Advanced Certification in AI in Tax Law (France)

Tax Law and Natural Language Processing

Artificial Intelligence: The field of computer science that creates systems capable of performing tasks that normally require human intelligence. Related terms: machine learning, deep learning. Example: AI models that predict tax audit risk. Challenges: bias in training data, regulatory compliance.

Algorithmic Transparency: The degree to which the inner workings of an algorithm are understandable to stakeholders. Related terms: explainability, black-box. Example: Publishing the logic used to flag potentially abusive tax arrangements. Challenges: protecting proprietary code while meeting disclosure obligations.

Anti-Abuse Rule: Legal provisions designed to prevent the misuse of tax benefits. Related terms: general anti-avoidance rule (GAAR), specific anti-avoidance rule. Example: Applying the French anti-abuse rule to a hybrid entity structure. Challenges: interpreting intent, balancing legitimate planning with deterrence.

API (Application Programming Interface): A set of protocols that allows software components to communicate. Related terms: REST, SOAP. Example: An API that feeds transaction data from ERP systems into a tax-compliance NLP engine. Challenges: versioning, security, data privacy.

Article 54 of the CGI: French tax code article governing the taxation of dividends. Related terms: Corporate Income Tax, double taxation. Example: Using NLP to extract dividend-related clauses from shareholder agreements. Challenges: interpreting nuanced language across jurisdictions.

Attenuation Principle: The concept that tax benefits should diminish when the underlying transaction loses economic substance. Related terms: substance over form, economic reality. Example: NLP models flagging contracts where the declared purpose diverges from the economic effect. Challenges: quantifying attenuation, data scarcity.

Automated Tax Decision Engine: Software that autonomously determines tax outcomes based on predefined rules. Related terms: rule-based system, decision tree. Example: An engine that classifies invoices as taxable or exempt using NLP-extracted line items. Challenges: maintaining rule sets, handling exceptions.

Base Erosion and Profit Shifting (BEPS): OECD initiative to curb tax avoidance strategies that exploit gaps in tax rules. Related terms: transfer pricing, IP-based profit shifting. Example: NLP analysis of intercompany agreements to detect BEPS-related language. Challenges: multilingual documents, evolving guidance.

Beneficial Ownership: The natural person who ultimately owns or controls a legal entity. Related terms: ultimate parent, control. Example: Using entity-resolution NLP to link shareholders across filings. Challenges: incomplete data, privacy restrictions.

Big Data: Extremely large datasets that require advanced processing techniques. Related terms: data lake, distributed computing. Example: Analyzing millions of VAT invoices with NLP to uncover compliance trends. Challenges: storage costs, data quality.

Binding Corporate Rules (BCR): Internal policies that allow multinational companies to transfer personal data within the group. Related terms: GDPR, data protection. Example: Ensuring that NLP-driven tax analytics respect BCR obligations. Challenges: cross-border data flows, auditability.

Blind Spot (NLP): Areas where a language model lacks coverage or accuracy. Related terms: out-of-vocabulary, domain drift. Example: Missing legal jargon specific to French tax statutes. Challenges: continuous model retraining, domain-specific corpora.

Blockchain Taxation: The application of tax rules to transactions recorded on distributed ledgers. Related terms: crypto-assets, ledger analytics. Example: NLP parsing of smart-contract code to determine taxable events. Challenges: anonymity, rapid regulatory change.

Business Taxonomy: Structured classification of business activities for tax purposes. Related terms: NACE, NAF. Example: Training an NLP classifier to assign NAF codes from company descriptions. Challenges: ambiguous language, evolving industry terms.

Capital Gains Tax (CGT): Tax on the profit realized from the sale of assets. Related terms: realisation, exemption. Example: NLP extraction of asset disposal dates from contracts to compute CGT liability. Challenges: multi-jurisdictional rules, timing nuances.

Chatbot Compliance Assistant: Conversational AI that helps taxpayers understand obligations. Related terms: virtual assistant, dialogue system. Example: A French-language chatbot that answers queries on VAT filing deadlines. Challenges: accurate legal content, language nuance.

