

How Reliable Are LLM Annotators for Online Manipulation? Evidence from Phishing and Abuse Corpora

Senaid Popovic^{1,2} Yannick Parmentier¹ Fabien Lauer¹

Damien Riquet² Maxime Meyer²

¹Université de Lorraine, CNRS, LORIA, F-54000 Nancy, France

²Hornetsecurity

senaid.popovic@hornetsecurity.fr, damien.riquet@hornetsecurity.fr,
maxime.meyer@hornetsecurity.fr, fabien.lauer@loria.fr, yannick.parmentier@loria.fr

Abstract

In this paper, we evaluate an LLM-assisted process for annotating persuasion intent in the context of online manipulation. We focus on the DIDEKO taxonomy, a two-tier framework of explicit and implicit manipulation categories based on Speech Act Theory and Cialdini’s influence principles. Our question is not whether such annotations support a downstream classifier, but whether they are reliable enough to be used as labels in the first place. We measure reliability as Cohen’s κ between a model and human gold labels, and between models, evaluating GPT-4o, Claude Sonnet 4.5, and Llama 3.3-70B on a synthetic corpus. On DIDEKO, models agree more with one another than with human gold, suggesting shared annotation heuristics that diverge from human judgement. On an unlabeled real-world corpus, hierarchical prompting increases inter-model agreement for two of three model pairs, but this result establishes consistency rather than correctness. On a separate human-annotated real-world corpus, we evaluate Claude Sonnet 4.5, Gemma 3-27B, and Qwen 2.5-7B at both top-level and sub-label resolution. Claude and Gemma exceed human agreement at the top level, whereas Claude is the only model to exceed held-out human benchmarks across the full 23-label taxonomy; Gemma also does so when sparse sub-labels are excluded. These results tend to show that LLM annotation reliability depends on the corpus and label granularity, and that inter-model agreement cannot substitute for human validation. We release the prompts, code, and sub-label definitions.

1 Introduction

Online manipulation, including phishing, social engineering, disinformation, and extremist recruitment, relies on a common mechanism: exploiting psychological influence principles to bypass rational decision-making (Cialdini, 1993; Stajano and Wilson, 2011). Detecting these tactics needs anno-

tated corpora, but human annotation of persuasion intent is slow and inconsistent. We measure this with Cohen’s κ (Cohen, 1960), the standard chance-corrected agreement coefficient: it compares the agreement two annotators actually reach with the agreement they would reach by chance given their label frequencies, so $\kappa = 1$ is perfect agreement and $\kappa = 0$ is chance level. Values around 0.4 to 0.6 are usually read as moderate, though in computational linguistics the thresholds are contested and depend on the task (Artstein and Poesio, 2008). On the DIDEKO dataset (Popovic et al., 2026), trained annotators reach only $\kappa = 0.111$ on implicit intent categories, despite detailed guidelines and multiple annotation rounds.¹

LLMs are a cheaper alternative for annotation (Gilardi et al., 2023; Tan et al., 2024), but their use for online-abuse data depends first on a basic reliability question: do off-the-shelf models reproduce trained human judgments on a theoretically grounded persuasion-intent taxonomy, or do their errors form a shared bias that any downstream classifier would inherit? We study this question across synthetic and real-world email corpora, comparing model-vs-human agreement, inter-model agreement, and the effect of label granularity. We also test open-weight model families as API-based capability checks; we do not evaluate local inference or deployment (Appendix A).

The DIDEKO taxonomy (Popovic et al., 2026) operationalises Cialdini’s principles (Cialdini, 1993) and Speech Act Theory (Searle, 1969), separating explicit requests from the implicit mechanisms that make compliance more likely, building on the persuasion-principle framings developed for phishing (Ferreira and Teles, 2019) and social engineering (Gragg, 2003), alongside a wider literature

¹This low agreement may come from the difficulty of the task itself, from the taxonomy, or even from the guidelines; we do not try to separate them here, though Section 4 shows that revised guidelines raise implicit agreement.

on detecting manipulative language in disinformation (Rashkin et al., 2017) and extremism (Alatawi et al., 2021). Closest to our setting, Nahmias et al. (2024) also prompt an ensemble of LLMs to quantify persuasion principles in emails, but use the resulting vectors as features for a downstream detector rather than evaluating the annotations themselves against human judgment.

Our contributions are:

- A reproducible evaluation of LLM-assisted persuasion-intent annotation across synthetic and real-world email corpora, with a public release of prompts, code, and an annotation tool.
- Evidence that, on the synthetic DIDEKO corpus, models agree with each other more than with human gold, and evidence that this ordering reverses on a real-world corpus, which bounds how far the shared-bias reading can be generalised.
- A cross-corpus inter-model agreement analysis showing that hierarchical prompting increases consistency on unlabeled real-world phishing without establishing label correctness.
- A human-labeled real-world evaluation that reports the same 100-email corpus at both collapsed top-level and sub-label resolution, including a held-out benchmark for the model-vs-human comparison.

