Strategy
www.tno.nl
+31 88 866 00 00
info@tno.nl

TNO-2026-16193 – 13 april 2026

LLM analyse

WP 4 van het Werkplan 2024
TNO Impact en Data-infrastructuur

Auteurs	Marcel de Heide Ruud Goorden

Rubricering verslag	TNO Public
Titel	TNO Public
Aantal pagina's	30
Aantal bijlagen	1
Opdrachtgever	Ministerie van Economische Zaken
Programmanaam	EZK Programmafinanciering Data Infrastructuur
Projectnaam	EZK Impact en Data-infrastructuur
Projectnummer	060.61215

Inhoudsopgave

1 Inleiding: beschrijving effecten onderzoek TNO

TNO heeft een werkplan opgesteld met een opzet van activiteiten voor de verbetering van de data-infrastructuur van TNO: het “Impact en Data-infrastructuur – Werkplan 2024 (TNO 2024 R12266 Impact en Data-Infrastructuur)”.[1] Hierin worden vier werk pakketten (WP’s) gedefinieerd: (i) Impact Stories professionaliseren; ii) PPS-data IT-systeem; iii) Theoretisch beleidskader impact TNO; iv) AI/Large Language Models inzetten om data te creëren uit ongestructureerde data.

Dit rapport beschrijft de eerste resultaten van WP4: een analyse van bestaande projectdata middels AI/LLM. Middels een pilot wordt de bruikbaarheid van deze technieken getest om sleuteltechnologieën te identificeren in projectdossiers.
Kader 1: Impact en Data-infrastructuur – Werkplan 2024 (TNO 2024 R12266).

1.1 Waardecreatiemodel: *Input - Impact*

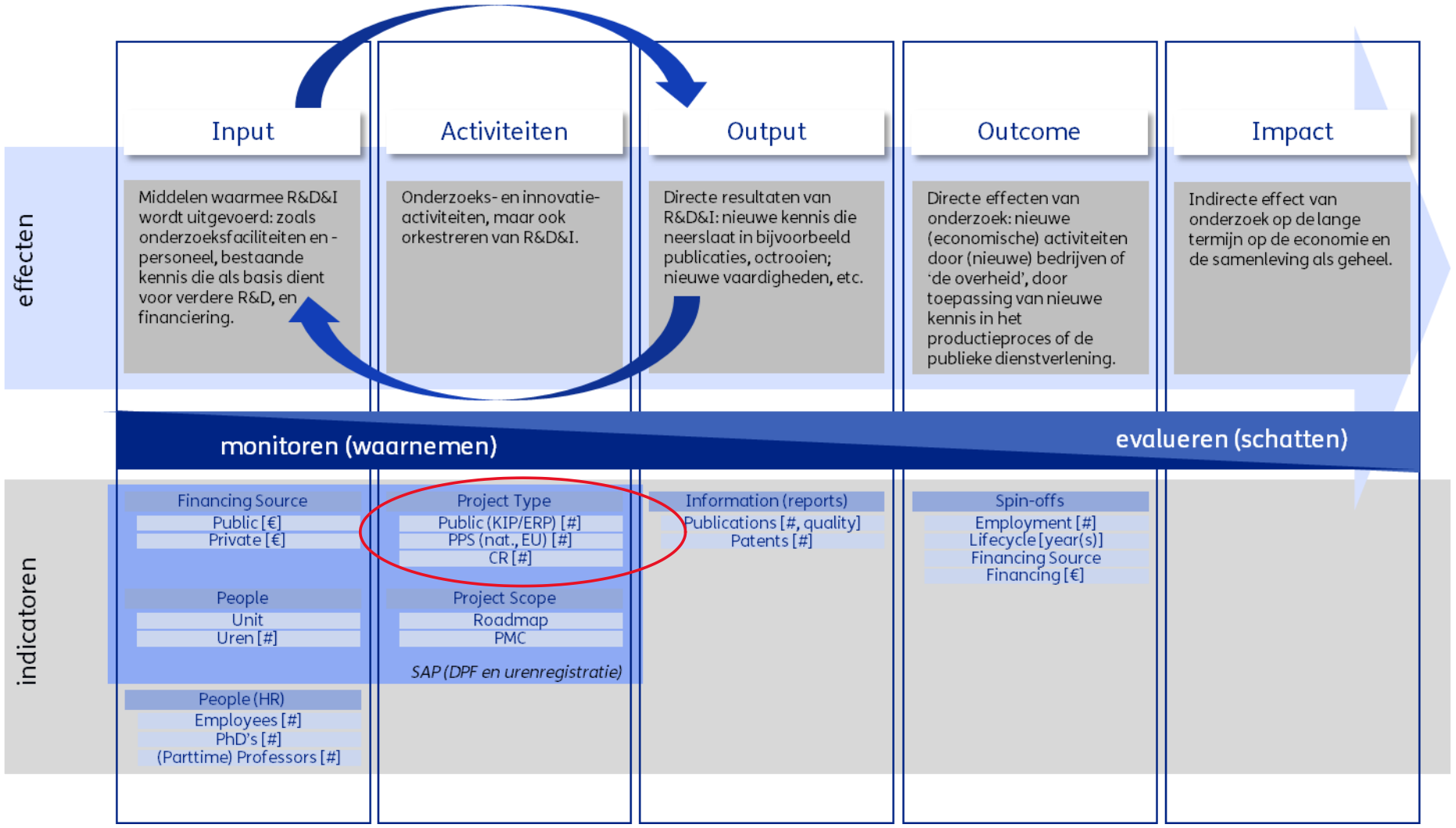
TNO hanteert als basis voor het beschrijven van de effecten van onderzoek het waardecreatiemodel (zie Figuur 1.1).

