Models

Large language models aren't the only kind of AI. The same basic approach — find patterns in large amounts of data, make predictions — has been applied to almost every kind of information humans produce.

**Image generation models** (Midjourney, Adobe Firefly) are trained on millions of image-and-caption pairs. They learn the relationship between words and visual patterns, then generate an image from a text description and other information.

**Video generation models** (Sora, Runway) extend this to motion, predicting how a scene looks and moves across time. Still visibly imperfect, but improving quickly.

**Voice and speech models** convert text to natural-sounding audio, transcribe speech to text, or — in more specialised applications — detect and classify accents.

Derm Foundation is a machine learning model for skin image analysis for dermatology applications.

WeatherNext is a family of state-of-the-art AI weather forecasting models from Google DeepMind and Google Research

MARS8.1 is the latest generation in CAMB.AI's MARS series of speech synthesis models that creates speech in 200+ languages.

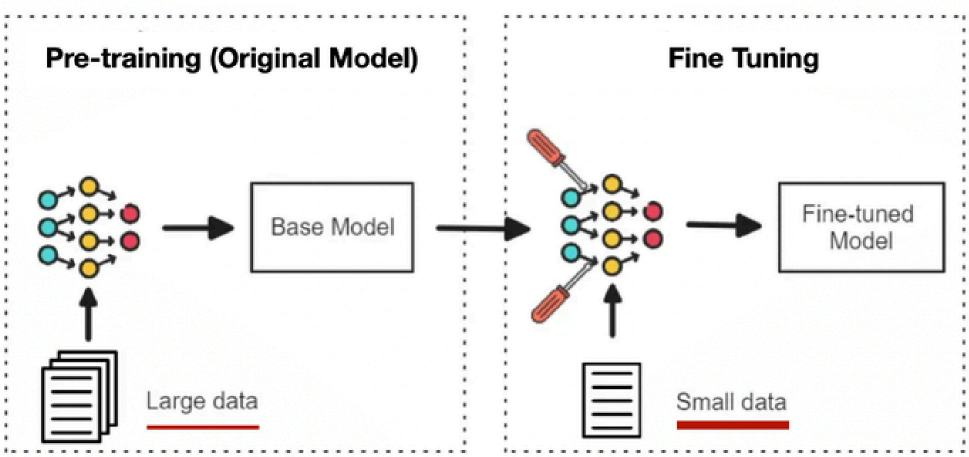
Purpose-built models are trained on narrow, domain-specific datasets to do one thing well. Dermatology models trained on clinical photographs have matched specialist accuracy at detecting certain skin cancers. Models have also detected early signs of diabetic retinopathy, or flagged anomalies in radiology scans.

The tradeoff is always the same: a purpose-built model outperforms a general model on its specific task, but can do nothing else. When you hear "AI can do X," it's worth asking which AI — a dermatology model and a general-purpose assistant are quite different things, even if both get called "AI."

# Fine-Tuning a LLM — Why It Changes Everything

Training a large model from scratch requires vast computing resources and data — the kind of investment that only very large technology companies can sustain. It cost $500M to $1B for OpenAI to develop GPT-5, and it cost Anthropic about $1B to develop its Opus model.

Fine-tuning is a different, more accessible process: taking an already-trained model and continuing to train it on a smaller, more specific dataset, to specialise its behaviour for a particular domain or task:



Think of it this way: **the base model is a highly educated generalist. Fine-tuning is specialist training**. A doctor, a lawyer, and an accountant all have some shared educational foundation — but each has undergone specific training that shapes how they apply general knowledge within their professional domain.

Fine-tuning in other industries shows how much specialization matters:

A medical AI that detects tumours in radiology scans is fine-tuned on hundreds of thousands of labelled medical images, reviewed by expert radiologists. The same base image model, without that fine-tuning, would not perform reliably in a clinical setting — and using it there would be dangerous.

A legal document AI trained on millions of contracts, case law records, and regulatory filings behaves very differently from a general-purpose text model when reviewing a lease agreement. It recognises clause structures, flags unusual terms, and applies legal conventions a general model would not.

A financial services AI trained on regulatory guidance, risk frameworks, and market data can identify compliance issues that a general-purpose model would completely miss. A genreal model could give you an overview or summary, but not rigorously identify them.

In each case, the domain-specific dataset is what creates the specialist. The base model provides language fluency; the fine-tuning provides expertise.