2 Annotation Process

Because the DIDEKO annotation protocol is central to interpreting our evaluation, we briefly summarize the published process and the guideline changes used in this study. Originally, to build DIDEKO (Popovic et al., 2026) annotators were asked to pick, for each email, among 23 fine-grained sub-labels. Human annotators disagreed heavily on that task (inter-annotator agreement, IAA, of $\kappa_{\text{imp}} = 0.111$ on implicit intent), which we read as a sign that the fine-grained single-choice format is too hard to annotate reliably. We therefore ask a set of yes/no questions instead: for each category, is it present or not? This gives boolean labels over the 9 DIDEKO categories (4 explicit, 5 implicit) and, optionally, their 23 sub-labels, so an email can carry several intents at once. Implicit categories need inference about psychological mech-

anism rather than recognition of an expressed request, and on DIDEKO they are the harder tier for annotators; Section 4 shows that this ordering is not fixed, since it reverses under revised guidelines.

All model calls use the OpenRouter API. Which models run on each corpus varies for cost reasons (for instance, a smaller Claude model on Nazario (Nazario, 2024)) and is stated with each experiment below. The open-weight models we test (Gemma 3-27B and Qwen 2.5-7B) use the same API, so they are capability checks rather than evidence of local deployment (Appendix A). Structured boolean output is enforced with DSPy (Khatab et al., 2023) for a consistent schema across models. No fine-tuning is used; every model runs as-is at temperature 0.0.

We compare two prompting strategies, flat and hierarchical. **Flat** classifies all 9 categories in a single call, as one multi-label decision. **Hierarchical** works in three steps, each a smaller labelling problem: explicit categories, then implicit categories (with a note that human IAA on these is $\kappa = 0.111$, to push toward conservative annotation), then sub-labels for the categories already marked positive. Both use single-pass prompting, with no retrieval, no multi-turn reasoning such as ReAct (Yao et al., 2023), and no few-shot examples; we leave these strategies to future work. Prompt templates are in Appendix B. F_1 scores are macro-averaged from per-email, per-category counts, the same unit used for κ .

All agreement coefficients are computed over per-email, per-category boolean decisions. The categories used for κ are the nine top-level DIDEKO taxonomy categories: four explicit-intent categories and five implicit-intent categories; at sub-label resolution, the same procedure is applied to the 23 taxonomy sub-labels. For κ , we compute Cohen’s κ separately for each category over the N emails and then average across categories, so every category contributes equally regardless of its prevalence; Krippendorff’s α is computed over the pooled set of (email, category) items when all three annotators are compared at once.

Table 4 in Appendix C summarises the three experiments, which differ in corpus, model set, and whether human labels exist; the model set is not held constant across them, for cost reasons noted above, so cross-experiment differences should not be read as model comparisons.

3 Evaluation on DIDEKO and Nazario

In this section, we describe evaluations done on two different corpora: a labeled synthetic corpus, for which we can measure agreement with gold-standard human labels, and an unlabeled real-world corpus, where we can only measure agreement between models. This corresponds to the first two rows of Table 4 (Appendix C). Throughout, we report Cohen’s κ over the four explicit categories (κ_{exp}), the five implicit categories (κ_{imp}), and all nine together (κ_{all}).

We use 60 LLM-generated spear-phishing emails (Nahmias et al., 2024) from the DIDEKO dataset (Popovic et al., 2026), with majority-vote gold labels from six human annotators. The human agreement figures we compare against ($\kappa_{\text{exp}} = 0.577$, $\kappa_{\text{imp}} = 0.111$, $\alpha = 0.407$) are those reported by Popovic et al. (2026) over the 169 emails with two or more annotators, under the original single-choice scheme; we do not recompute them for these 60 emails in the boolean label space, so they indicate the difficulty of the annotation task rather than a like-for-like ceiling for our setup. We also use 120 real-world, unlabeled emails from Nazario (Nazario, 2024) (2021–2024), where the Nazario experiment measures cross-corpus inter-model agreement, not label correctness. Claude Haiku 4.5 replaces Claude Sonnet 4.5 on Nazario for cost reasons.

