Multi-Source Emotion Annotation in Children's Language: When LLM Consensus Diverges from Human Judgment

Farida Saïd[1], **Jeanne Villaneau**[2]
[1]LMBA, Université de Bretagne Sud, France
farida.said@univ-ubs.fr

[2]IRISA, Université de Bretagne Sud, France
jeanne.villaneau@univ-ubs.fr

Abstract

Automated emotion annotation increasingly relies on inter-LLM agreement as a proxy for label quality. We test this assumption on 2,106 clause-level segments from interviews with French-speaking children (ages 6–11) about parental roles, a setting where affect is often implicit rather than lexically explicit. Using a 500-segment expert gold standard, we show that internal consensus can be seriously misleading: Dawid–Skene, a probabilistic label aggregation method, estimates GPT-5.2 valence accuracy at 90.7%, whereas evaluation against human gold yields 71.0%, revealing substantial overestimation driven by shared neutralization bias. Conversely, Dawid–Skene underestimates Claude Sonnet 4, reversing model ranking. Majority Vote, Dawid–Skene, and MACE produce near-identical consensus labels, suggesting that the main source of error lies in shared annotator bias rather than in the aggregation rule itself. We release the expert gold subset and the probabilistic corpus to support future work. Our results show that high inter-LLM agreement cannot replace external human validation for affect annotation.

Keywords: emotion annotation, child language, implicit affect, LLM annotation, annotation aggregation, human validation, gold standard, valence

1. Introduction

Emotion analysis in short, transcribed oral responses by children is challenging: telegraphic syntax, disfluencies, and pervasive implicit affect limit the reliability of sentiment tools trained on adult corpora (Ogren and Sandhofer, 2021; Sagae et al., 2010; Liu and Chen, 2024). Yet understanding how children emotionally frame family relationships is relevant to developmental psychology, education, and child welfare, while automated annotation for this population remains underexplored (Pérez-Espinosa et al., 2020).

We study French transcriptions of children's oral interview responses about the parental roles Father and Mother (Bellachhab et al., 2025). The dataset contains 2,106 clause-level segments requiring emotion (valence, intensity) labels, and exhaustive expert annotation is cost-prohibitive. Before scaling an automated pipeline, we ask: can inter-LLM consensus be trusted as a quality signal for emotion annotation, or can shared model bias yield deceptive agreement?

We construct a 500-segment stratified expert gold standard and use it to assess both LLM performance and Dawid–Skene as a quality estimator. We show that when annotators share systematic biases (notably over-neutralization), aggregation can mistake correlated errors for consensus.

Our experimental design exploits a structural asymmetry: Dawid–Skene estimates annotator reliability from internal consensus, while our expert gold standard provides an independent external measure. When these two signals diverge — as they do substantially here — the gap reveals not noise but systematic bias. We use this dissociation to identify the mechanism, not merely the magnitude, of LLM annotation error in this domain.

Our contributions are:

1. **DS / Gold dissociation:** Dawid–Skene overestimates GPT-5.2 by 19.7 points and underestimates Claude Sonnet 4 by 6.8 points on valence, reversing model ranking.

2. **Aggregation is not the bottleneck:** Majority Vote, Dawid–Skene, and MACE yield near-identical consensus labels (98.6% three-way agreement; 493/500), implicating annotator bias rather than the aggregation method.

3. **Implicit affect dominance:** Using Stanza lemmatization and a 61-entry simplified FEEL lexicon as a conservative probe, 76% of segments lack overt affect markers, creating conditions for systematic over-neutralization.

4. **Expert gold standard:** 500-segment expert annotation with structured disagreement resolution ($\kappa = 0.682$–0.731), validated via Dawid–Skene aggregation of human-only annotations ($\kappa = 0.983$–0.991).

Proceedings of Computational Affective Science (CAS) @ LREC 2026, pages 125–135
16 May 2026.

2. Related Work

Emotion annotation in NLP has been shaped by shared tasks (Strapparava and Mihalcea, 2007) and lexical resources such as FEEL (Abdaoui et al., 2017) and FANCat (Syssau et al., 2021) for French — both derived from adult corpora, which raises immediate questions about applicability to child-produced language. Beyond resource availability, the field has increasingly recognized that emotion annotation is inherently subjective: disagreement between annotators often reflects genuine perceptual differences rather than error (Uma et al., 2021; Plank, 2022), a perspective known as *perspectivism*. This is especially true when affect is implicit — when no explicit emotion word anchors the judgment and annotators must rely on pragmatic inference. In such cases, disagreement concentrates precisely where the signal is weakest, as our results confirm. Related work has begun extending annotation to child-facing settings (Seo et al., 2024; Pérez-Espinosa et al., 2020), but the reliability of automated methods for this population remains underexplored.