Classification Accuracy: Metric measuring the proportion of correct predictions in a categorisation task. Related terms: precision, recall. Example: Evaluating an NLP model that classifies documents as “tax-relevant” vs. “non-relevant.” Challenges: class imbalance, overfitting.

Clustering (NLP): Grouping similar textual items without predefined labels. Related terms: unsupervised learning, topic modeling. Example: Clustering tax rulings to discover emerging interpretative trends. Challenges: determining optimal cluster number, interpretability.

Code of Ethics (AI): Guidelines governing responsible AI development and deployment. Related terms: fairness, accountability. Example: Ensuring that tax-risk models do not discriminate against small enterprises. Challenges: translating abstract principles into concrete checks.

Compliance Monitoring: Ongoing observation of activities to ensure adherence to regulations. Related terms: audit trail, risk scoring. Example: Real-time NLP scanning of invoices for missing tax identifiers. Challenges: false positives, system latency.

Corporate Income Tax (CIT): Tax levied on the profits of companies. Related terms: tax base, effective tax rate. Example: Using NLP to extract deductible expenses from financial statements. Challenges: differing fiscal years, statutory limits.

Cross-Border Taxation: Tax rules applicable to transactions involving multiple jurisdictions. Related terms:

double taxation treaty, withholding tax. Example: NLP analysis of cross-border service contracts to identify treaty benefits. Challenges: divergent legal language, treaty hierarchy.

Data Anonymization: Process of removing personally identifiable information from datasets. Related terms: pseudonymisation, privacy-by-design. Example: Stripping taxpayer IDs from a corpus used to train a tax-compliance model. Challenges: re-identification risk, utility loss.

Data Governance: Framework for managing data availability, usability, integrity, and security. Related terms: data stewardship, metadata. Example: Defining policies for storing extracted tax clause metadata. Challenges: cross-department alignment, regulatory updates.

Data Pipeline: Sequence of processes that move data from source to destination. Related terms: ETL, stream processing. Example: Ingesting raw PDF rulings, applying OCR, then NLP tagging. Challenges: error propagation, scaling.

Data Quality: The condition of data based on accuracy, completeness, and consistency. Related terms: data cleansing, validation. Example: Verifying that extracted VAT rates match official tables. Challenges: manual correction costs, source heterogeneity.

Data Lake: Central repository that stores raw data in its native format. Related terms: schema-on-read, big data. Example: Housing all tax-related documents for downstream NLP analysis. Challenges: governance, searchability.

Data Minimisation: Principle of collecting only data necessary for a specific purpose. Related terms: GDPR, purpose limitation. Example: Limiting the NLP model to extract only tax-relevant entities, not full text. Challenges: balancing model performance with privacy.

Deep Learning: Subset of machine learning using neural networks with many layers. Related terms: transformer, convolutional network. Example: Fine-tuning a BERT model on French tax rulings. Challenges: computational cost, interpretability.

Denial-of-Service (DoS) Attack: Malicious attempt to disrupt services by overwhelming resources. Related terms: cybersecurity, rate limiting. Example: Protecting an online tax-question answering portal from DoS attacks. Challenges: maintaining availability while allowing legitimate traffic.

Digital Signature: Cryptographic method to verify authenticity of electronic documents. Related terms: PKI, non-repudiation. Example: Signing XML tax filings generated by an NLP engine. Challenges: key management, cross-border recognition.

Document Classification: Process of assigning categories to texts. Related terms: text categorisation, supervised learning. Example: Classifying tax notices as “assessment,” “reminder,” or “notice of appeal.” Challenges: overlapping categories, evolving terminology.

Domain Adaptation: Technique for transferring a model trained on one domain to another. Related terms: transfer learning, fine-tuning. Example: Adapting a general-purpose French language model to the niche

vocabulary of French tax law. Challenges: limited domain data, catastrophic forgetting.