Models vs. human gold. Table 1 reports κ and F_1 for GPT-4o, Claude Sonnet 4.5, and Llama 3.3-70B under both prompt strategies. Claude Sonnet 4.5 under flat prompting is the only run to exceed the human implicit baseline ($\kappa = 0.168 > 0.111$); GPT-4o leads on explicit intents ($\kappa = 0.238$ under flat prompting). Hierarchical prompting raises overall κ for all three models, mostly by improving explicit-intent agreement for Claude and Llama, but every model remains far below the human explicit baseline (0.577).

Inter-model agreement. This low agreement with human gold is not random disagreement. Pairwise inter-model κ ranges from 0.414 to 0.513, well above any model’s agreement with human gold (0.04–0.16): the models converge on shared heuristics that do not match the taxonomy, rather than failing independently. Model majority vote would therefore propagate shared biases rather than the intended categories. Section 4 revisits this on a different corpus, where the pattern does not hold.

Across the two corpora, inter-model agreement drops from $\kappa = 0.449$ on DIDEKO to $\kappa = 0.305$ on the Nazario emails under flat prompting, mostly on implicit intents ($0.513 \rightarrow 0.289$). Hierarchical prompting raises agreement for two of the three pairs (Table 2): GPT-4o vs. Llama from 0.318 to 0.491, and Claude (Haiku 4.5) vs. Llama from 0.254 to 0.508. The third pair, Claude vs. GPT-4o, drops slightly ($0.344 \rightarrow 0.328$). Splitting the task into steps seems to help most when the models start out far apart: the pair that does not improve is also the one that already agreed most under flat prompting. Since the Nazario emails have no gold labels, this tells us about consistency between models, not about whether either model is right.

4 Sub-Label Evaluation Against Human Annotations

We now turn to checking the process against independent human labels on real-world data.

4.1 Data

We build a 100-email corpus: 33 from Nazario (2021–2024), 33 from TExtPhish (Qachfar and Verma, 2024), and 34 legitimate Enron emails (Klimt and Yang, 2004) as a negative class, with revised sub-label guidelines adding positive/negative examples for each of the 23 sub-labels. Three annotators independently label all 100 emails at sub-label resolution with a dedicated tool (Appendix E); majority vote gives the gold standard. All three are domain experts in email security: one PhD student in NLP (an author of this paper), one senior researcher, and one further PhD student. All are fluent but non-native English speakers. Before starting, each annotator studied the revised guidelines and completed a short practice pass on fictitious generated emails that are not part of the corpus. The guidelines were revised from the original DIDEKO instructions by one of the authors, based on discussion and feedback with the annotators, adding one positive and one negative example per sub-label. Annotators did not see each other’s labels and disagreements were not discussed or adjudicated afterwards, so the agreement figures below are for independent judgements.

4.2 Results

For the top-level comparison below, both human and model labels are collapsed from the same sub-label annotations by disjunction over the children of each DIDEKO category.

Table 1: Model vs. human gold (DIDECO, 60 emails), flat vs. hierarchical prompting. † exceeds the human implicit baseline ($\kappa = 0.111$). Human $\kappa_{\text{all}} = 0.355$ is mean pairwise Cohen’s κ ; DIDECO also reports Krippendorff’s $\alpha = 0.407$.

Model	Strat.	κ_{exp}	κ_{imp}	κ_{all}	$F_{1,\text{exp}}$	$F_{1,\text{imp}}$
GPT-4o	Flat	0.238	0.078	0.149	0.486	0.415
	Hier.	0.235	0.103	0.161	0.471	0.352
Claude Sonnet 4.5	Flat	−0.049	0.168 [†]	0.072	0.443	0.606
	Hier.	0.199	0.101	0.145	0.440	0.533
Llama 3.3-70B	Flat	0.019	0.063	0.043	0.506	0.531
	Hier.	0.235	0.079	0.148	0.469	0.478
Human IAA	—	0.577	0.111	0.355	N/A	

Table 2: Inter-model κ , Nazario (120 emails), flat vs. hierarchical. Claude here is Haiku 4.5, not Sonnet 4.5.

Pair	Strat.	κ_{all}	κ_{exp}	κ_{imp}
Claude vs GPT-4o	Flat	0.344	0.309	0.371
	Hier.	0.328	0.398	0.273
Claude vs Llama	Flat	0.254	0.358	0.171
	Hier.	0.508	0.518	0.500
GPT-4o vs Llama	Flat	0.318	0.308	0.326
	Hier.	0.491	0.575	0.424

At sub-label resolution, pairwise human-human agreement ranges from $\kappa = 0.294$ to $\kappa = 0.389$ (average $\kappa = 0.349$), higher than DIDECO’s original implicit baseline ($\kappa_{\text{imp}} = 0.111$) but still only fair to moderate. Collapsing the same annotations to 9 top-level categories raises pairwise human agreement to $\kappa = 0.509$, showing that label granularity materially changes the evaluation. Krippendorff’s α across all three annotators is 0.604 overall ($\alpha_{\text{exp}} = 0.516$, $\alpha_{\text{imp}} = 0.655$). On this corpus, implicit agreement is actually higher than explicit, and the same ordering holds for pairwise κ in Table 3. This reverses DIDECO’s pattern of implicit being the harder tier, and suggests the revised guidelines and worked examples help most on the implicit categories, though fine-grained intent annotation remains hard for trained humans.

