

Supplementary material for
SpecLLM: LLM-Based VNIR Spectral Interpretation

Michał Romaszewski, Anna Strzoda, Marzena Halama,
Przemysław Głomb, and Piotr Masłowski

Supplementary Note 1: Enlarged class and spatial-sampling maps

Figures S1 and S2 provide full-page versions of the sampling maps summarised in Figure 3 of the manuscript. For each scene we show a simplified RGB rendering, the class map of the evaluated substances, and the spatially disjoint training and test patches with their guard zones. Scene F contains seven evaluated classes; Scene E contains six, because beetroot juice is not present in that scene. The maps were generated from the public HyperBlood annotation and the stored split metadata.

(a) Scene F: simplified RGB rendering



(b) Scene F: class map of evaluated substances



(c) Scene F: spatial sampling, split 00

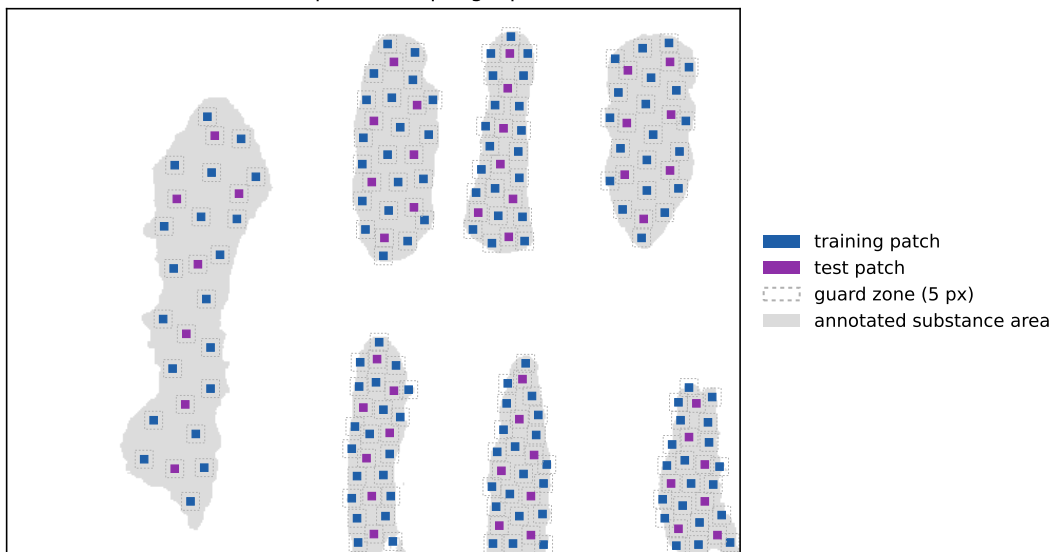


Figure S1: Scene F: (a) simplified RGB rendering, (b) class map of the seven evaluated substances, and (c) the spatially disjoint training and test patches with their guard zones.