De ‘maatschappelijke discussie’ over de rol en relevantie van TNO gaat in de praktijk (voornamelijk) over hoe effectief TNO is in het creëren van *Outcome*, en daarmee *Impact* door het doen van onderzoek. *Outcome* gaat in die context over effecten op microniveau (het niveau van ‘de overheid’ en innoverende bedrijven die kennis toepassen in het ‘productieproces’): omvang van toegevoegde waarde, werkgelegenheid, etc. *Impact* gaat over de effecten op macroniveau (het niveau van de economie en de maatschappij ‘als geheel’): omvang van bbp-effecten maar ook ‘maatschappelijke’ impact zoals veiligheid.

De ‘politieke discussie’ over de rol en relevantie van TNO gaat (daarnaast) over hoe efficiënt TNO is (als één van de publieke instrumenten om R&D en innovatie aan te jagen) in het ‘omzetten’ van *Input* in *Outcome* / *Impact* - om de publieke financiering te verantwoorden. Voor EZ vormt inzicht in de effectiviteit én efficiëntie in die context daarnaast de basis om te trachten de totale beleidsmix verder te optimaliseren.

Voor TNO zelf kan juist inzicht in hoe *Input* leidt tot *Activities* tot *Output* en zo verder van *Outcome* tot *Impact* dienen als een basis voor het efficiënt sturen van bijvoorbeeld (keuzes voor) onderzoeksactiviteiten.

[1] Dit programma wordt gefinancierd door het Ministerie van Economische Zaken.



Figuur 1.1: Waardecreatiemodel.

1.2 Tekortkoming van huidige set indicatoren

Figuur 1.1 geeft ook een overzicht van de huidige informatie die TNO verzamelt in de context van het waardecreatiemodel. De basis voor de informatieverzameling is het SAP-systeem (software voor het managen van bedrijfsprocessen), en het bijbehorende *Digital Project Forms* (met informatie over onderwerp (titel), financieringsbron, scope, etc.) van het project en de 'urenregistratie' (met informatie over tarieven en tijdschrijven van onderzoekers). De informatie wordt (tijdens de uitvoering) verzameld en opgeslagen op het niveau van individuele projecten.

De indicatoren die worden verzameld in die context adresseren voornamelijk *Input*, *Activiteiten* en *Output* van onderzoek, maar niet tot nauwelijks de effecten verder in het waardecreatiemodel. Daarnaast zijn de indicatoren van links naar rechts in het waardecreatiemodel niet 'gelinkt' - waarmee ontwikkeling en toepassing van kennis (bijvoorbeeld in een bepaald technologieveld) in de tijd niet is te volgen. Zo is het met de huidige manier van verzamelen en opslaan van informatie bijvoorbeeld niet mogelijk om een project te koppelen aan een vervolgproject waarbij de *Output van* de één gebruikt is als *Input* van de ander. Daarnaast is het nu niet (altijd '1:1') duidelijk is welke projecten ten grondslag liggen aan een bepaald patent. Dit maakt het onmogelijk is om (met behulp van data) een compleet beeld te geven van de rol en relevantie van TNO.

1.3 Analyse projectdata middels AI/LLM

Maar TNO beschikt naast de indicatoren zoals hierboven beschreven ook over 'ongestructureerde' gegevens over projecten die bijvoorbeeld zijn opgenomen in digitale projectdossiers - zoals projectvoorstellen, voortgangsrapportages, etc. In deze tekstuele documenten staat veel informatie opgenomen over de inhoud van het onderzoek, samenwerkingsvormen, en uitkomsten.

Tot op heden was informatie in deze vorm niet eenvoudig bruikbaar voor het analyseren van de impact van TNO. Door de ontwikkelingen op het gebied van Large Language Models (LLM), zijn er nu nieuwe mogelijkheden beschikbaar gekomen om alsnog deze informatie te gebruiken voor een nadere analyse.