Entity Recognition (NER): Identifying and classifying named entities in text. Related terms: tokenisation, span extraction. Example: Extracting VAT numbers, dates, and monetary amounts from invoices. Challenges: ambiguous entity boundaries, multilingual entities.

Entity Resolution: Matching records that refer to the same real-world entity. Related terms: deduplication, record linkage. Example: Linking a taxpayer's different filings across years despite variations in naming. Challenges: false matches, scalability.

European Union VAT Directive: Directive governing the common VAT system across EU member states. Related terms: VAT OSS, intra-EU supply. Example: NLP parsing of cross-border e-commerce contracts to apply the correct VAT rate. Challenges: divergent national implementations, frequent amendments.

Exhaustive Search: Technique that evaluates every possible solution to find the optimal one. Related terms: brute force, combinatorial optimisation. Example: Searching all possible tax-treatment combinations for a complex restructuring. Challenges: computational infeasibility for large problem spaces.

Explainable AI (XAI): Methods that make AI decisions understandable to humans. Related terms: interpretability, model-agnostic. Example: Providing a rationale for why an NLP model flagged a clause as "potentially abusive." Challenges: balancing fidelity with simplicity, regulatory acceptance.

Fact-Finding Interview: Structured dialogue to gather factual information. Related terms: questionnaire, knowledge elicitation. Example: An AI-driven interview that collects transaction details for tax compliance. Challenges: ensuring completeness, avoiding leading questions.

FATCA (Foreign Account Tax Compliance Act): US law requiring foreign financial institutions to report on US account holders. Related terms: CRS, information exchange. Example: NLP extraction of account holder names from bank statements for FATCA reporting. Challenges: data volume, cross-border privacy.

Feature Engineering: Process of creating input variables for machine learning models. Related terms: feature extraction, dimensionality reduction. Example: Deriving "sentence length" and "presence of tax keywords" as features for a classification model. Challenges: domain expertise needed, over-fitting risk.

Fiscal Year: The 12-month period used for accounting and tax reporting. Related terms: financial year, tax period. Example: Aligning NLP-derived transaction dates with the appropriate fiscal year for CIT calculations. Challenges: differing start dates across entities.

FON (Fiscal Object Number): French identifier for the object of a tax assessment. Related terms: tax reference, assessment number. Example: Extracting FON values from scanned assessment letters using OCR-combined NLP. Challenges: varied document layouts, OCR errors.

Forward-Looking Tax Risk Model: Predictive system that anticipates future tax compliance issues. Related terms: risk scoring, scenario analysis. Example: Using NLP-derived contract clauses to forecast audit likelihood. Challenges: data lag, uncertainty quantification.

GAAR (General Anti-Avoidance Rule): Broad rule that allows tax authorities to ignore transactions lacking economic substance. Related terms: anti-abuse rule, substance over form. Example: NLP detecting “substance-less” language in intercompany financing agreements. Challenges: subjective interpretation, litigation risk.

Graph Neural Network (GNN): Neural architecture that processes data represented as graphs. Related terms: node embedding, message passing. Example: Modelling relationships between entities (taxpayer, subsidiary, jurisdiction) to detect circular structures. Challenges: data sparsity, explainability.

Hybrid Tax Model: Combination of rule-based and machine-learning components. Related terms: ensemble, knowledge-driven AI. Example: A system that first applies deterministic VAT rules, then uses NLP to resolve ambiguous cases. Challenges: integration complexity, maintenance overhead.

Information Retrieval (IR): Process of obtaining relevant documents from a large collection. Related terms: search engine, ranking. Example: An IR system that returns relevant tax rulings based on a user query. Challenges: query ambiguity, relevance feedback.

Inference Engine: Component that applies logical rules to derive conclusions. Related terms: expert system, forward chaining. Example: An inference engine that combines extracted tax clauses to determine overall liability. Challenges: rule explosion, conflict resolution.

Input Normalisation: Standardising raw data into a consistent format before processing. Related terms: pre-processing, tokenisation. Example: Converting varied date formats in contracts to ISO-8601 before NLP analysis. Challenges: handling exceptions, locale variations.