At the collapsed top level (Table 3), Gemma 3-27B and Claude Sonnet 4.5 exceed the pairwise human-human average ($\kappa = 0.689$ and 0.648 vs. 0.509), while Qwen 2.5-7B remains below it ($\kappa = 0.456$). At sub-label resolution, Claude remains strongest ($\kappa = 0.469$), Gemma is closer to the human average ($\kappa = 0.372$), and Qwen falls behind ($\kappa = 0.232$).

The model-vs-majority comparison is still not apples-to-apples on its own: an aggregate label is inherently easier to match than a single annotator’s

Table 3: Model agreement with three-annotator majority gold on the 100-email real-world corpus. Top-level labels are collapsed from the same 23 sub-label annotations. The human row is pairwise human-human agreement averaged over annotator pairs, not majority-vote agreement.

Comparison	Top-level κ			Sub-label κ			$F_{1,\text{sub}}$
	all	exp	imp	all	exp	imp	
Claude Sonnet 4.5	0.648	0.695	0.611	0.469	0.561	0.419	0.809
Gemma 3-27B	0.689	0.694	0.685	0.372	0.454	0.328	0.781
Qwen 2.5-7B	0.456	0.589	0.349	0.232	0.369	0.159	0.602
Human-human avg	0.509	0.458	0.550	0.349	0.329	0.359	0.693

judgment. To control for this, we hold out each annotator, build a two-person consensus from the other two under a strict (unanimous) and a loose (either annotator) rule, and measure the held-out annotator’s sub-label κ against that consensus. Held-out humans average $\kappa = 0.299$ under the unanimous rule and $\kappa = 0.382$ under the either-annotator rule. Claude’s sub-label result ($\kappa = 0.469$) exceeds both held-out averages, so its edge over the human benchmarks is not just an artifact of comparing against an easier, averaged target. The effect is specific to Claude, though: Gemma falls between the two rules and Qwen is below both, so this is not a general property of LLMs. The same control at top-level resolution gives $\kappa = 0.505$ (unanimous) and $\kappa = 0.503$ (either), close to the pairwise figure of 0.509, so the top-level conclusion that Claude and Gemma exceed human agreement does not depend on which of the two baselines is used.

This pattern reverses the DIDECO results, where models agree with each other more than with human gold. On this corpus, sub-label LLM-LLM agreement averages only $\kappa = 0.269$, below pairwise human agreement ($\kappa = 0.349$) and average model-vs-gold agreement ($\kappa = 0.357$).

Sparse sub-labels. Nine of the 23 sub-labels have fewer than five positive instances in the gold standard and two have none at all, so their κ is either unstable or, where a rater shows no variance, undefined; our implementation returns zero in that case, which pulls the macro-average down. Table 6 therefore repeats the comparison over the 14 sub-labels with at least five positives. Every value rises substantially, Claude from 0.469 to 0.632 and pairwise human agreement from 0.349 to 0.450, and the ordering between models and the human average is unchanged: Claude stays above it, Qwen below it, and LLM-LLM agreement stays below human-human. The held-out comparison does change, however. Recomputing the benchmark over the same 14 sub-labels gives $\kappa = 0.393$ (unanimous) and $\kappa = 0.493$ (either), and Gemma’s 0.585 now exceeds both, so the claim that only Claude clears the held-out benchmark holds at full taxonomy resolution but not once sparse sub-labels are removed. We keep the all-23 figures in Table 3 as the headline numbers because they cover the full taxonomy, and read Table 6 as a check on how much of the low absolute agreement is a prevalence artifact. Per-sub-label values are in Appendix D.

Agreement by source. Splitting the mixed corpus by source (Table 7) shows that model-vs-gold agreement exceeds inter-model agreement on Nazario (0.427 vs. 0.254), TExtPhish (0.510 vs. 0.391), and Enron (0.393 vs. 0.306). The comparison with inter-annotator agreement varies by source: models fall below annotators only on TExtPhish, where annotation is easiest. Human agreement also varies more across sources than model agreement, so corpus difficulty may contribute to the observed reversal; with only 33 or 34 emails per source, these estimates are uncertain.