In de context van WP4 is daarom een pilot uitgevoerd (bestaande uit verschillende experimenten), met als doel om de effectiviteit van LLM wat betreft het structureren van informatie te toetsen. Daartoe zijn met behulp van specifieke 'tools' bepaalde 'aspecten' (van de effecten) van onderzoek geïdentificeerd in projectdossiers. Specifiek was het doel om projecten, als een vorm van *Activiteiten*, te kunnen 'labelen' volgens de sleuteltechnologieën van de NTS - om zo de bijbehorende 'micro-data' uit het SAP systeem te kunnen aggregeren, en indicatoren te kunnen samenstellen op het niveau van specifieke technologieën.[2] Deze pilot is uitgevoerd door het *Data and Analytics* team van TNO.

[2] Zie: De Nationale Technologiestrategie. Bouwstenen voor strategisch technologiebeleid | Rapport | Rijksoverheid.nl.

2 Methodiek

De centrale vraag die in de context van deze pilot wordt geadresseerd is: *aan welke sleutel- en sub technologieën (Key Enabling Technologies, KET's) draagt TNO bij?* Door de specifieke aard van KET's, het ontbreken van formele / gestructureerde registratie, en het feit dat mogelijk relevante informatie in deze context verspreid is geregistreerd in meerdere bronnen en documenten in het projectdossier is deze vraag nu niet te beantwoorden.

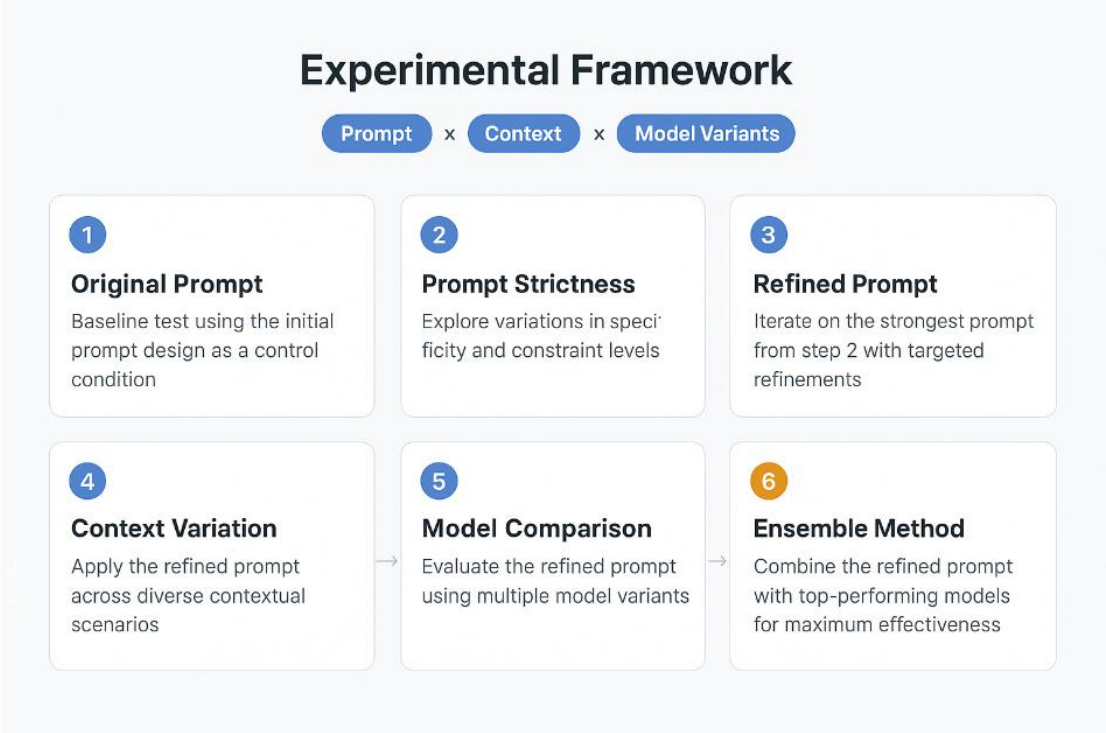
Als eerste aanzet om tot nadere inzichten te komen in deze context richt het onderzoek zich op het automatisch classificeren van TNO's bijdragen aan KET's via Large Language Models (LLMs). De brondata bestaat uit KIP-documenten (Kennis- en Innovatieprojecten), die in verschillende formaten en op diverse locaties (SharePoints) zijn opgeslagen. Deze documenten bevatten onder andere projectinformatie, doelstellingen, partners, resultaten en impact.