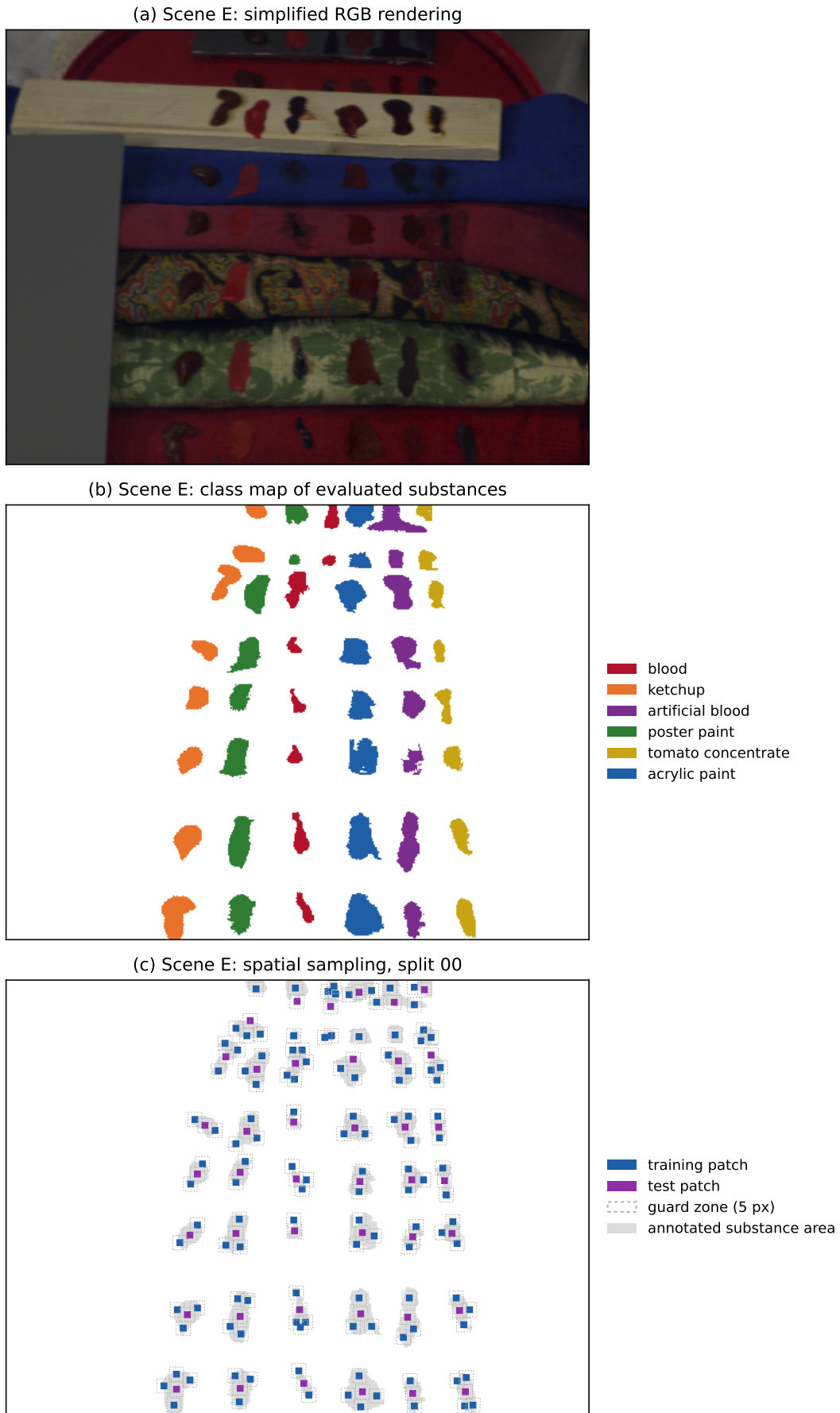


Figure S2: Scene E: (a) simplified RGB rendering, (b) class map of the six evaluated substances, and (c) the spatially disjoint training and test patches with their guard zones. Beetroot juice is not annotated in this scene.

Supplementary Note 2: Model-reported confidence

Cell-level confidence diagnostics for the natural-language feature, JSON feature and raw CSV representations, computed as described in the Methods section of the manuscript.

Each cell covers the valid outputs that also carried a parsable confidence value, which excludes 3 (NL features), 5 (JSON features) and 7 (raw CSV) of the 3,900 outputs per representation whose confidence field was missing or unusable.

Table S1: Confidence diagnostics in Scene F. Accuracy is calculated among valid outputs. Gap denotes the absolute difference between mean confidence and accuracy.

Model	Representation	Mean confidence	Accuracy	ECE	Gap	AUROC
Llama4-scout	NL features	0.780	0.659	0.122	0.121	0.576
	JSON features	0.780	0.628	0.176	0.152	0.557
	raw CSV	0.799	0.438	0.366	0.360	0.613
Qwen14b	NL features	0.848	0.603	0.248	0.245	0.524
	JSON features	0.856	0.629	0.237	0.227	0.595
	raw CSV	0.870	0.264	0.606	0.605	0.499
gpt-oss-20b	NL features	0.771	0.770	0.048	0.001	0.736
	JSON features	0.776	0.703	0.081	0.073	0.717
	raw CSV	0.787	0.671	0.120	0.115	0.666

Table S2: Confidence diagnostics in Scene E. Accuracy is calculated among valid outputs. Gap denotes the absolute difference between mean confidence and accuracy.