LLMs have emerged as a promising alternative to crowd-sourced annotation: ChatGPT matches or exceeds crowd-worker quality on several text classification tasks (Gilardi et al., 2023), and GPT-4 reaches expert-level performance on political content coding (Törnberg, 2023). For emotion annotation specifically, LLMs often align well with human judgments but exhibit systematic perceptual biases, motivating hybrid human–LLM pipelines (Niu et al., 2024, 2025). A natural question is how to assess LLM annotation quality without exhaustive human validation — Dawid–Skene (Dawid and Skene, 1979) is typically used for this, modelling each annotator's confusion matrix via EM to recover latent true labels and estimate per-annotator reliability. Agreement metrics such as Fleiss' κ (Fleiss, 1971; Cohen, 1960), Krippendorff's α (Krippendorff, 2011), and their properties are discussed in (Artstein and Poesio, 2008). However, Dawid–Skene assumes conditional independence of annotator errors — an assumption that may fail when multiple prompts of the same model share architectural biases. Whether inter-LLM consensus is a reliable quality signal under these conditions, particularly when affect is largely implicit, is the question we address.

3. Data and Annotation Schema

3.1. Corpus

We use an anonymized corpus of transcribed individual interviews with French-speaking children (corpus and collection protocol available under CC-BY-4.0), collected in four primary schools in France between March 2021 and March 2022. Participants (N=184 children, ages 6–11, mean 9.2 years; 51% boys) were asked to describe the concepts FATHER and MOTHER. The corpus contains 2,106 segments: 1,053 for FATHER and 1,053 for MOTHER (mean length 8.7 words, SD 5.3).

3.2. Annotation Schema

We designed a two-level annotation schema capturing valence and intensity:

E1 – Valence and Intensity: `valence` ∈ {positive, negative, neutral, mixed}; `intensity` ∈ {0 (none), 1 (weak), 2 (moderate), 3 (strong)}; `evidence_span` citing the supporting text fragment; `negation_flag` for negation constructions.

This paper focuses primarily on `E1.valence` as the primary dimension, with `E1.intensity` as a secondary analysis demonstrating LLM limitations on ordinal annotation scales. All annotations include a self-reported confidence score (0.0–1.0).

4. Automated Annotation Methods

4.1. Lexicon baselines

FEEL Complete (14,126 entries; raw). We evaluate the full FEEL lexicon (Abdaoui et al., 2017) (derived from adult corpora) without modification.

FEEL Complete (filtered). To mitigate domain mismatch, we test an aggressively filtered variant (stopword removal, POS filtering, length constraints, and stricter mixed rules); results are reported in Section 6.1.

FEEL Simplified (61 entries). Finally, we use a manually curated 61-item subset of FEEL as a conservative proxy for overt affect (29 pos., 30 neg., 2 contextual), excluding frequent but context-dependent role terms (e.g., *travaille*, *ménage*, *chef*) (Appendix B).

4.1.1. Annotation procedure

All lexicon-based methods use Stanza (Qi et al., 2020) French lemmatization (`tokenize`, `mwt`, `pos`, `lemma` processors) to normalize word forms before lexicon lookup. Negation markers (*ne*, *pas*, *jamais*, *sans*) within a 3-token window preceding an affective lemma flip its valence. Final valence is determined by majority vote over matched lemmas, defaulting to neutral if no matches occur.

4.2. LLM-Based Annotation

We treat each LLM call as a noisy annotator. Two models, Claude Sonnet 4 (`claude-sonnet-4-20250514`, Anthropic) and GPT-5.2 (`gpt-5.2-2025-12-11`, OpenAI), annotated the full 2,106-segment corpus through their respective batch APIs. All runs used temperature 0, `max_tokens`=500, and structured JSON outputs constrained by a predefined schema.

Prompting setup. Each model was queried with three zero-shot prompts (P1–P3), yielding six pseudo-annotators in total (2 models $\times$ 3 prompts). The prompts shared the same system instruction and annotation schema, and differed only in annotation posture: conservative, implicit-affect sensitive, or lexically grounded. The full prompt templates are provided in Appendix A.

Implementation details. Claude Sonnet 4 was queried via the Anthropic Message Batches API on 2025-12-25, and GPT-5.2 via the OpenAI Batch API on 2025-12-23. For each segment, the user message contained the segment metadata and the JSON Schema specifying the fields used in this study (valence, intensity, evidence span, negation flag, and confidence).

4.3. Aggregation Methods

All aggregation methods (MV, DS, and MACE) are applied only to the six LLM pseudo-annotators; lexicon-based methods are evaluated separately as baselines.

Majority/Plurality Vote (MV). MV assigns the plurality label. In case of ties, we use a uniform distribution over tied labels, yielding confidence $1/n_{\text{tied}}$.

Dawid–Skene (DS). DS is a probabilistic aggregation model (Dawid and Skene, 1979) that estimates latent true labels and annotator-specific confusion matrices via EM, using a Dirichlet prior $\alpha = 1$. It assumes conditional independence of annotator errors, an assumption that may be violated when LLM annotators share correlated biases (Section 6.6).

Ordinal DS for E1 intensity ($y \in \{0, 1, 2, 3\}$). For ordinal intensity, we replace the standard categorical noise model with:

$$P(l = t \mid y = k, a) \propto \exp(-\beta_a |k - t|),$$

normalized over t, where β_a controls annotator a's tolerance to ordinal distance.

MACE. MACE is an EM-based aggregation model (Hovy et al., 2013) in which annotators either output the true label with competence θ_a or sample from a global spam distribution ξ with probability $1 - \theta_a$, thereby down-weighting unreliable annotators.