Het onderzoek is als volgt opgezet:

- **Data:** KIP-documenten zijn met behulp van *Azure Document Intelligence* omgezet naar '.md-bestanden' voor verdere analyse.
- **Labeling:** Om tot een eerste inzicht te komen is een 'feasibility-sample' van 113 documenten samengesteld uit een set van 'direct beschikbare' documenten (zogenaamde KIP rapportages zoals hierboven beschreven). Deze sample is vervolgens handmatig 'gelabeld' door strategy-analisten (niet de domeinexperts). Daartoe is samen met deze strategy-analisten een (eerste aanzet tot) een uniforme definitie van 'bijdrage aan KET' geformuleerd. De 'vindplaats' van KET bijdrage in de documenten is ook bij labeling geregistreerd (t.b.v. context).
- **Proces:**
  - Tekstsegmenten zijn geëxtraheerd en per tekstsegment met een LLM geclassificeerd. Voor KET classificatie waren dat de secties met omschrijvingen doel en resultaten. Dit impliceert dat de relevante stukken tekst waarin een mogelijke KET classificatie te herleiden was zijn gebruikt en aangeboden aan de LLM
  - Documenten met ontbrekende tekst segmenten voor KET classificatie zijn in het geheel aan de LLM aangeboden.[3]
  - KET- en subtechnologie-classificatie gebeurt met specifieke prompts. Dit wil zeggen dat we instructies geven aan de LLM over hoe en wanneer een KET toe te wijzen
  - Subtechnologies werden dynamisch verwerkt a.h.v. geclassificeerde KET's. Indien er een KET is toegewezen wordt m.b.v. prompting wederom aan de LLM gevraagd om er een of meerdere subtechnologieën aan toe te wijzen
  - Resultaten worden gelogd en samengevat in Excel, zodat een beeld werd verkregen hoe vaak er met specifieke context kon worden geprocessed. Dit gaf een beeld over of we vaak volledige documenten moesten aanbieden of juist de tekstuele delen die specifiekere context bevatte. Je wil bij een LLM juist de specifieke context kunnen aanbieden. Daarom hebben we dit overzicht gemaakt
- ***Prompt engineering*:** Verschillende promptvarianten zijn getest (strikt, gebalanceerd, inclusief), met variatie in strengheid, context en modeltype (o.a. GPT-4, GPT-5 en varianten).

[3] Ontbrekende tekstsegmenten zijn het resultaat van oneigenlijk gebruik van de formats voor de rapportages: tekstsegmenten zijn weggehaald, hernoemd, etc. door de 'rapporteurs'.

De onderzoeksaanpak is samengevat in Figuur A.1, en gedetailleerd beschreven in Bijlage A. Het experimenteel raamwerk gaat over de assen van prompts (striktheid van toewijzingen, i.c.m. toewijzing o.b.v. drempelwaarden), context (tekstsegmenten versus hele documenten), model varianten (b.v. GPT-4, GPIT-5, etc.).



Figuur A.1: Opzet pilot: 6 experimenten.

3 Resultaten

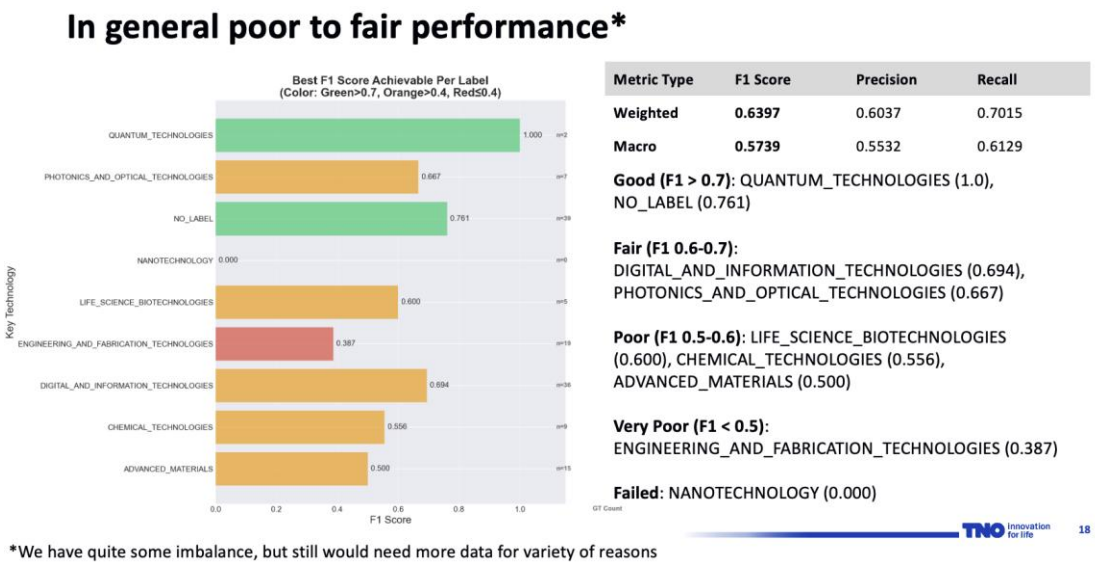
De experimenten van het raamwerk zijn als het ware 'opvolgend', in dat de resultaten van een eerder experiment de vormgeving van het daaropvolgend experiment mede bepaald. In die context zijn de resultaten van experiment 2 en 6 het meest relevant wat betreft de onderzoeksvraag zoals gedefinieerd in Hoofdstuk 2. De resultaten van deze experimenten en enkele meer generieke waarnemingen zijn hier onder beschreven - de quotes zijn hierbij van het onderzoeksteam. Voor de volledigheid zijn de overige resultaten beschreven in Bijlage A.