**What "fine-tuned for education" should mean**: it should mean the model has been trained on:

- high-quality educational materials — curriculum frameworks, standards documents, exemplary lesson plans
- learning science research
- pedagogical models
- teacher professional development literature
- And more.

# Pedagogy in AI Models Matters!

LLMs such as ChatGPT and Gemini are built to generate coherent, relevant, fluent text in response to prompts. That is a meaningfully different goal from supporting a ten-year-old's development of reading comprehension, or scaffolding a high school student's analytical writing, or providing a first-year teacher with professional feedback aligned to a specific instructional framework.

Generic models struggle to be deeply pedagogical by default. A model trained on the general internet has learned the broad patterns of human language. It has not learned what a developmentally appropriate explanation for a Year 2 student looks like compared to a Year 6 student. It hasn't learned how to scaffold complex concepts using evidence-based practices. These are pedagogical distinctions that require pedagogical training data.

Generic models are not grade-level aware. Without specific fine-tuning and configuration, a general-purpose model generates text at whatever level seems most natural based on the prompt — which typically defaults to adult-appropriate language and cognitive demand. Ask it to explain a concept "for students" and it may lower the reading level somewhat, but it will not reliably calibrate to the specific vocabulary range, conceptual readiness, and developmental stage of a particular year group.

## Quality K12 AI Requires Deliberate Pedagogical Design

This means the training data should include evidence-based instructional practices, learning science frameworks, curriculum standards, teacher professional development models, assessment design principles, and so much more! It means outputs should be validated against those frameworks — not just checked for grammatical fluency. It means the model's outputs for a Grade 3 classroom should look meaningfully different from its outputs for a Grade 10 classroom, in ways that reflect genuine understanding of what students at each level are ready for.

| Pedagogically-aware, well-built AI | General or poorly built K12 AI |
|---|---|
| Lesson plans are aligned to specific standards and include Bloom's-level objectives appropriate to the year group | A lesson plan that could apply to any grade, any subject, and any school with minimal adjustment |
| Formative assessments are calibrated to the cognitive demand expected at that level — not just recall when analysis is required | A quiz that tests recall when the standard requires analysis or application |
| Differentiated materials maintain the same core learning goal while adjusting access scaffolding — not simply shortening the text | A "differentiated" version that is simply shorter rather than more accessible or that doesn't actually challenge an advanced learning, but gives students more practice |
| Professional development suggestions reference named instructional frameworks and published research | A professional development suggestion that references no specific framework |

**The student growth implications are real.**

At scale — across a classroom, a school, a district — the difference between pedagogically-designed AI and generic AI compounds. Students who engage repeatedly with AI-generated content set at the wrong cognitive level, misaligned to standards, or pedagogically incoherent are not simply getting less than they could. In some cases, they are practising the wrong things. **Educators who understand this distinction are in a position to protect their students from it.**

# Using AI As An Educator

We have established the foundations: AI is pattern recognition and prediction, not reasoning. It can be fine-tuned for specific purposes, with significant variation in quality. Pedagogical design matters — and the absence of it has real consequences in the classroom.

So where does that leave you, as an educator, in a world where AI tools are proliferating rapidly? The answer is not to avoid AI. The answer is to engage with it as the professional you are — with the same critical judgment you bring to textbook selection, technology adoption, and curriculum design.

Here are some common use cases for the four stakeholder groups most directly affected by AI in schools:

### As Students:

- Receiving personalised explanations of concepts calibrated to their level
- Practising skills with immediate, specific feedback
- Getting help drafting and improving written work
- Getting help in starting an assignment if stuck
- Accessing content in their home language
- Using AI as a research starting point — not a final source


## As Teachers:

- Drafting lesson plans and unit outlines aligned to standards, to district curriculum, to specific requirements, and based on a class's previous performance
- Creating differentiated versions of materials for different learner profiles
- Generating formative assessments, rubrics, and exit tickets
- Drafting parent communications and newsletters
- Synthesising student performance data to identify intervention needs
- Accessing targeted professional development support and coaching prompts

## As School Administrators:

- Drafting staff communications, newsletters, and board updates
- Generating and collecting surveys and feedback data, creating summaries
- Supporting scheduling and administrative correspondence
- Analysing student usage of AI tools
- Generating curricula documentation for all class, semesters
- Supporting teacher professional development planning, curriculum, content and activity development

