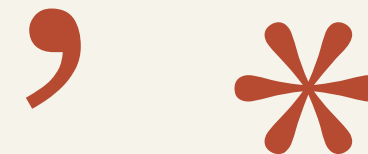


SurfacePrint

Typographic Fingerprints
of LLMs

Yuntian Deng

CIFAR meeting, October 14, 2026



The words match. The apostrophes differ.

don't

Straight apostrophe

don't

Curly apostrophe

People, models and editing software can produce either

Illustrative example

A formatting habit becomes a measurement

Five apostrophes in an illustrative text

‘ ‘ ‘ ’ ’

3 / 5

60% curly

One of 97 measurements from raw text

Example feature: curly apostrophes divided by all apostrophes. These are illustrative counts.

A small classifier scores the formatting



Higher scores indicate more model-like formatting

No language-model calls when scoring a new text

The classifier learns from labeled human and generated examples

We test new texts from 17 known model sources

Train

Human text + outputs
from **17 AI models**

Test

Held-out texts
from those same sources

About 8,000 human and 8,000 generated training examples

Programming / Math / Explanations / Fiction

Human posts from StackExchange and Reddit. All 17 models are represented in training.

AUROC measures how well the scores rank texts

How often does generated text
receive a higher score than human text?

0.5

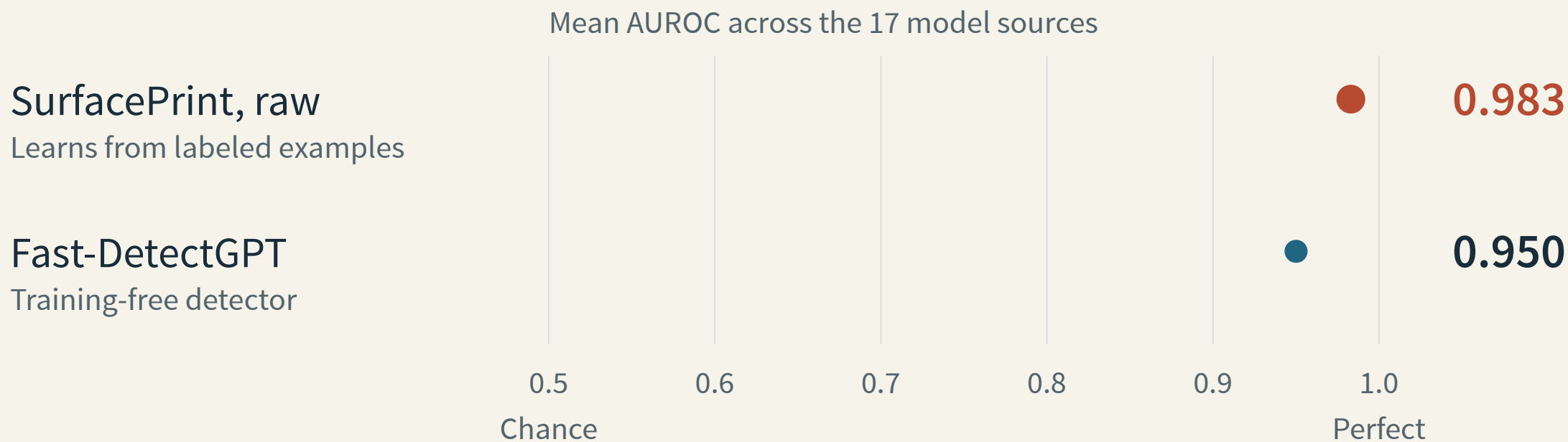
Chance ranking

1.0

Perfect ranking

An evaluation over many texts. A single document's probability of AI authorship requires more evidence.

Formatting strongly separates the benchmark texts



Both methods use the same held-out test texts. They use different amounts of supervision.

A signal remains after whitespace cleanup

Two variants, each trained on its own version of the text

Raw text

First paragraph.

Second paragraph.

0.983 AUROC

Whitespace standardized

First paragraph.