- **Labeldistributie:** De resultaten tonen grote verschillen in labelaanduidingen. Enkele KET's, zoals nanotechnologie, werden niet gelabeld omdat ze ontbraken in de documentenset. Circa 30% van de documenten bood geen inzicht in bijdragen aan KET's ("management KIP's"). Sommige KET's kwamen slechts incidenteel voor (1-10 keer), wat de betrouwbaarheid beïnvloedt: één fout geeft relatief een grote afwijking.
- **Promptvergelijking (experiment 2, zie voorbeeld** Figuur A.1**):**
  - Strikte prompts te zijn conservatief en missen veel bijdragen (lage *recall*, hoge precisie). Dat wil zeggen: van de KET's die gelabeld waren, werden er relatief weinig toegekend, m.a.w. we missen toewijzingen van labels. "Als we 10 labels technologie A hebben werden er maar 3 toegekend". De labels die wel werden toegekend, waren relatief wel juist toegekend: "als het technologie A is, dan is het ook technologie A"
  - Inclusieve prompts zijn te ruim en geven veel valse positieven (hoge *recall*, lage precisie).Dit betekent: "veel labels van technologie A aan werden toegekend, maar tegelijkertijd zorgde dat ervoor dat er ook labels die niet technologie A werden toegekend."
  - Gebalanceerde prompt biedt de beste balans tussen precisie en *recall*, met de hoogste F1-score (0.52).[4] "Dit is de balans tussen *veel* labels van technologie A toekennen en ze ook *juist* toekennen, d.w.z. technologie B heeft geen label toegekend gekregen dat het technologie A is"
- **Modelprestaties:** prestaties variëren van matig tot redelijk, met duidelijke ruimte voor verbetering, vooral bij minderheidsklassen. Model prestaties zijn beoordeeld o.b.v. de F1-score
- **Ensemble-aanpak (experiment 6, zie** Figuur A.2**):**
  - Voor elk label kun je een specifiek model en prompting kiezen. Dit verbetert de prestaties licht, maar sommige categorieën blijven last houden van over-predictie, waarbij te vaak onterecht een label wordt toegekend.
  - Bovendien blijft de overall performance (F1- score, zie boven) te laag om dit om historische documenten op schaal daadwerkelijk toe te passen.

[4] De **F1-score** (of F-maat) is een statistische maatstaf die wordt gebruikt om de prestaties van een classificatiemodel (machine learning) te evalueren. Het is het harmonisch gemiddelde van **precisie** (precision) en **herinnering** (recall), waardoor het een betrouwbaarder beeld geeft dan enkel nauwkeurigheid (accuracy), vooral bij onevenwichtige datasets. Hoge F1-score (dicht bij 1): Goede balans tussen precisie en recall. Lage F1-score (dicht bij 0): Slechte prestaties, waarbij ofwel de precisie of de recall (of beide) laag is.

Feature	Strict (Original)	Balanced (Version 1)	Inclusive (Version 2)
Primary Goal	Identify high-impact, novel R&D.	Capture meaningful, valuable technical contributions.	Cast a wide net for any technical component.
Core Requirements	**Substantial** Technical Innovation. Requires a high-depth, novel advancement.	**Significant** Technical Contribution. Accepts meaningful improvements.	**A** Technical Contribution. Any technical work is sufficient.
Evidence Validation	Requires answers to 4 specific questions, including "Advancement beyond state-of-the-art" and "Substantial Impact."	Requires answers to 4 questions, but rephrases "Advancement" to "improves existing approaches" and "Impact" to "meaningful."	Requires answers to only 2 questions, focused on identifying a technical component and evidence. No "Advancement" or "Impact" test.
Confidence Threshold	**> 0.5** to classify as a KET.	**> 0.4** to classify as a KET.	**No threshold.** The score is for descriptive purposes only.
Final Rule	"When in doubt, classify as **NO_KET**."	"If evidence is present but not overwhelming, err on the side of giving a **KET**."	"Give a **KET** for any project that is not explicitly excluded."
Exclusions	Highly detailed list. For mixed projects, technical component must be **central and substantial**.	Highly detailed list. For mixed projects, technical component must be **central and significant**.	Simplified list. A project is **NOT** excluded if it has **ANY** technical component.