4.3.1. Robustness to the Aggregation Method

On the 500 gold segments, all three methods converge strongly (98.6% three-way agreement, 493/500) and yield near-identical performance (MV: 69.2%, DS: 69.0%, MACE: 69.2%; $\kappa \approx 0.475$). This makes the aggregation rule itself an unlikely source of error and instead points to the LLM annotator pool as the main source of systematic bias (Section 6.6). For brevity, the remaining analyses use DS, which also provides annotator-specific confusion matrices for the bias analysis in Section 6.6.

5. Human Gold Standard Construction

5.1. Stratified Sampling

To validate automated annotations, we constructed a human gold standard on 500 segments selected by stratified sampling from the full corpus, providing $\pm$3% confidence intervals at the 95% level.

Four complementary strata were designed to maximize validation informativeness (Table 1).

Stratum	%	n	Objective
DS Confidence	40	200	Calibrate DS scores vs. human[a]
Label-proportional	30	150	Corpus representativeness[b]
LLM Disagreements	20	100	Maximally informative cases
Demographic	10	50	Age/gender/concept robustness
Total	**100**	**500**	

[a] Equal sampling across three confidence levels: high (≥ 0.9), medium ($0.7 \leq x < 0.9$), and low (< 0.7).
[b] Sampling proportional to the LLM-predicted label distribution (positive 53.7%, neutral 40.8%, negative 4.4%, mixed 1.1%). Gold-standard results later confirmed that this distribution differs from the true one.

Table 1: Stratified sampling design (N=500)

Since stratification targeted *structural properties* (confidence zones, inter-LLM disagreement, demographic balance) rather than specific LLM predictions on individual instances, and since

the resulting gold converges with an independent Dawid–Skene aggregation of human-only annotations ($\kappa = 0.991$, Section 5.4), selection bias is minimal. Because this subset is not a simple random sample, results are reported by stratum where relevant, and global estimates are interpreted with caution.

5.2. Annotation Protocol

Three independent expert annotators with backgrounds in education and child development annotated all 500 segments, following a shared annotation manual with prototypical examples and evidence-span justifications. A 30-segment calibration pilot and a 50-segment quality checkpoint preceded the full annotation. Disagreements were resolved in structured discussion sessions; final labels were determined by majority vote (2/3 agreement). Segments were flagged with confidence levels (HIGH / MEDIUM / LOW) based on the degree of initial annotator agreement. The resulting gold is predominantly positive (68.4%), with neutral (26.0%), negative (4.6%), and mixed (1.0%) classes; full label distributions are reported in Table 11.

5.3. Inter-Annotator Agreement

Inter-annotator reliability on the 500-segment subset was assessed using Fleiss' κ over the three expert annotators. Agreement is substantial for valence ($\kappa = 0.731$) and moderate-to-substantial for intensity ($\kappa = 0.682$), consistent with the known difficulty of ordinal intensity scales (Conventions in (Artstein and Poesio, 2008)). Disagreements concentrate on short or contextually ambiguous segments.

Relative to the adjudicated (vote-based) gold, Annotator 3 shows lower agreement on valence ($\kappa = 0.719$) than the other two annotators, reflecting a more conservative interpretation of implicit affect. Annotator 3 was outvoted on 76 segments during disagreement resolution; accordingly, the gold reflects the majority (more liberal) boundary and may underestimate the positive/neutral boundary difficulty (Section 8.1).

5.4. Gold Standard Validation

To validate the collaborative resolution procedure, we applied Dawid–Skene independently to the three individual human annotations and compared the resulting labels to the vote-and-discussion gold.

The two procedures converge at 98.8–99.6% ($\kappa = 0.983$–0.991), confirming label equivalence and providing the baseline contrast for Section 6.6:

Dimension	Agreement	κ	Divergences
E1.valence	99.6%	0.991	2
E1.intensity	98.8%	0.983	6

Note. All 8 divergences fall on segments flagged as MEDIUM or LOW confidence in the vote-based gold. No divergence occurs on HIGH-confidence segments.

Table 2: Agreement between vote-and-discussion gold and Dawid–Skene human gold

applied to truly independent annotators, DS produces near-perfect alignment; applied to LLM prompts sharing architectural biases, it does not.

6. Results

6.1. Lexicon Baseline

Table 3 compares lexicon methods against the human gold.

Method	Acc.	κ	Gap
FEEL Simplified (61)	44.0%	0.162	-25.0 pp
FEEL Complete (raw)	46.2%	-0.032	-22.8 pp
FEEL Complete (filt.)	58.6%	0.023	-10.4 pp
LLM Aggregation	69.0%	0.475	baseline

Note. Gap = difference from the LLM aggregation baseline (69.0%).

Table 3: Lexicon baseline comparison for valence accuracy and κ on the human gold ($N = 500$)

FEEL Simplified is the stronger baseline despite its 61-entry size: its 76% zero-hit rate defaults to neutral, correctly handling implicitly affective text, while FEEL Complete (raw) over-predicts *mixed* (113 vs 5 gold) due to context-dependent lemmas (*papa [dad]*, *maman [mom]*) marked positive in adult corpora. Filtering recovers 12.4pp (+46.2%→58.6%) but the 10.4pp gap to LLM aggregation persists, confirming that lexicons cannot capture the 76% of segments expressing affect implicitly (see Appendix C for illustrative examples).