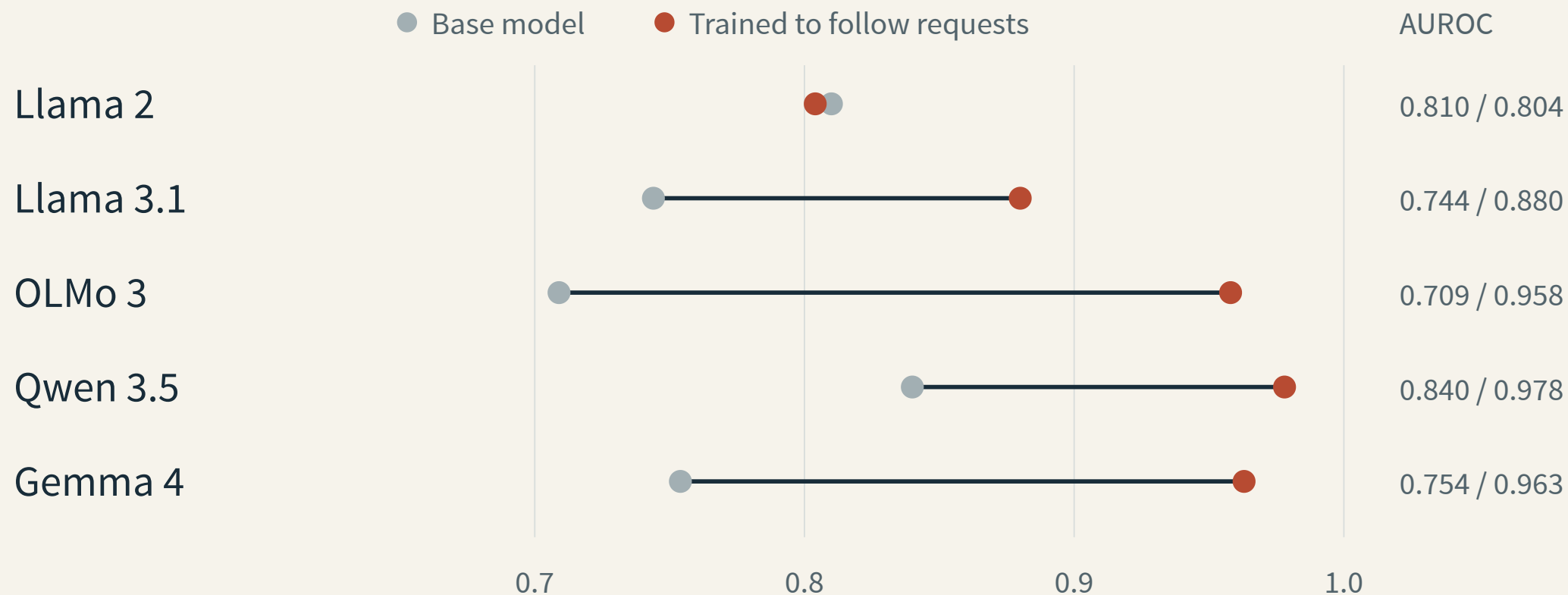
Second paragraph.

0.889 AUROC

Illustration: repeated blank lines become one. SP-WS also standardizes line endings and trailing whitespace.

Instruction-tuned text is often easier to detect

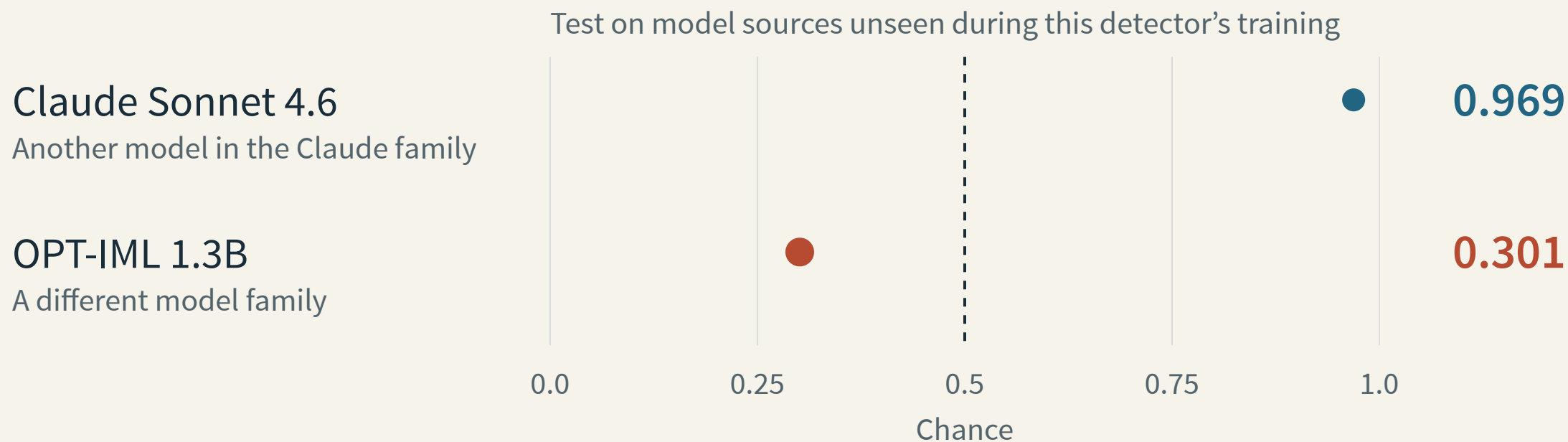
Separate whitespace-standardized detectors for each condition



Higher AUROC means easier to distinguish from human text. Four of five model families improve.

Transfer to new model sources is uneven

New experiment: train on human text + Claude Sonnet 4 only



One SP-WS detector. AUROC below 0.5 means the learned ranking often runs in the wrong direction.

Formatting edits change what we measure

Original text

A simple answer

It doesn't change.

After a formatting edit

A simple answer

It doesn't change.

The message remains while the measured features change

Illustration only. We have not measured the detector's response to this particular edit.

Practical use needs a measured false-positive rate

How often would human documents be wrongly flagged?

A score cutoff and a representative test set make this measurable

We have not yet measured that error rate in the intended application setting.