Behaviour on legitimate mail. On the 34 legitimate Enron emails (Table 8), models over-fire on *call to action* and *trust establishment*. Claude and Gemma respectively flag the latter on 47.1% and 55.9% of legitimate emails, against a human rate of 17.6%; at sub-label resolution, ordinary signature blocks drive many *legitimacy creation* positives.

5 Conclusion

We evaluated an LLM-assisted process for persuasion-intent annotation across synthetic phishing, real-world phishing, and legitimate email. On DIDEKO, models agree more with one another

than with human gold, whereas this pattern does not hold on the real-world corpus. On the human-annotated corpus, reliability depends on label granularity: Claude Sonnet 4.5 and Gemma 3-27B exceed human agreement at the top level, but only Claude exceeds the held-out benchmark across the full taxonomy; Gemma also does so when sparse sub-labels are excluded. Models over-fire on trust-related categories in legitimate mail. These results support treating LLM annotations as candidate labels requiring human review, and reporting a held-out human baseline beside model-vs-majority-vote scores. Future work should test few-shot and agentic prompting, local inference, and downstream effects of shared biases.

We release the sub-label definitions, prompts, and evaluation code, including scripts that build the 100-email corpus.²

Acknowledgements

This work was carried out as part of a CIFRE doctoral partnership between Hornetsecurity and LORIA (Université de Lorraine, CNRS), supported by the Association Nationale de la Recherche et de la Technologie (ANRT). We thank the annotators for their work on the corpus described in Section 4, and the anonymous reviewers for their comments.

Limitations

Our limitations include the majority-vote gold standard over LLM-generated DIDEKO emails, the small DIDEKO sample, and varying model sets and Claude versions across corpora. Inference used OpenRouter rather than local hardware, and provider-side route changes remain a reproducibility risk (see Appendix C). The 100-email sub-label corpus has sparse categories and only three held-out comparisons. Expert annotators, one of whom revised the guidelines, provide an expert rather than crowdworker ceiling; all are non-native English speakers. Conservative prompting may depress implicit positives. We report point estimates, do not test few-shot or agentic prompting, and evaluate English only.

Ethics Statement

The public corpora may contain privacy-sensitive or reusable phishing text; DIDEKO is synthetic. We use all corpora for defensive research only.

²<https://github.com/Sneusneu/WOAH2026>

References

- Hind S. Alatawi, Areej M. Alhothali, and Kawthar M. Moria. 2021. [Detecting white supremacist hate speech using domain specific word embedding with deep learning and BERT](#). *IEEE Access*, 9:106363–106374.
- Ron Artstein and Massimo Poesio. 2008. [Survey article: Inter-coder agreement for computational linguistics](#). *Computational Linguistics*, 34(4):555–596.
- Robert B. Cialdini. 1993. *Influence: The Psychology of Persuasion*. Morrow.
- Jacob Cohen. 1960. [A coefficient of agreement for nominal scales](#). *Educational and Psychological Measurement*, 20(1):37–46.
- Ana Ferreira and Soraia Teles. 2019. [Persuasion: How phishing emails can influence users and bypass security measures](#). *International Journal of Human-Computer Studies*, 125:19–31.
- Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. [ChatGPT outperforms crowd-workers for text-annotation tasks](#). *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.
- David Gragg. 2003. [A multi-level defense against social engineering](#). SANS Institute Reading Room.
- Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. 2023. [DSPy: Compiling declarative language model calls into self-improving pipelines](#). *arXiv preprint arXiv:2310.03714*.
- Bryan Klimt and Yiming Yang. 2004. [The Enron corpus: A new dataset for email classification research](#). In *Proceedings of the 15th European Conference on Machine Learning (ECML 2004)*, pages 217–226. Springer.
- Daniel Nahmias, Gal Engelberg, Dan Klein, and Asaf Shabtai. 2024. [Prompted contextual vectors for spear-phishing detection](#). *arXiv preprint arXiv:2402.08309*.
- Jose Nazario. 2024. [Phishing corpus](#). <https://monkey.org/~jose/phishing/>. Accessed 2026.
- Senaid Popovic, Damien Riquet, Maxime Meyer, Fabien Lauer, and Yannick Parmentier. 2026. DIDECO: An annotated dataset for intent detection in digital communications. In *Proceedings of the 15th Biennial Conference on Language Resources and Evaluation (LREC)*, Palma de Mallorca, Spain. European Language Resources Association. <https://hal.science/hal-05616045v1>.
- Fatima Zahra Qachfar and Rakesh M. Verma. 2024. [All your LLMs belong to us: Experiments with a new extortion phishing dataset](#). In *Proceedings of the 38th IFIP Annual Conference on Data and Applications Security and Privacy (DBSec)*, pages 241–258, San Jose, CA, USA.
- Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. [Truth of varying shades: Analyzing language in fake news and political fact-checking](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2931–2937, Copenhagen, Denmark. Association for Computational Linguistics.
- John R. Searle. 1969. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press.
- Frank Stajano and Paul Wilson. 2011. [Understanding scam victims: Seven principles for systems security](#). *Communications of the ACM*, 54(3):70–75.
- Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Man-sooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. [Large language models for data annotation and synthesis: A survey](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 930–957, Miami, Florida, USA. Association for Computational Linguistics.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. [ReAct: Synergizing reasoning and acting in language models](#). In *The Eleventh International Conference on Learning Representations (ICLR)*, Kigali, Rwanda.