International Tax Treaty: Agreement between two countries to avoid double taxation. Related terms: tax convention, Article 12. Example: NLP extraction of treaty-benefit clauses from bilateral agreements. Challenges: treaty hierarchy, language differences.

Knowledge Graph: Structured representation of entities and their relationships. Related terms: semantic network, ontology. Example: Building a graph linking tax concepts, statutes, and case law for query answering. Challenges: ontology alignment, data freshness.

Legislative Text Mining: Applying NLP techniques to statutes, regulations, and official bulletins. Related terms: text analytics, semantic parsing. Example: Mining the French Tax Code to identify changes affecting digital services. Challenges: legal drafting style, amendment tracking.

Legal Ontology: Formal specification of concepts and relationships in law. Related terms: taxonomy, semantic web. Example: Defining entities such as “taxable event,” “exemption,” and “penalty” for NLP annotation. Challenges: consensus building, extensibility.

Lexical Ambiguity: Situation where a word has multiple meanings. Related terms: polysemy, homonymy. Example: The term “tax” could refer to a levy or to a tax-related document. Challenges: disambiguation requires context, especially in short snippets.

Linear Regression: Statistical method for modelling the relationship between a dependent variable and one or more independents. Related terms: predictive modeling, least squares. Example: Estimating the impact of a tax rate change on revenue using historic data. Challenges: assumptions of linearity, outlier sensitivity.

Machine Translation (MT): Automatic conversion of text from one language to another. Related terms: neural MT, post-editing. Example: Translating German tax rulings into French for comparative analysis. Challenges: preserving legal nuance, domain-specific terminology.

Metamodel: Model that defines the structure of other models. Related terms: meta-ontology, schema. Example: A metamodel describing how tax-related entities should be represented in a knowledge graph. Challenges: ensuring flexibility without loss of coherence.

Model Drift: Degradation of model performance over time due to changes in data distribution. Related terms: concept drift, retraining. Example: An NLP model trained on 2020 tax forms misclassifying 2024 forms after regulatory updates. Challenges: monitoring, timely retraining.

Monte Carlo Simulation: Technique that uses random sampling to estimate statistical properties. Related terms: probabilistic modeling, risk analysis. Example: Simulating multiple tax-audit scenarios to assess exposure. Challenges: computational intensity, input distribution assumptions.

Named Entity Disambiguation (NED): Resolving which specific entity a mention refers to. Related terms: entity linking, coreference resolution. Example: Distinguishing “SNCF” as a railway operator versus a corporate subsidiary in a tax document. Challenges: limited context, overlapping entity names.

Neural Machine Translation (NMT): Deep-learning approach to MT using encoder-decoder architectures. Related terms: transformer, attention mechanism. Example: Translating French tax notices to English for multinational compliance teams. Challenges: rare legal terms, domain adaptation.

Neural Network: Computation model inspired by the human brain, consisting of interconnected nodes. Related terms: deep learning, backpropagation. Example: A feed-forward network predicting the probability of a tax audit based on transaction attributes. Challenges: overfitting, interpretability.

Noise-Robust Training: Techniques that improve model resilience to erroneous or noisy inputs. Related terms: data augmentation, regularisation. Example: Training an NLP model on OCR-derived tax documents that contain misspellings. Challenges: balancing robustness with precision.

OCR (Optical Character Recognition): Technology that converts scanned images of text into machine-readable characters. Related terms: image preprocessing, handwriting recognition. Example: Extracting VAT numbers from scanned receipts. Challenges: low-resolution scans, complex layouts.

Ontology Alignment: Process of mapping concepts from different ontologies to each other. Related terms: semantic mapping, schema matching. Example: Aligning a French tax ontology with an EU-wide fiscal ontology for data interoperability. Challenges: differing granularity, term equivalence.

Outlier Detection: Identifying data points that deviate markedly from the norm. Related terms: anomaly

detection, robust statistics. Example: Flagging unusually high VAT refunds for further audit. Challenges: defining normal thresholds, false alarms.