Model	Representation	Mean confidence	Accuracy	ECE	Gap	AUROC
Llama4-scout	NL features	0.727	0.498	0.260	0.229	0.555
	JSON features	0.766	0.433	0.345	0.332	0.551
	raw CSV	0.871	0.192	0.679	0.679	0.588
Qwen14b	NL features	0.810	0.523	0.298	0.288	0.550
	JSON features	0.830	0.540	0.292	0.290	0.608
	raw CSV	0.883	0.199	0.684	0.684	0.554
gpt-oss-20b	NL features	0.653	0.514	0.172	0.139	0.632
	JSON features	0.709	0.516	0.205	0.193	0.649
	raw CSV	0.722	0.412	0.344	0.310	0.489

Mean confidence changed much less than accuracy across representations. AUROC ranged from 0.489 to 0.736 and was mostly below 0.62, showing that confidence was generally only weakly predictive of correctness. The lower ECE observed in some feature-based cells should therefore not be interpreted as an independent advantage: much of the change reflects higher accuracy at a similar mean confidence.

Table S3: Reliability bins pooled descriptively across the six model–scene cells for each representation. The cell-level metrics remain those in Tables S1–S2. Empty bins are marked with a dash.

NL features				JSON features				Raw CSV			
Bin	n	Conf.	Acc.	Bin	n	Conf.	Acc.	Bin	n	Conf.	Acc.
0.0–0.1	2	0.050	0.500	0.0–0.1	1	0.050	0.000	0.0–0.1	0	–	–
0.1–0.2	17	0.142	0.353	0.1–0.2	6	0.150	0.333	0.1–0.2	7	0.150	1.000
0.2–0.3	63	0.226	0.333	0.2–0.3	42	0.218	0.429	0.2–0.3	21	0.238	0.333
0.3–0.4	44	0.327	0.477	0.3–0.4	33	0.323	0.333	0.3–0.4	25	0.322	0.400
0.4–0.5	52	0.416	0.538	0.4–0.5	30	0.418	0.400	0.4–0.5	22	0.426	0.455
0.5–0.6	124	0.548	0.492	0.5–0.6	89	0.547	0.438	0.5–0.6	58	0.549	0.414
0.6–0.7	543	0.621	0.477	0.6–0.7	434	0.627	0.468	0.6–0.7	283	0.630	0.417
0.7–0.8	481	0.741	0.559	0.7–0.8	492	0.743	0.508	0.7–0.8	319	0.742	0.433
0.8–0.9	1888	0.824	0.629	0.8–0.9	2037	0.824	0.572	0.8–0.9	1667	0.828	0.347
0.9–1.0	650	0.919	0.715	0.9–1.0	715	0.920	0.771	0.9–1.0	1380	0.915	0.353

Supplementary Note 3: Saved-prompt character counts

Prompt size was measured on the complete saved prompt files, including system and user messages, for all 700 Scene-F test samples for gpt-oss-20b.

Table S4: Character counts for saved Scene-F prompts.

Representation	n	Mean	Median	Range
raw XML	700	13,316	13,316	13,316–13,316
raw JSON	700	5,786	5,786	5,774–5,796
raw CSV	700	5,769	5,769	5,757–5,779
statistics + features JSON	700	5,040	4,979	4,807–5,556
features JSON	700	4,972	4,910	4,740–5,488
statistics + features NL	700	4,534	4,484	4,340–4,952

Supplementary Note 4: What the model receives

This note illustrates the input representation on two individual test spectra. Figure S3 shows the preprocessed spectra with the extrema detected by the deterministic encoder, together with the wavelengths of the class-archetype features that the prompt contains. The caption reproduces, for the blood spectrum, the complete text supplied to the model and the answer returned by it, so that the evidence available to the model can be read directly.