Figuur A.1: Promptvergelijking - experiment 2



Figuur A.2: Ensemble-aanpak - experiment 6.

4 Conclusies

Op basis van de pilot kan het volgende geconcludeerd worden:

- Automatische KET-classificatie met LLM's is mogelijk haalbaar, maar de prestaties zijn nog niet optimaal in gegeven huidige setting.
- Er is sprake van *class imbalance* en ondervertegenwoordiging van bepaalde KET's, wat meespeelt in de resultaten: meer data is nodig. Dat betekent dat niet elk label goed vertegenwoordigd is qua aantallen (sommige kwamen veel voor, anderen niet of nauwelijks).
- Prompt engineering is van essentieel belang: een zorgvuldig geformuleerde prompt levert optimale resultaten onder de huidige omstandigheden. Zelfs minimale aanpassingen in de opdracht aan een LLM kunnen bij ieder model tot uiteenlopende resultaten leiden.
- Menselijke beoordeling blijft essentieel voor het waarborgen van de kwaliteit binnen de golden dataset. Het is om die reden niet aan te raden om labels door slechts één persoon te laten toewijzen en deze vervolgens als absolute waarheid ("de gouden standaard") te beschouwen. Dit geldt des te meer wanneer er ruimte is voor discussie over de precieze definitie van een 'KET-bijdrage'.

De huidige aanpak levert onvoldoende resultaat op. Verdere analyse van data (en peer review) en optimalisatie van prompts is noodzakelijk voor betere uitkomsten.

1. **Algemene overeenstemming** over **criteria** om tot **KET toewijzing** te komen (liggen al aantal templates uit experiment welke kunnen worden besproken). Om tot een initiële verbetering te komen is echter stap (2) ook van belang. Dit is een definitie kwestie.
2. Met in stap (1) **overeengekomen prompt** de **gelabelde documenten opnieuw bekijken** (in peers) (inspanning van 1 + 2 is middelmatig, en kan relatief snelle kwaliteitswinst opleveren). Dit betekent dus m.b.v. de opgestelde prompt EN de opnieuw gelabelde documenten opnieuw evalueren of de kwaliteit toereikend blijkt.
3. **Automatische promptoptimalisatie** kan systematisch de resultaten verbeteren, mits stap (1) en (2) zijn uitgevoerd. Deze relatief nieuwe techniek houdt het programmatisch herformuleren van prompts in. Onze latere optimalisatiepogingen leverden geen beter resultaat op, waarschijnlijk omdat stap 1 en 2 niet waren aangepast.
4. **Label extra data in ondervertegenwoordigde categorieën** via *active learning*: minimaal 30 documenten per KET (in totaal 150-200). Dit verbetert de labelverdeling en voegt meer subtechnologieën toe aan de dataset. Daarna kan automatische promptoptimalisatie opnieuw worden ingezet.

5 Literatuur

TNO (2024). Impact en Data-Infrastructuur. TNO 2024 R12266.

Bijlage A

Onderzoeksopzet en Resultaten


LLM Strategy POC
Results of experiments KET classification
2-10-2025
TNO
TNO innovation for life


Background
AS TNO WE ARE OBLIGED TO REPORT ON OUR IMPACT TO THE GOVERNMENT (AND OTHER BODIES)
FOR THAT WE HAVE A LOT OF INFORMATION WE COLLECT EITHER STRUCTURED AND NON-STRUCTURED
INFORMATION REQUESTS ARE SOMETIMES NOT DIRECTLY STORED IN OUR DATA GATHERING PROCEDURES / SYSTEMS WHICH MAKES IT HARD
TNO innovation for life

Key challenge

TNO

To which **key- and sub-technologies** does TNO contribute?

* Key Enabling Technologies (KET = Sleuteltechnologie)

TNO innovation for life 3

Some background on KET*

It might be difficult given the **specific nature** of the key technologies (moreover the sub-technologies (40))

No formal registration, but we have KIP Documents: documents on research results, and have multiple informative text sections scattered around the document

What is a contribution to? Sometimes it might be formulated, sometimes not, etc.