Implicit affect dominance. FEEL Simplified's 76% zero-hit rate (after Stanza lemmatization) quantifies the core challenge: three out of four transcribed segments about parents lack explicit emotion words yet convey clear sentiment. Appendix C provides illustrative examples of affect expression mechanisms that lexicons cannot capture: pragmatic implicature (*qui prend soin de nous* → positive), counterfactual exclusion (*si tu es un garçon ton papa te comprend mieux* → negative), and frequency qualifiers (*on n'est pas d'accord parfois* →

Method	Valence		Intensity	
	Acc.	κ	Acc.	κ
FEEL Simplified	44.0	0.162	–	–
Surface Rules	39.2	0.118	–	–
GPT-5.2	71.0	0.484	41.0	0.107
Claude Sonnet 4	79.2	0.606	55.2	0.370
Annotator 1	98.0	0.956	90.8	0.865
Annotator 2	96.0	0.914	87.2	0.819
Annotator 3	87.0	0.719	78.0	0.682

Note. Accuracy is the proportion of exact label matches. LLM scores are computed from Dawid–Skene aggregated labels over the three prompts for each model. Human annotator performance is evaluated against the vote-and-discussion gold label. Lexicon methods are not evaluated for intensity because their deterministic rules are not designed for ordinal scales.

Table 4: Annotation performance against the human gold standard ($N = 500$)

mixed). This finding empirically justifies LLM deployment for this domain.

6.2. LLM Internal Consistency

Both LLMs achieved high internal consistency across their three prompts (Krippendorff's $\alpha > 0.8$, Table 6): Claude Sonnet 4 mean $\alpha = 0.901$, GPT-5.2 mean $\alpha = 0.858$. GPT-5.2 shows 92.2% unanimous agreement across prompts; Claude Sonnet 4 shows 93.2%. These figures suggest high annotation quality. Section 6.6 shows they do not.

6.3. Performance Against Human Gold Standard

Table 4 reports annotation performance against the 500-segment human gold standard.

Several findings emerge. First, both LLMs substantially underperform human annotators on valence (14–27 percentage-point gap) and on intensity (23–50 point gap). Second, Claude Sonnet 4 outperforms GPT-5.2 on both dimensions, the reverse of Dawid-Skene estimates for valence (Section 6.6). Intensity results are analysed in Section 6.5.

Class-level performance. Table 5 reports per-class metrics, revealing asymmetric error patterns masked by overall accuracy.

GPT-5.2 achieves high precision on positive (0.969) but poor recall (0.632), reflecting systematic over-neutralization: 118 gold-positive segments incorrectly predicted neutral (34.5% of positive instances). Claude achieves better balance (precision=0.959, recall=0.749) with only 76 such errors (22.2%). Macro-F1 scores—which weight all classes equally—show Claude's advantage more clearly than raw accuracy: 0.660 vs 0.592 (+0.068). Both models struggle with the rare mixed class (F1=0.308–0.333, n=5), but Claude substantially outperforms GPT on negative (F1=0.766 vs 0.619), despite negative being only 4.6% of the corpus.

6.4. Aggregation Methods: Robustness

Three aggregation methods (Majority Vote, Dawid-Skene regularized, MACE) produce nearly identical results on the 500 gold segments, demonstrating that aggregation algorithm choice is not the performance bottleneck. Table 6 presents the comparison.

This near-perfect agreement across aggregation methods (98.6%, 493/500 segments) indicates that aggregation algorithm choice is not the main limitation in this setting. Although the three methods differ substantially in complexity, they produce functionally equivalent labels on the present corpus. The limiting factor is therefore not algorithmic sophistication, but shared annotator bias.

The three methods nevertheless serve different secondary purposes. When prioritising human review, MV is preferable because its vote margin flags 108/500 segments (21.6%) as low-confidence (< 0.7), directly identifying cases for adjudication. By contrast, DS flags 0/500 segments and MACE only 3/500, making both largely uninformative for uncertainty-based review on this corpus. When confidence scores are not required, MACE achieves the best Macro-F1 (0.640 vs. 0.604 for MV), suggesting slightly better handling of rare classes. DS, however, remains indispensable for annotator bias analysis because it is the only method that estimates annotator-specific confusion matrices. This capability makes it possible to compare internal reliability with external gold performance, which is the central analytical objective of this study (Section 6.6).

Given the goals of this study, we therefore adopt DS for the present corpus.

6.5. Intensity: Ordinal Silver vs Gold

Individual LLMs perform poorly on intensity annotation ($\kappa = 0.107$–0.370, exact accuracy 41.0–55.2%), far below human annotators ($\kappa = 0.682$–0.865). Ordinal DS aggregation substantially narrows this gap (Table 7), but reveals a systematic downward bias.