## As District Leaders:

- Synthesising performance data across schools and cohorts
- Generating board-facing reports and presentations
- School improvement plans
- Budgeting and resource allocation tools
- Supporting curriculum coherence analysis across the district
- Identifying professional development gaps across a teaching workforce

## Reading AI Outputs Critically

Being a critical consumer of AI outputs is a professional skill that develops with practice. It's not enough to ask "does this look right?" The better questions are:

- Is this appropriate for my students? Consider grade level, vocabulary, cognitive demand, and cultural context.
- Is this accurate? AI generates plausible text. Check factual claims independently before using them with students.
- Does this serve the right learning goal? A well-presented activity that doesn't align to the learning objective is still the wrong activity.
- Has this been calibrated to pedagogy, or is it generic? A lesson plan that could work for any grade, any subject, and any school is a signal that the tool hasn't been fine-tuned for education.
- Would I sign my name to this as a professional? You remain responsible for what you use in your classroom. If the answer is not yet — edit before you deploy.

Your professional role hasn't changed. AI is a tool — and not necessarily a good one. Your judgment about your students, your curriculum, your standards, and your community remains the primary input in every instructional decision. AI can inform, generate, draft, and suggest. You decide.

# Distinguishing High Quality from Low Quality AI in K-12

Not all AI tools are created equal. In the education market in particular, the gap between a carefully built educational AI platform and a general-purpose AI with an educational interface can be enormous — even when both look polished in a demonstration. Here is what to look for.

**Transparency about limitations.** The tool is honest about what AI can and cannot do, and is designed to support teacher oversight rather than bypass it.

**Published research or evidence.** There are peer-reviewed studies, published pilot outcomes, or independent evaluations of the tool's effectiveness. Not testimonials, but real research.

**Documented training methodology.** The company can explain what data the model was trained on, who curated that dataset, and how outputs were evaluated.

**Visible grade-level calibration.** Outputs are noticeably different across grade levels — not just simpler vocabulary, but cognitive demands, scaffolding approaches, and objective alignment.

**Pedagogical alignment.** Lesson plans reference named instructional frameworks. Assessments are mapped to specific Bloom's or DOK levels. Differentiated materials scaffold access to the same learning goal.

**Educator, administrator, and district leader involvement.** The company worked with teachers, students, administrators, superintendents, CTOs/CIOs, principals, and education researchers in building the model — not just the product visuals.

## Signs of a generic tool dressed up in K-12 language:

**Vague claims with no evidence.** "AI-powered learning," "personalised for every student," "designed for education" — with no explanation of how, and no published research or data.

**No data documentation.** When you ask about FERPA compliance, student data handling, or training methodology, the answers are vague, delayed, or unavailable.

**Outputs that require significant teacher correction.** Every output needs substantial editing before it is classroom-ready. Output lacks high enough quality and fails nuances, details, or adjusting to specific students and classes.

**Heavy reliance on prompt engineering.** The tool requires you to specify grade level, standard, Bloom's level, tone, and format in every prompt. A K-12 tool should already know these things and apply them by default.

Try Our Free AI Mini Class!

# From Understanding to Action

You've just done something not every educator has done: taken the time to understand what AI actually is, how it works, and what separates good & bad tools.

That knowledge is immediately practical. AI is not going away. In the coming years, the tools will become more capable, more present, and more integrated into schools. Educators who understand the fundamentals will be better positioned to shape how AI is used in their classrooms and institutions, and to ask smart questions, such as choosing between:

1. A generic AI model trained on the general internet, configured with school-friendly language, and presented through an educational interface?
2. An AI model trained on a curated dataset of instructional best practices, peer-reviewed pedagogical research, evidence-based teaching frameworks, and teacher professional development models — validated by education researchers and designed specifically for the K-12 context?

The difference matters!

Now is your chance to explore with fresh eyes. **Try the Colleague AI platform** — built by researchers, with educators, grounded in learning science — and see for yourself what the difference looks like in practice.

Or try our free AI mini-class to start using AI tools.

colleague.ai

**Questions?**
**Contact us: hello@colleague.ai**

Colleague.ai
Kirkland, WA

## In Partnership With:

colleague.ai

AmplifyLearn.AI

National Science Foundation

IES Institute of Education Sciences

U.S. Department of Education