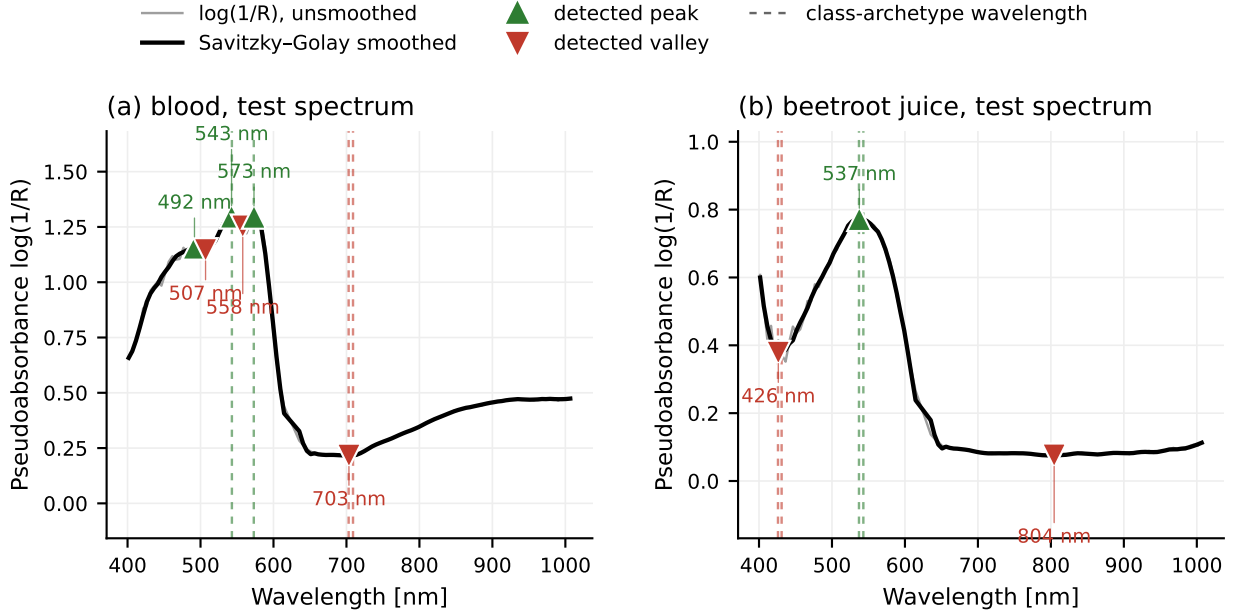


Figure S3: What the model receives. (a) A blood test spectrum and (b) a beetroot-juice test spectrum from Scene F, after the $\log(1/R)$ transformation (grey) and Savitzky-Golay smoothing (black). Triangles mark the extrema detected by the encoder and are labelled with their wavelength; dashed vertical lines mark the wavelengths of the selected class-archetype features of the corresponding class. The corresponding heights, widths and prominences or depths are listed in full below. The two blood peaks near 543 and 573 nm and the valley near 703 nm coincide with the blood archetype, whereas the beetroot-juice spectrum shares only the peak near 543 nm. For spectrum (a), the natural-language description generated by the encoder and supplied in the user message is, verbatim: *Pseudoabsorbance_log1r spectrum with minimum 0.2153, maximum 1.297 and mean value 0.6238, with prominent features: 491.72 nm peak (height 1.1568, width 12.4219 nm, prominence 0.0121), 506.91 nm valley (minimum 1.1447, width 14.8061 nm, depth 0.0121), 542.54 nm peak (height 1.297, width 166.5051 nm, prominence 0.639), 557.89 nm valley (minimum 1.2518, width 18.4396 nm, depth 0.0447), 573.28 nm peak (height 1.2965, width 12.9731 nm, prominence 0.0447) and 703.4 nm valley (minimum 0.2153, width 168.1968 nm, depth 0.2585).* The blood archetype contained in the system prompt is: *Pseudoabsorbance ($\log(1/r)$) spectrum with minimum 0.221, maximum 1.1072 and mean value 0.5643, with prominent features: 543.0 nm peak (height 1.1025, width 119.447 nm, prominence 0.3106), 573.0 nm peak (height 1.0983, width 53.7761 nm, prominence 0.1401), 703.0 nm valley (minimum 0.215, width 167.4714 nm, depth 0.225) and 709.0 nm valley (minimum 0.2262, width 170.9208 nm, depth 0.182).* The saved output of gpt-oss-20b for this spectrum is *label = blood, confidence = 0.95, reason = "Peak and valley positions match blood's signature."* The description also lists two near-threshold extrema (prominence and depth 0.0121, against the detection threshold $p_{\min} = 0.01$): the encoder reports the most prominent detected extrema without filtering them further.