Parallel Corpus: Collection of texts in two languages that are translations of each other. Related terms: bilingual data, alignment. Example: Using a French-English parallel corpus to train a legal-domain MT model. Challenges: scarcity of high-quality legal parallel data.

Parsing (Syntax): Analyzing the grammatical structure of a sentence. Related terms: dependency parsing, constituency tree. Example: Parsing a tax clause to identify the subject, predicate, and conditional phrase. Challenges: complex legal syntax, long sentences.

Penalty Calculation Engine: Software that computes statutory penalties based on violations. Related terms: interest accrual, late filing fee. Example: An engine that uses NLP-extracted breach dates to calculate applicable penalties. Challenges: handling multiple penalty regimes, rounding rules.

Performance Metrics: Quantitative measures used to evaluate model effectiveness. Related terms: F1 score, AUC. Example: Reporting precision, recall, and F1 for a tax-document classification model. Challenges: selecting metrics aligned with business impact.

Personal Data: Any information relating to an identified or identifiable natural person. Related terms: GDPR, data subject. Example: Redacting taxpayer names from a public dataset used for model training. Challenges: balancing utility with privacy, consent management.

Phishing Detection: Identifying fraudulent emails that aim to steal credentials. Related terms: spam filter, malware. Example: An NLP filter that blocks tax-related phishing attempts targeting accountants. Challenges: evolving tactics, false negatives.

Pipeline Orchestration: Coordinating multiple processing steps in a data workflow. Related terms: Airflow, Dag. Example: Orchestrating OCR → tokenisation → NER → storage for tax document processing. Challenges: error handling, resource optimisation.

Post-Training Quantisation: Reducing model size by lowering numerical precision after training. Related terms: model compression, edge deployment. Example: Deploying a compact NLP model on a taxpayer-portal server. Challenges: maintaining accuracy, hardware compatibility.

Precision: Ratio of correctly predicted positive observations to total predicted positives. Related terms: accuracy, recall. Example: Measuring precision of an NLP model that flags “potential tax evasion” clauses. Challenges: high precision may reduce recall.

Privacy-Preserving Machine Learning: Techniques that protect sensitive data during model training. Related terms: federated learning, differential privacy. Example: Training a tax risk model across multiple firms without sharing raw data. Challenges: communication overhead, model convergence.

Probabilistic Topic Model: Statistical model that discovers abstract topics in a collection of documents. Related terms: LDA, latent Dirichlet allocation. Example: Identifying dominant tax themes (e.g., “digital

services," "transfer pricing") in rulings. Challenges: choosing number of topics, interpretability.

Prompt Engineering: Crafting inputs to guide large language models toward desired outputs. Related terms: few-shot learning, instruction tuning. Example: Designing prompts that elicit concise summaries of complex tax statutes from GPT-style models. Challenges: prompt brittleness, version control.

Quantitative Risk Assessment: Numerical evaluation of potential losses. Related terms: VaR, expected loss. Example: Estimating the monetary impact of a probable tax audit using historical data. Challenges: data availability, model assumptions.

RAG (Retrieval-Augmented Generation): Architecture that combines external knowledge retrieval with generative language models. Related terms: knowledge-grounded generation, contextualised LM. Example: A system that retrieves relevant tax articles before generating an answer to a user query. Challenges: retrieval latency, hallucination risk.

Regulatory Sandbox: Controlled environment that allows experimentation with innovative technologies under regulator supervision. Related terms: pilot program, innovation hub. Example: Testing an AI-driven tax compliance assistant with the French tax authority. Challenges: limited scope, compliance documentation.

Reinforcement Learning (RL): Learning paradigm where an agent interacts with an environment to maximise cumulative reward. Related terms: policy gradient, Q-learning. Example: Training an RL agent to optimise audit selection to balance revenue and fairness. Challenges: reward design, exploration-exploitation trade-off.