A API-Based Open-Weight Model Note

The Gemma 3-27B and Qwen 2.5-7B results in Table 3 are open-weight model-family capability checks, not local-deployment evidence. They were obtained through OpenRouter, the same API route used for the other models. A proper local study would rerun the same prompts on fixed hardware, report quantization, decoding settings, throughput, and cost, and compare local output against the API output reported here. Until then, these rows cannot support claims about privacy, hardware requirements, or local reproducibility.

B Prompt Templates

All prompts are implemented as DSPy (Khattab et al., 2023) signatures: a docstring giving task instructions, typed input fields, and one boolean output field per category with a natural-language description. The descriptions below are the literal field descriptions used at inference time.

B.1 Flat Strategy

System instruction: “Classify all explicit and implicit intents present in this spear-phishing email using the DIDEKO taxonomy. EXPLICIT intents (Speech Act Theory): mark True only when clear textual evidence supports the category. IMPLICIT intents (Ferreira & Teles 2019 framework): apply strict criteria. Human IAA on implicit categories is $\kappa = 0.111$, reflecting genuine difficulty. Only annotate when specific mechanism evidence is present. Surface-level professional language does NOT qualify.”

Input field: `email_text` — full email text to classify.

Output fields (one boolean per top-level category):

- `information_manipulation` (1.1) — direct requests for data or credentials, or instructions to manage system access
- `resource_acquisition` (1.2) — requests for money or bank details (direct), or gift cards, admin rights, or equipment (indirect)
- `trust_establishment` (1.3) — sender asserts identity or role, or invokes a known organisation, brand, or official process
- `call_to_action` (1.4) — an explicit immediate action instruction, or a preparatory step for a future action
- `authority` (2.1) — exploits hierarchical authority, uses credentials directly, or implies consequences for non-compliance
- `social_proof` (2.2) — herd mentality (“everyone did X”), diffusion of responsibility (“all employees”), or moral duty (compliance protects others)
- `liking_similarity_deception` (2.3) — general deception (fabricated scenarios), deceptive relationships (fake prior bond), or liking/similarity (shared traits exploited for rapport)
- `distraction` (2.4) — scarcity (time/resource urgency), overloading, strong affect (fear, alarm, excitement), or need and greed (exceptional reward appeal)

- `commitment_integrity_reciprocation` (2.5) — integrity signal (sender claims honesty), consistency (invokes past behaviour or role), commitment (obligation/coercion), or reciprocation (prior favour creates debt)

B.2 Hierarchical Strategy

Step 1: Explicit intents. System instruction: “Classify explicit intents in this email using the DIDEKO taxonomy, grounded in Speech Act Theory (Searle, 1969). Explicit intents are directly expressed communicative goals. Mark True only when clear textual evidence supports the category.” Input fields: `definitions` (the full explicit-intent definition block, including criteria and worked examples for each sub-label) and `email_text`. Output fields are the same four explicit categories as in the flat strategy (`information_manipulation`, `resource_acquisition`, `trust_establishment`, `call_to_action`), with identical field descriptions.

Step 2: Implicit intents. System instruction: “Classify implicit intents in this email using the DIDEKO taxonomy (Ferreira & Teles 2019 framework, synthesising Cialdini 1993, Gragg 2003, Stajano & Wilson 2011). Implicit intents are psychological persuasion mechanisms that influence behaviour indirectly, without stating the manipulation explicitly. STRICT ANNOTATION: human IAA on these categories was $\kappa = 0.111$, reflecting genuine difficulty. Apply conservative criteria, only annotate when specific mechanism evidence is present. Surface-level professional language does NOT qualify.” Input fields: `definitions` (the full implicit-intent definition block) and `email_text`. Output fields are the same five implicit categories as in the flat strategy, with identical descriptions.