The ± 1 accuracy of 95.0% and quadratic $\kappa = 0.654$ confirm ordinal plausibility—most errors are off by one level—but the label distribution exposes a strong downward compression (Table 8): the sil-

Class	GPT-5.2			Claude Sonnet 4		
	P	R	F1	P	R	F1
positive ($n = 342$)	0.969	0.632	0.765	0.959	0.749	0.841
neutral ($n = 130$)	0.494	0.954	0.651	0.597	0.923	0.725
negative ($n = 23$)	0.684	0.565	0.619	0.750	0.783	0.766
mixed ($n = 5$)	0.286	0.400	0.333	0.250	0.400	0.308
Macro-F1		0.592			0.660	
Weighted-F1		0.724			0.802	

Note. P = Precision, R = Recall, and F1 = F1-score. Macro-F1 weights all classes equally, whereas Weighted-F1 weights by class frequency. Support (n) is shown in parentheses. LLM scores are computed from Dawid–Skene aggregated labels over the three prompts for each model.

Table 5: Per-class performance for valence ($N = 500$)

Method	Accuracy	κ	Macro-F1	Weighted-F1	Agree. w/ others
Majority Vote	69.2%	0.475	0.604	0.704	98.8%
DS	69.0%	0.475	0.618	0.703	99.0%
MACE	69.2%	0.476	0.640	0.705	99.2%
Three-way agreement: 98.6% (493/500 segments)					

Note. All methods aggregate 6 LLM pseudo-annotators (2 models $\times$ 3 prompts). "Agree. w/ others" = proportion of segments where this method agrees with at least one other method. Three-way agreement computed over segments where all three methods produce identical labels (493/500 = 98.6%).

Table 6: Aggregation method comparison for valence performance on 500 gold segments

Metric	Value
Exact accuracy	55.2%
$\pm$1-level accuracy	95.0%
MAE	0.498
κ (quadratic)	0.654
Spearman ρ	0.740

Table 7: Comparison of ordinal DS silver and human gold for intensity ($N = 500$)

ver over-predicts levels 0–1 and severely underpredicts levels 2–3 (recall $\approx 28.6\%$ for level 3).

Level	Gold	Silver
0 (none)	130	198
1 (weak)	128	182
2 (moderate)	207	110
3 (strong)	35	10

Table 8: Intensity label distributions for gold and ordinal DS silver

This silver is suitable for pre-annotation and coarse-grained trend analysis, but should not be used for fine-grained intensity statistics, particularly when high-intensity segments are the focus.

Model	DS est.	Human gold	Δ
GPT-5.2	90.7%	71.0%	-19.7 pp
Claude Sonnet 4	72.4%	79.2%	$+6.8$ pp

Note. DS estimate = Dawid–Skene diagonal accuracy from LLM consensus on the full corpus ($N = 2{,}106$). Human gold = accuracy against expert annotation ($N = 500$). Δ = human gold minus DS estimate, in percentage points.

Table 9: Dissociation between DS estimates and human gold for valence

6.6. The DS / Human Dissociation

Table 9 contrasts Dawid-Skene estimates with human gold performance.

The dissociation is asymmetric: Dawid-Skene overestimates GPT-5.2 by 19.7 points while underestimating Claude Sonnet 4 by 6.8 points. Two distinct mechanisms explain this pattern.

6.6.1. GPT-5.2: Bias collusion inflates DS confidence

GPT-5.2 achieves 92.2% unanimous prompt agreement. However, 27.3% of unanimously-agreed segments are incorrect: all three GPT prompts predict "neutral" for segments the human gold labels "positive" (n=100 segments). This sys-

tematic over-neutralization is not random error—it reflects a shared model bias. Dawid–Skene cannot distinguish genuine consensus from shared bias: it interprets the unanimity as high reliability and assigns elevated confidence scores to precisely the segments where GPT is wrong. The Dawid–Skene independence assumption is therefore violated not merely formally but in a way that compounds the most frequent error type. Corpus-level evidence confirms the mechanism: the human gold contains 68.4% positive segments, while GPT predicts only 44.6%—a 23.8-point neutralization deficit that three correlated prompts cannot self-correct.

Claude Sonnet 4: diversity penalized by correlated prompt agreement. Claude shows 82.0% unanimously correct segments, compared to 72.7% for GPT—indicating higher annotation precision when prompts agree. However, Dawid–Skene aggregation over the full 6-annotator pool (2 models $\times$ 3 prompts) assigns lower weight to Claude because Claude and GPT diverge on difficult segments, while the three GPT prompt variants remain highly consistent with each other. Under the conditional-independence assumption, DS interprets this correlated within-model agreement as higher annotator reliability and therefore downweights Claude whenever the two model blocks disagree. When evaluated directly against the human gold, Claude's majority-vote accuracy (79.2%) matches its individual evaluation, confirming that the DS underestimation reflects correlated-prompt structure rather than genuine annotation weakness.

Confidence-based filtering. Despite the miscalibration at corpus level, LLM internal confidence scores remain locally predictive of quality. Segments with confidence ≥ 0.7 achieve 92.6–94.9% accuracy (Table 10), providing a practical threshold for annotation pipelines: apply LLM annotation to the full corpus, retain high-confidence segments, and prioritize low-confidence segments for human review.