Remote Sensing Data: Information collected from satellites or aerial platforms. Related terms: geospatial analytics, GIS. Example: Using satellite imagery to verify the existence of a declared industrial site for tax purposes. Challenges: data resolution, integration with textual records.

Repository Pattern: Software design pattern that abstracts data access. Related terms: DAO, data abstraction. Example: Implementing a repository that hides whether tax documents are stored in a relational DB or a document store. Challenges: performance tuning, consistency.

Rule-Based System: Engine that applies explicit IF-THEN statements to derive conclusions. Related terms: expert system, knowledge base. Example: A rule-based system that determines VAT applicability based on product categories. Challenges: rule maintenance, handling exceptions.

Scalable Vector Graphics (SVG): XML-based format for two-dimensional graphics. Related terms: vector image, web rendering. Example: Visualising tax-risk heat maps as interactive SVGs on a compliance dashboard. Challenges: browser compatibility, data binding.

Semantic Search: Retrieval technique that understands the meaning behind queries. Related terms: embedding, vector similarity. Example: Searching for "tax incentives for renewable energy" and retrieving relevant clauses even if phrasing differs. Challenges: embedding quality, latency.

Sentiment Analysis: Process of determining the emotional tone behind words. Related terms: opinion

mining, subjectivity detection. Example: Analyzing taxpayer feedback on the ease of filing to inform service improvements. Challenges: domain specificity, sarcasm detection.

Sequence-to-Sequence (Seq2Seq): Model architecture that maps an input sequence to an output sequence. Related terms: encoder-decoder, attention. Example: Translating a French tax provision into plain English for layperson comprehension. Challenges: preserving legal precision, handling long inputs.

Service-Oriented Architecture (SOA): Design paradigm where services communicate over a network. Related terms: microservices, SOAP. Example: Exposing a tax-calculation service that can be called by external ERP systems. Challenges: versioning, security.

Shapley Value: Concept from cooperative game theory that fairly distributes payoff among participants. Related terms: feature importance, explainability. Example: Using Shapley values to explain which document features most influenced a tax-risk score. Challenges: computational cost, interpretability for non-technical users.

Side-Channel Attack: Exploiting indirect information (e.g., timing, power consumption) to compromise a system. Related terms: cryptanalysis, hardware security. Example: Protecting an AI model that processes confidential tax data from timing attacks. Challenges: mitigation strategies, detection.

Similarity Metric: Quantitative measure of likeness between two items. Related terms: cosine similarity, Jaccard index. Example: Measuring similarity between two tax rulings to suggest precedent. Challenges: selecting appropriate metric for legal text.

Single-Sign-On (SSO): Authentication method that permits a user to log in with one set of credentials across multiple applications. Related terms: OAuth, SAML. Example: Allowing tax consultants to access the AI compliance portal via corporate SSO. Challenges: federation trust, session management.

Soft-Skill AI: Artificial intelligence that assists with interpersonal aspects such as communication and negotiation. Related terms: conversational AI, empathetic bots. Example: A virtual assistant that explains tax obligations in a friendly tone. Challenges: cultural sensitivity, tone calibration.

Spacy: Open-source library for advanced NLP in Python. Related terms: tokeniser, pipeline. Example: Using Spacy's French model to perform tokenisation and POS tagging on tax documents. Challenges: custom entity addition, version compatibility.

Statistical Significance: Measure of whether an observed effect is likely due to chance. Related terms: p-value, confidence interval. Example: Testing whether a new AI-driven audit selection method yields a higher revenue capture rate. Challenges: sample size, multiple testing.

Stemming: Reducing words to their root form. Related terms: lemmatisation, morphological analysis. Example: Converting "taxable," "taxation," and "taxes" to the stem "tax" for indexing. Challenges: over-stemming leading to loss of meaning, language-specific rules.

Supervised Learning: Training models on labeled data to predict outcomes. Related terms: classification,

regression. Example: Training a classifier to label documents as “tax-relevant” using a manually annotated corpus. Challenges: label quality, class imbalance.