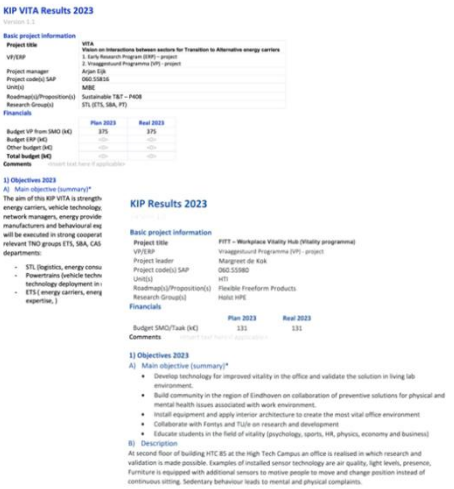
~hundreds of documents per year

* Key Enabling Technologies (Sleutel technologieen)

TNO innovation for life 4

Character of KIP documents

- PDF and .docx documents scattered around in Sharepoints
- Describes meta data and research done, partners involved, results, etc.
- Different formats per year and even within a year



KIP VITA Results 2023

Basic project information

Project title	VITA Vision on Interactions between sectors for Transition to Alternative energy carriers
VP/ERP	1. Early Research Program (ERP) – project 2. Vraaggestuurd Programma (VP) - project
Project manager	Arjan Eijk
Project code(s) SAP	060.55816
Unit(s)	MBE
Roadmap(s)/Proposition(s)	Sustainable T&T – P408
Research Group(s)	STI, ETS, SBA, PT)

Financials

	Plan 2023	Real 2023
Budget VP from SMO (k€)	375	375
Budget ERP (k€)		
Other budget (k€)		
Total budget (k€)		

Comments

1) Objectives 2023

A) Main objective (summary)*

The aim of this KIP VITA is strength energy carriers, vehicle technology, network managers, energy provide manufacturers and behavioural eq will be executed in strong cooperat relevant TNO groups ETS, SBA, CAS departments:

- STL (logistics, energy consu
- Powertrains (vehicle techn technology deployment in
- ETS (energy carriers, energ expertise,)

KIP Results 2023

Basic project information

Project title	FITT – Workplace Vitality Hub (Vitality programma)
VP/ERP	Vraaggestuurd Programma (VP) - project
Project leader	Margreet de Kok
Project code(s) SAP	060.55580
Unit(s)	HTI
Roadmap(s)/Proposition(s)	Flexible Freeform Products
Research Group(s)	Holst HPE

Financials

	Plan 2023	Real 2023
Budget SMO/Taak (kC)	131	131

Comments

1) Objectives 2023

A) Main objective (summary)*

- Develop technology for improved vitality in the office and validate the solution in living lab environment.
- Build community in the region of Eindhoven on collaboration of preventive solutions for physical and mental health issues associated with work environment.
- Install equipment and apply interior architecture to create the most vital office environment
- Collaborate with Fontys and TU/e on research and development
- Educate students in the field of vitality (psychology, sports, HR, physics, economy and business)

B) Description

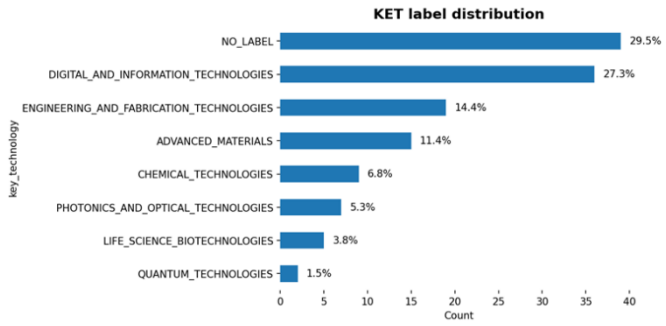
At second floor of building HTC 85 at the High Tech Campus an office is realised in which research and validation is made possible. Examples of installed sensor technology are air quality, light levels, presence, Furniture is equipped with additional sensors to motive people to move and change position instead of continuous sitting. Sedentary behaviour leads to mental and physical complaints.

* Processed with Azure Doc Intelligence to .md files (Note: came out best in small experiment with open-source alternative)