Confidence	GPT-5.2	Claude Sonnet 4
< 0.5	62.7% (n=67)	70.5% (n=44)
0.5–0.7	57.7% (n=253)	70.6% (n=279)
0.7–0.9	92.6% (n=176)	94.9% (n=175)
≥ 0.9	100% (n=4)	100% (n=2)

Note. Confidence is the mean self-reported score across three prompts. Accuracy is measured against the human gold standard ($N = 500$).

Table 10: LLM confidence vs. human gold accuracy

Method	Pos.	Neu.	Neg.	Mix.
Human Gold	68.4%	26.0%	4.6%	1.0%
FEEL Simplified	19.0%	77.0%	3.0%	1.0%
FEEL Complete (filt.)	80.6%	8.2%	7.0%	4.2%
LLM Aggregation	39.8%	53.0%	5.4%	1.8%
GPT-5.2	44.6%	50.2%	3.8%	1.4%
Claude	53.4%	40.2%	4.8%	1.6%

Note. LLM Aggregation corresponds to MACE silver ($N = 500$). DS and MV differ by less than 0.4 pp. FEEL Complete (filt.) over-predicts the positive class by 12.2 pp, whereas LLM Aggregation over-predicts the neutral class by 27.0 pp.

Table 11: Valence label distributions ($N = 500$ gold segments)

6.7. Complementary Bias Patterns: FEEL vs LLMs

FEEL and LLMs exhibit opposite systematic biases in label distribution, revealing fundamentally different annotation mechanisms. FEEL Complete over-detects positive (+12.2pp) through keyword sensitivity, while LLMs over-predict neutral (+27.0pp) through conservative bias on implicit affect. These opposite errors suggest potential value in hybrid pipelines (Section 8). Table 11 presents the full contrast.

7. Annotation Strategy for the Full Corpus

Our validation translates into four operational decisions for the remaining 1,606 segments. For valence, we apply MV silver labels and retain them as-is where vote confidence ≥ 0.7 — segments falling below this threshold across the full corpus (142/2,106) are submitted for human review, prioritising the positive/neutral boundary where both models concentrate their errors. Intensity is annotated manually for all segments: ordinal DS silver is suitable as pre-annotation but recall at levels 2–3 is too low ($\approx 28.6\%$ for level 3) to trust without human verification. Throughout, silver and human-reviewed labels are kept separate to preserve downstream flexibility.

Transferability caveat. The 0.7 threshold was estimated on the same 500 segments used for all comparisons. It is a working estimate requiring validation as human review of the remaining segments produces additional evidence. Researchers applying this framework to other corpora should construct a domain-specific gold sample before adopting these thresholds.

8. Discussion

Lexicons and LLMs fail in opposite directions — simplified FEEL achieves 95.8% precision on explicit affect but produces no signal for 76% of segments, while LLMs cover that gap at the cost of systematic over-neutralization. These complementary weaknesses point toward a hybrid pipeline: use the simplified lexicon for overt affect, apply LLMs to zero-hit segments, flag disagreements for expert review, and annotate intensity manually throughout.

Beyond the present corpus, our results call into question a common assumption in LLM annotation studies: that high inter-LLM agreement (or high DS confidence) reliably signals high annotation quality. This assumption can fail systematically when multiple prompts of the same model share architectural biases — a plausible scenario in practice, since prompt variation does not eliminate pretraining biases. In our experiments, the 19.7-point overestimation for GPT-5.2 is large enough to reverse model ranking and materially change conclusions about annotation quality. We therefore recommend that LLM annotation studies include at least one independent external human validation set, however small, to guard against shared biases invisible to internal metrics (Pangakis et al., 2023). Confidence calibration curves against ground truth should be reported alongside internal agreement metrics (Guo et al., 2017; Kadavath et al., 2022), and DS consensus is better treated as an upper bound on LLM quality than as a direct estimate, particularly when annotators are known or suspected to share biases (Paun et al., 2018).

8.1. Limitations

1. **Gold standard size.** 500 segments represent 23.7% of the corpus. Per-stratum performance should be interpreted with appropriate confidence intervals.

2. **Annotator heterogeneity.** Annotator 3 was outvoted on 76 segments during disagreement resolution, producing a gold that reflects two more liberal annotators' interpretation of implicit affect. Applications requiring conservative annotation of the positive/neutral boundary should treat the gold as an optimistic estimate.

3. **Stratification informed by LLM predictions.** The label-proportional stratum (30%) reflects the LLM-predicted distribution rather than the true distribution. Post-hoc comparison confirms systematic differences (gold: 68.4% positive vs. LLM-predicted 53.7%), meaning the gold sample slightly under-represents easy positive segments—a conservative bias for LLM evaluation.

4. **Dawid–Skene independence assumption.** Three prompts per model share architectural biases, violating conditional independence. We quantify rather than eliminate this limitation, and recommend alternative aggregation methods (e.g., item-response theory) for future work.

5. **Confidence threshold estimated on gold segments.** The 0.7 threshold was derived from the same 500 segments used for all comparisons. It is a working estimate requiring validation on held-out data.