Tax Administration: Government agency responsible for tax collection and enforcement. Related terms: direction générale des finances publiques (DGFiP), audit. Example: Deploying AI tools within the French tax administration to streamline audit selection. Challenges: legacy systems, change management.

Tax Avoidance: Legal strategies used to minimise tax liabilities. Related terms: tax planning, anti-abuse rule. Example: NLP detection of clauses that suggest aggressive tax planning. Challenges: distinguishing legitimate optimisation from abusive schemes.

Tax Credit: Amount that reduces tax liability dollar-for-dollar. Related terms: tax incentive, reduction. Example: Extracting eligibility criteria for research-and-development credit from statutory text. Challenges: interpreting eligibility thresholds, interaction with other deductions.

Tax Evasion: Illegal practice of not paying taxes owed. Related terms: fraud, concealment. Example: Using anomaly detection to uncover under-reported income in electronic filings. Challenges: evidentiary standards, privacy constraints.

Tax Information Exchange Agreement (TIEA): Bilateral pact facilitating the exchange of tax-relevant information. Related terms: OECD, information sharing. Example: NLP processing of exchanged reports to extract relevant taxpayer identifiers. Challenges: data format heterogeneity, confidentiality.

Tax Law Corpus: Curated collection of statutes, regulations, case law, and commentary. Related terms: training data, knowledge base. Example: Building a French tax law corpus of 10,000 documents for language model fine-tuning. Challenges: licensing, keeping the corpus up-to-date.

Taxonomy Alignment: Mapping between different classification schemes. Related terms: ontology mapping, crosswalk. Example: Aligning the French NAF taxonomy with the EU-wide NACE classification for reporting purposes. Challenges: divergent granularity, semantic gaps.

Temporal Reasoning: Understanding and processing time-related information. Related terms: time-line extraction, event sequencing. Example: Determining the effective date of a tax amendment from legislative text. Challenges: ambiguous date expressions, relative temporal references.

Tokenisation: Splitting text into smaller units such as words or sub-words. Related terms: sentence segmentation, byte-pair encoding. Example: Tokenising a French tax decree before feeding it to a transformer model. Challenges: handling compound words, punctuation.

Transfer Learning: Reusing a pre-trained model on a new, related task. Related terms: fine-tuning, domain adaptation. Example: Adapting a generic French BERT model to the tax domain with a small annotated set. Challenges: catastrophic forgetting, over-fitting.

Triple-Store: Database optimized for storing and retrieving RDF triples. Related terms: graph database, SPARQL. Example: Storing tax ontology statements in a triple-store for semantic queries. Challenges: query

performance, data ingestion pipelines.

Unstructured Data: Information that does not follow a pre-defined data model. Related terms: free text, PDF. Example: Processing scanned tax rulings that lack consistent fields. Challenges: extraction accuracy, metadata generation.

Unsupervised Learning: Modeling techniques that find patterns without labeled outcomes. Related terms: clustering, dimensionality reduction. Example: Discovering hidden groupings of tax cases using autoencoders. Challenges: evaluation without ground truth, interpretability.

Version Control: System for tracking changes to code and data. Related terms: Git, branching. Example: Managing iterations of the tax-NLP model with Git to ensure reproducibility. Challenges: large binary files, data lineage.

Virtual Private Cloud (VPC): Isolated network environment within a public cloud. Related terms: network segmentation, security groups. Example: Deploying the AI tax compliance platform within a VPC to meet French data-sovereignty requirements. Challenges: configuration complexity, cost monitoring.

Weighted Ensemble: Combining multiple models where each contributes proportionally to its performance. Related terms: stacking, bagging. Example: Merging rule-based, statistical, and deep-learning predictions to improve tax risk scores. Challenges: correlation among models, ensemble maintenance.

Zero-Shot Learning: Ability of a model to perform tasks it has never seen during training. Related terms: prompting, few-shot. Example: Asking a large language model to summarise a new tax amendment without explicit fine-tuning. Challenges: reliability, domain specificity.