Step 3: Sub-labels. System instruction: “Assign fine-grained sub-labels for the active high-level categories. Only classify sub-labels belonging to categories already confirmed True above.” Input fields: `rules` (sub-label assignment rules per category), `email_text`, and `active_categories` (a comma-separated list of the top-level categories confirmed True in Steps 1–2). Output fields are 23 booleans, one per sub-label, each described by a short phrase, for example:

scarcity_time (“deadline, urgency marker, time limit”), deceptive_relationships (“fake prior relationship, fabricated referral, false familiarity”), or reciprocation (“prior favour/service creates felt obligation to return it”).

B.3 Sub-Label Strategy

The evaluation in Section 4 uses a separate single-pass signature over all 23 sub-labels. Unlike the two strategies above, its instruction does not describe the input as a phishing email, since the corpus mixes phishing and legitimate mail. System instruction: “Annotate this email using the DIDEKO persuasion intent taxonomy. Answer True or False for each sub-label based on textual evidence only. EXPLICIT INTENTS (Speech Act Theory — directly expressed communicative goals): mark True only when clear textual evidence supports the sub-label. Default to False if uncertain. IMPLICIT INTENTS (Cialdini 1993 via Ferreira & Teles 2019 — psychological persuasion): apply STRICT criteria. Only annotate when specific mechanism evidence is present. Surface-level professional language does NOT qualify. Human IAA on implicit categories is $\kappa = 0.111$ — be conservative.”

Each of the 23 output fields carries a description with a positive and a negative example, mirroring the revised annotation guidelines. For instance, `information_gathering` is described as “1.1.1 — Direct or indirect requests for data, credentials, or identity confirmation. YES: ‘Can you confirm your banking details?’, ‘Click here to verify your identity’. NO: mentioning information in passing, providing info without requesting it.” The full set of 23 descriptions is in the released code.

B.4 Worked Example

The following illustrates how both strategies annotate the same email:

“Your account has been locked due to suspicious activity. Click here to verify your identity immediately to prevent permanent suspension.”

Expected explicit labels: `information_manipulation` (identity verification request), `call_to_action` (click here immediately). Expected implicit labels: `authority` (institutional tone, consequence of non-compliance), `distraction` (time scarcity: “immediately”; strong affect: fear of permanent suspension).

Table 4: Overview of the three experiments. The model set differs across corpora for cost reasons, so numbers are comparable within a row, not across rows.

Corpus	Models	Human labels	What is measured
DIDEKO, 60 synthetic emails	GPT-4o, Claude Sonnet 4.5, Llama 3.3-70B	Yes, 6 annotators, majority vote	Model vs. gold, and model vs. model, at top level
Nazario, 120 real emails	GPT-4o, Claude Haiku 4.5, Llama 3.3-70B	No	Model vs. model only: consistency, not correctness
Mixed, 100 real emails (Nazario, TExtPhish, Enron)	Claude Sonnet 4.5, Gemma 3-27B, Qwen 2.5-7B	Yes, 3 annotators, majority vote	Model vs. gold at top level and sub-label level, against a held-out human benchmark

C Inference Settings

All calls go through OpenRouter with DSPy-enforced structured boolean output, temperature 0.0, and no retries on parse failure; runs with a parse failure are discarded and re-issued. Table 5 lists the routes. [†]The route for the DIDEKO Claude run was not logged at the time; we report the identifier used by the sub-label run, which the output filenames and the shared model configuration indicate is the same, but we could not verify it after the fact. The DIDEKO hierarchical runs were issued on 2026-05-29 and the flat runs on 2026-06-01, so the two strategies in Table 1 were not run in the same session.

Table 5: OpenRouter routes per experiment. DIDEKO and Nazario were run on 2026-06-01, the mixed corpus on 2026-06-09.

Model	Route (OpenRouter)
<i>DIDEKO, 60 emails</i>	
GPT-4o	openai/gpt-4o
Claude Sonnet 4.5	anthropic/claude-sonnet-4-5 [†]
Llama 3.3-70B	meta-llama/llama-3.3-70b-instruct
<i>Nazario, 120 emails</i>	
GPT-4o	openai/gpt-4o
Claude Haiku 4.5	anthropic/claude-haiku-4-5
Llama 3.3-70B	meta-llama/llama-3.3-70b-instruct
<i>Mixed corpus, 100 emails</i>	
Claude Sonnet 4.5	anthropic/claude-sonnet-4-5
Gemma 3-27B	google/gemma-3-27b-it
Qwen 2.5-7B	qwen/qwen-2.5-7b-instruct

D Per-Sub-Label Agreement

Table 9 gives κ for each of the 23 sub-labels against the three-annotator majority gold, together with gold prevalence. The nine sub-labels marked * fall below five positive instances; two of them have none, so no model comparison on them is informative. Agreement is high on frequent, lexically

Table 6: Sub-label κ over all 23 sub-labels and over the 14 with at least five positive instances in the gold standard. Nine sparse sub-labels contribute values that are near zero or undefined, so the all-23 column understates agreement. The ordering between the models and the human average is unchanged, but the held-out benchmark is not: recomputed over the 14 labels it rises to 0.393 (unanimous) and 0.493 (either), which Gemma also clears.