Formatting offers an additional detection clue

Its reliability depends on the source and processing of the text

Questions

An AUROC score does not set the false-positive rate

**A cutoff turns a score
into a flag**

Different cutoffs trade missed AI text against wrongly flagged human text

AUROC summarizes ranking across cutoffs. It does not select an operating point.

The classifier combines many small decision trees

300

trees

6

maximum depth

3

minimum leaf size

97 measurements, balanced class weights, fixed seed

A decision path can use only formatting tests

- 1 Markdown bold present?
 Yes
- 2 Double-backtick inline code present?
 No

Result Tree predicts **LLM**

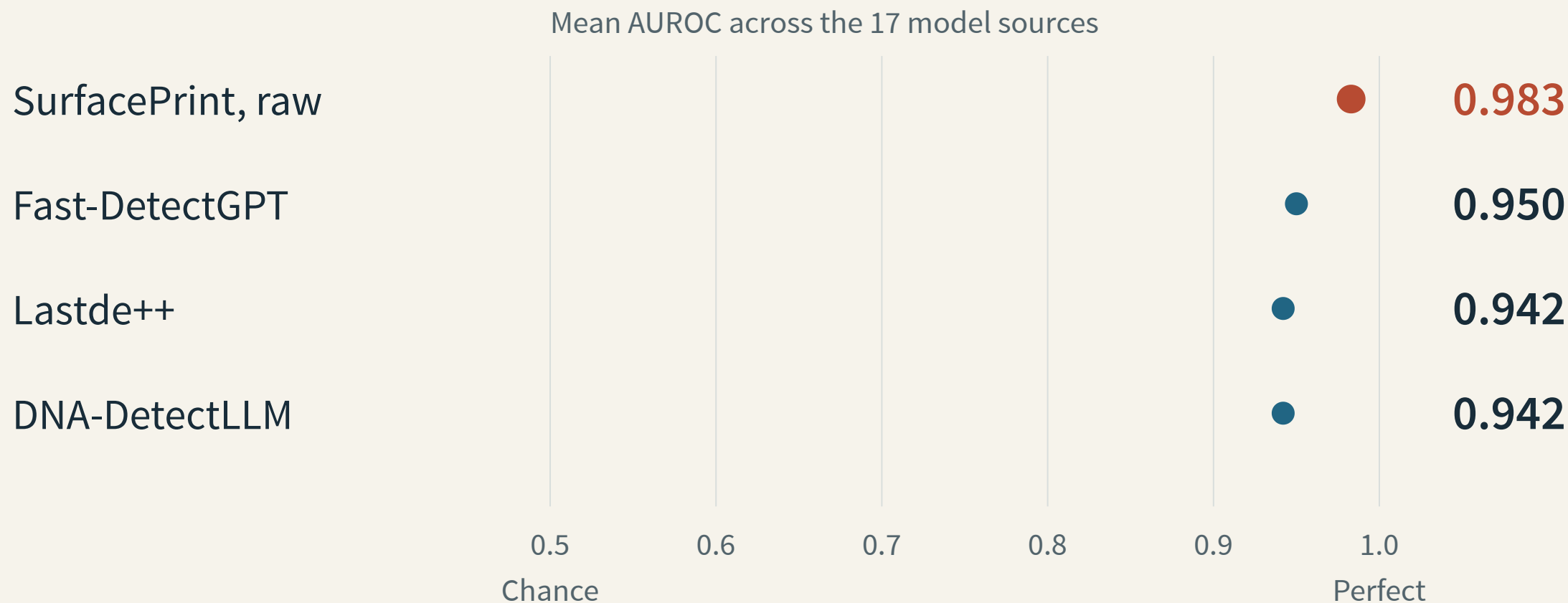
One actual path in the illustrative depth-3 tree. Its AUROC is 0.7497; the main forest is a different model.

The 97 measurements cover five surface families

FAMILY	EXAMPLE	COUNT
Prose typography	' '	22
Markdown	**bold** <u>__bold__</u>	25
Math markup	$\$x\$$ $\backslash(x\backslash)$	24
Code formatting	Tabs spaces	20
Whitespace	Blank lines	6

Features include rates, densities and indicators computed from raw text

The main reference-model comparators



Same held-out texts. SurfacePrint learns from labels; these baselines are training-free detectors.

The source pool and detector sample differ

Source pool, per domain

2,000 train / 500 validation / 1,000 test prompts

Main detector's training sample

8,000 human posts + 470 outputs per AI model

$17 \times 470 = 7,990$ generated examples

Four source domains. Held-out test examples do not enter detector training.

SP-WS standardizes selected whitespace conventions

Line breaks	Use one line-break encoding
Spaces at line ends	Remove them
Repeated blank lines	Keep one blank line
Final line break	Remove it

Punctuation and markup remain in the input

SP-WS is a separately trained variant. Some layout information remains.

A next evaluation would test real document handling

Representative documents

Human and generated text from the intended setting

Real processing steps

Include ordinary editing, copying and extraction

Held-out error measurement

Report false positives and detection at a fixed cutoff

Proposed follow-up evaluation, not a completed experiment