6. **Zero-shot prompting.** The LLM annotation uses zero-shot prompts, consistent with realistic large-scale annotation practice. Whether few-shot or retrieval-augmented setups would reduce or merely displace the shared biases documented here remains an open question.

7. **Domain and language specificity.** Findings apply to French child language about family; generalization to other domains, languages, or age groups requires separate validation.

9. Conclusion

We set out to answer a practical question—which annotation strategy to use for a 2,106-segment corpus of French child language—and uncovered a methodological problem that extends beyond our corpus.

Three findings stand out. First, DS consensus is not a substitute for human validation: Dawid–Skene overestimates GPT-5.2 valence accuracy by 19.7 points due to shared neutralization bias across all three GPT prompts, while simultaneously underestimating Claude Sonnet 4 by 6.8 points. The gap reverses model ranking. Second, aggregation algorithm is not the bottleneck: Majority Vote, Dawid–Skene, and MACE converge at 98.6% agreement—better algorithms cannot compensate for biased annotators. Third, implicit affect dominance: 76% of transcribed segments require context-aware methods; lexicons alone are insufficient, but LLMs' conservative bias toward neutral precisely targets these implicit segments.

Practically, we adopt MV valence labels for the full corpus (vote confidence threshold 0.7 for human review prioritisation), DS for per-annotator bias analysis, and ordinal DS silver labels as pre-annotation for intensity with human review at levels 2–3. We recommend that any LLM annotation study include at least a small external human validation set, and treat Dawid–Skene confidence as an internal aggregation score rather than calibrated quality evidence.

The validated 500-segment gold standard and full 2,106-segment probabilistic corpus will be re-

leased to support future research on emotion expression in child language.

Ethics and Reproducibility

We use a publicly available anonymized corpus of interview transcripts with minors (Bellachhab et al., 2025). The data contain no personally identifiable information, and our annotations target text properties (affective content) rather than individual assessment. The corpus and collection protocol are released under CC-BY-4.0, which supports reproducibility.

10. Bibliographical References

Ron Artstein and Massimo Poesio. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34:555–596, 2008.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

A. Philip Dawid and Allan M. Skene. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society: Series C*, 28:20–28, 1979.

Joseph L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76:378–382, 1971.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. ChatGPT outperforms crowd-workers for text-annotation tasks. arXiv:2303.15056, 2023.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML 2017)*, pages 1321–1330. PMLR, 2017.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. Learning whom to trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130, Atlanta, Georgia, 2013. Association for Computational Linguistics.

Saurav Kadavath et al. Language models (mostly) know what they don't know. arXiv preprint arXiv:2207.05221, 2022.

Klaus Krippendorff. Computing Krippendorff's alpha-reliability. Departmental Papers (ASC), Annenberg School for Communication, University of Pennsylvania, 2011.

Chao Liu and Charis Chen. Text mining and sentiment analysis: A new lens to explore the emotion dynamics of mother–child interactions. *Social Development*, 33, 2024. https://doi.org/10.1111/sode.12733.

Minxue Niu, Mimansa Jaiswal, and Emily Mower Provost. From text to emotion: Unveiling the emotion annotation capabilities of LLMs. In *Proceedings of Interspeech 2024*, pages 2650–2654, 2024. https://doi.org/10.21437/Interspeech.2024-2282.

Minxue Niu, Yara El-Tawil, Amrit Romana, and Emily Mower Provost. Rethinking emotion annotations in the era of large language models. *IEEE Transactions on Affective Computing*, 16 (04):2668–2679, 2025.

Marissa Ogren and Catherine M. Sandhofer. Emotion words in early childhood: A language transcript analysis. *Cognitive Development*, 60: 101122, 2021. https://doi.org/10.1016/j.cogdev.2021.101122.

Nicholas Pangakis, Samuel Wolken, and Neil Fasching. Automated annotation with generative AI requires validation. arXiv preprint arXiv:2306.00176, 2023. https://doi.org/10.48550/arXiv.2306.00176.

Silviu Paun, Bob Carpenter, Jon Chamberlain, Dirk Hovy, Udo Kruschwitz, and Massimo Poesio. Comparing bayesian models of annotation. *Transactions of the Association for Computational Linguistics*, 6:571–585, 2018.

Barbara Plank. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, 2022. https://doi.org/10.18653/v1/2022.emnlp-main.731.

Peng Qi et al. Stanza: A python NLP toolkit for many human languages. In *Proceedings of ACL 2020: System Demonstrations*, pages 101–108, 2020.

Kenji Sagae, Eric Davis, Alon Lavie, Brian MacWhinney, and Shuly Wintner. Morphosyntactic annotation of CHILDES transcripts. *Journal of Child Language*, 37(3):705–729, 2010.

Woosuk Seo, Chanmo Yang, and Young-Ho Kim. Chacha: Leveraging large language models to prompt children to share their emotions about personal events. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–20. ACM, 2024. CHI '24.

Carlo Strapparava and Rada Mihalcea. Semeval-2007 task 14: Affective text. In *Proceedings of SemEval-2007*, pages 70–74, 2007.

Petter Törnberg. ChatGPT-4 outperforms experts and crowd workers in annotating political Twitter messages. arXiv:2304.06588, 2023.