Comparison	all 23	14 labels
Claude Sonnet 4.5	0.469	0.632
Gemma 3-27B	0.372	0.585
Qwen 2.5-7B	0.232	0.351
<i>Human-human avg</i>	0.349	0.450
<i>LLM-LLM avg</i>	0.269	0.410

Table 7: Sub-label κ within each source corpus, over the 14 sub-labels with adequate prevalence. Inter-model agreement is below agreement with the gold in all three sources; it falls below inter-annotator agreement only on TExtPhish.

Source	n	hum.-hum.	LLM-LLM	avg vs. gold
Nazario	33	0.225	0.254	0.427
TExtPhish	33	0.539	0.391	0.510
Enron	34	0.280	0.306	0.393

marked sub-labels such as *direct financial gain*, *scarcity* and *strong affect*, and low on *overloading* and *information control*, where annotators also disagree.

E Annotation Tool

Human annotations were collected with a purpose-built command-line tool (implemented in Go) that presents one email at a time alongside all 23 sub-label questions, grouped by their parent category. For each sub-label, the tool shows the category definition and one positive and one negative example drawn from the revised guidelines. Annotators record a binary judgement (present or absent) for each sub-label; top-level category labels are derived automatically by disjunction over their sub-labels and are not entered directly. Sessions are saved incrementally and can be resumed, and each annotator’s responses are stored independently to avoid cross-contamination before majority voting. The tool, the guideline text shown to annotators, and the underlying email corpus are all included in the data release described in Section 4.

Table 8: Positive rate (%) on the 34 legitimate Enron emails, human majority gold vs. each model, at top-level resolution, with the false-positive rate on legitimate vs. phishing emails in parentheses for the two stronger models. Over-firing on legitimate mail is concentrated in *trust establishment* and *call to action*.

Category	gold	Claude	Gemma	Qwen
information manipulation	20.6	8.8 (0/8)	23.5 (15/45)	2.9
resource acquisition	2.9	5.9 (3/19)	2.9 (0/0)	0.0
trust establishment	17.6	47.1 (36/35)	55.9 (50/0)	20.6
call to action	44.1	70.6 (47/20)	50.0 (11/0)	32.4
authority	11.8	11.8 (0/12)	17.6 (7/4)	14.7
social proof	5.9	17.6 (19/0)	5.9 (6/2)	5.9
liking, similarity, deception	2.9	0.0 (0/56)	0.0 (0/11)	0.0
distraction	29.4	29.4 (12/8)	35.3 (12/17)	8.8
commitment, integrity, recip.	26.5	14.7 (8/6)	14.7 (0/0)	0.0

Table 9: Per-sub-label κ against the three-annotator majority gold, with gold prevalence out of 100 emails. * marks the nine sub-labels below the prevalence threshold of five; “-” marks a comparison in which no variance is present and κ is undefined. The human column is the mean over annotator pairs.

Sub-label	prev.	Claude	Gemma	Qwen	human
<i>Explicit</i>					
information gathering	19	0.647	0.319	0.233	0.265
information control	8	0.457	-0.056	0.174	0.123
direct financial gain	34	0.935	0.978	0.864	0.880
indirect resource acquisition*	1	0.273	0.000	0.000	0.163
identity establishment*	3	0.147	0.370	0.492	0.538
legitimacy creation	34	0.680	0.644	0.526	0.104
direct action requests	72	0.788	0.844	0.660	0.461
indirect action facilitation	12	0.559	0.530	0.000	0.102
<i>Implicit</i>					
authority	44	0.774	0.689	0.110	0.534
herd mentality*	0	-	0.000	-	-0.008
diffusion of responsibility*	2	-0.031	0.000	0.000	0.140
moral duty	9	0.185	0.497	0.430	0.318
general deception	48	0.801	0.717	0.553	0.660
deceptive relationships*	2	0.310	-0.029	0.000	0.092
liking and similarity*	1	0.000	0.000	0.000	0.045
scarcity	45	0.840	0.838	0.623	0.696
overloading	24	0.178	0.240	0.000	0.363
strong affect	57	0.899	0.820	0.506	0.714
need and greed*	4	0.740	0.022	-0.074	0.702
integrity*	0	0.000	0.000	-	-0.022
consistency	5	0.322	0.322	0.000	0.362
commitment	53	0.782	0.802	0.234	0.723
reciprocation*	1	0.492	0.000	0.000	0.070