Alexandra Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470, 2021. https://doi.org/10.1613/jair.1.12752.

11. Language Resource References

Amine Abdaoui, Jérôme Azé, Sandra Bringay, and Pascal Poncelet. FEEL: a french expanded emotion lexicon. *Language Resources and Evaluation*, 51(3):833–855, 2017.

Abdelhadi Bellachhab, Olga Galatanu, Frédéric Pugniere-Saavedra, Valérie Rochaix, Farida Saïd, and Jeanne Villaneau. Projet COSMOS : Transcriptions anonymisées d'interviews d'enfants sur la construction sémantique et les stéréotypes. HAL: hal-04893220, 2025.

Humberto Pérez-Espinosa, Juan Martínez-Miranda, Ismael Espinosa-Curiel, Josefina Rodríguez-Jacobo, Luis Villaseñor-Pineda, and Himer Avila-George. IESC-Child: An interactive emotional children's speech corpus. *Computer Speech & Language*, 59:55–74, 2020. https://doi.org/10.1016/j.csl.2019.06.006.

Arielle Syssau, Adil Yakhloufi, Edouard Giudicelli, Catherine Monnier, and Royce Anders. FANCat: French affective norms for ten emotional categories. *Behavior Research Methods*, 53:447–465, 2021.

A. Prompt Design and Schema

We used a shared system prompt combined with three posture variants in order to induce controlled diversity across annotations while preserving a common task definition. Each segment was submitted with the same structured user message and the same JSON schema constraints.

Base prompt (system, shared across P1–P3). You are an expert annotator for emotion analysis of children's short answers. You MUST output ONLY valid JSON that conforms to the provided schema. Important constraints: `evidence_span` fields MUST be exact substrings copied from `segment_text` (character-exact, including accents and punctuation). Do not invent evidence spans. If uncertain, use confidence in [0,1] accordingly and keep `evidence_span` minimal but valid. Do not add any extra keys.

Posture variants (appended to the base prompt).
P1 (conservative): *"Annotate carefully and conservatively. Prefer fewer labels if ambiguous."*
P2 (implicit): *"Annotate based on the child's expressed feeling/attitude in the text, even if implicit."*
P3 (lexical): *"Annotate with emphasis on explicit lexical cues; if none, keep confidence low."*

User message (per segment).

```
{
  "segment_id":   <int>,
  "child_id":     <int|null>,
  "concept":      "father|mother",
  "question":     "<question_type>",
  "segment_text": "<exact child
   ↪  utterance>"
}
```

JSON schema (E1 fields used in this paper).

```
E1.valence: "positive"|"negative"|"
 ↪  neutral"|"mixed"
E1.intensity: 0 | 1 | 2 | 3
E1.evidence_span: exact substring of
 ↪  segment_text
E1.negation_flag: true | false
confidence: 0.0 - 1.0  (root level)
```

B. Simplified FEEL Lexicon

Category	Lemmas (in French)
Positive (29)	*aide*, *aider*, *aimer*, *amour*, *apprécier*, *bien*, *bisou*, *bon*, *bonne*, *câlin*, *confiance*, *content*, *douce*, *doux*, *gentil*, *gentille*, *heureux*, *joie*, *meilleur*, *merveilleux*, *protection*, *protéger*, *raison*, *réconfort*, *réconforter*, *soutenir*, *soutien*, *super*, *tendresse*
Negative (30)	*abandon*, *abandonner*, *colère*, *difficile*, *douleur*, *dur*, *dure*, *détester*, *énervé*, *fâché*, *haine*, *haïr*, *inquiet*, *inquiétude*, *mal*, *mauvais*, *mauvaise*, *méchant*, *méchante*, *peur*, *peureux*, *pleurer*, *seul*, *souffrance*, *souffrir*, *sévère*, *tort*, *triste*, *tristesse*
Contextual (2)	*parfois*, *quelquefois*

Table 12: Simplified FEEL lexicon

Example	Gold valence	Why standard lexicons fail / How affect is expressed
qui prend soin de nous *[who takes care of us]*	positive	Implicit affect: no affective word appears, but "taking care" pragmatically implies trust and safety. FEEL Simplified yields 0 hits and therefore defaults to neutral.
si tu es un garçon ton papa te comprend mieux *[if you are a boy, your dad understands you better]* *(a girl about "Father")*	negative	The conditional implies exclusion: the child suggests that, as a girl, she is less understood. No explicit negative lexical cue is present.
on n'est pas d'accord parfois *[sometimes we disagree]*	mixed	The qualifier *parfois [sometimes]* weakens absolute disagreement and shifts interpretation toward a conditional or mixed valence. This requires pragmatic reasoning rather than lexical matching.
qui va au magasin *[who goes to the store]*	neutral	This is a purely descriptive action. FEEL Complete incorrectly treats *aller [to go]* and *magasin [store]* as positive, producing a false positive.

Table 13: Illustrative examples of affect expression in child language about parents

C. Examples of Affect Expression

Table 13 illustrates several cases in which affect is conveyed through implicature, syntax, and pragmatics rather than through explicit emotion words. The fourth example further shows that expanding the lexicon may introduce false positives.