TNO innovation for life

KIP Results 2023 Version 1.0	KIP Results 2023 Version 1.1	KIP Results 2022 Version 1.0 EN	KIP Results 2022 Version 1.2 EN
Basic project information Project title VP/ERP Project leader Project code(s) SAP Unit(s) Roadmap(s)/Proposition(s) Research Group(s)	Basic project information Project title VP/ERP Project leader Project code(s) SAP Unit(s) Roadmap(s)/Proposition(s) Research Group(s)	Basic Project Information Project title Project Leader Project Code SAP Project Start Project End Unit Roadmap VP/ERP Research Group	Basic Project Information Project title Project Leader Project Code SAP Project Start Project End Unit Roadmap VP/ERP Research Group
Financials	Financials	Budgets Government Category of Research	Budgets
1) Objectives 2023 A) Main objective (summary)* B) Description	1) Objectives 2023 A) Main objective (summary)* B) Description	1) Objectives 2022	1) Objectives 2022
2) Partners Category | Name(s)	2) Partners Category | Name(s)	2) Progress of Project A) Plan of Work B) Progress Realized C) Financial Realization D) Cooperation E) Sponsoring	2) Progress of Project A) Plan of Work B) Progress Realized C) Financial Realization D) Cooperation E) Sponsoring
3) Results 2023 A) Short description (summary)* B) Listing of results Category | Result | Realized C) Assessment of results	3) Results 2023 A) Short description (summary)* B) Listing of results Category | Result | Realized C) Assessment of results	3) Realized contribution to new knowledge and technology A) New knowledge and technology B) Improvement technology position of TNO C) Relevance for governmental departments or companies/organizations	3) Realized contribution to new knowledge and technology A) New knowledge and technology B) Improvement technology position of TNO C) Relevance for governmental departments or companies/organizations
4) Expected impact of the results 2023 A) Relevance to public and/or private stakeholders B) Relevance to TNO's PMCs/PMC cluster/propositions	4) Expected impact of the results 2023 A) Relevance to public and/or private stakeholders B) Relevance to TNO's PMCs/propositions	4) Realized contribution to market position A) Improvement market position of TNO B) Actions for valorization C) Protection of IP	4) Realized contribution to market position A) Improvement market position of TNO B) Actions for valorization C) Protection of IP
5) Outreach activities 2023 Category | Activity	5) Outreach activities 2023 Category | Activity	5) Results Realized Number | Type of result | Results | initially planned | Realized | TRL/SRL from → to	5) Results Realized Number | Type of result | Results | initially planned | Realized | TRL/SRL from → to
6) Adjacent projects	6) Adjacent projects	6) Publicity Number | Type of publicity | Reference to publicity	6) Publicity Number | Type of publicity | Reference to publicity
		7) Continuation next years	7) Continuation next years

Statistics and learnings from labelling KET contribution



Labeller 1: 75%, Labeller 2: 18%, Labeller 3: 6%, Labeller 4: 1%

Procedure

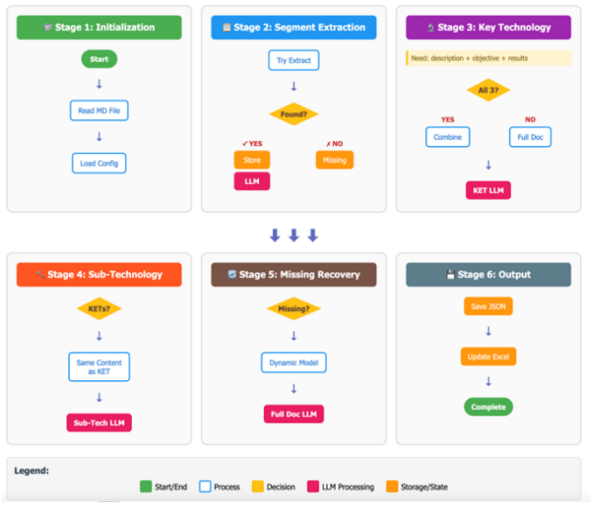
- Started with **feasibility** sample(113) since we didn't have all KET documents in one place (scattered across Sharepoints in the organization)
- Labelling was set out at Strategy Analysts (**Not THE domain experts** which have done the studies)
- Analysts were also asked where **(section(s)/context)** they based their conclusion on
- In essence there was a guidance on the technologies, but no real uniform definition was put **on paper** regarding "**contribution**" to a KET

Result

- KIP documents we also no KET assigned (management documents, vision documents, etc.)

TNO innovation for life 7

Modular process flow

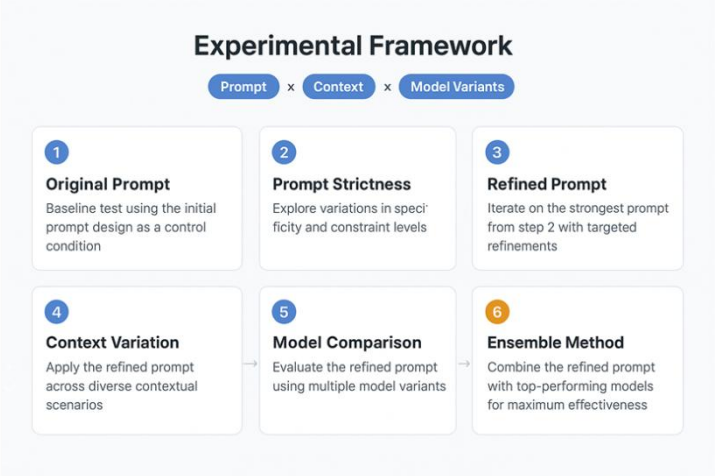


- Configuration of text segment extraction and Pydantic classes are loaded
- Segments are extracted
- If segments are found processed with LLM and their specific prompt and text segment
- Missing segments are saved and processed later with full document and their respective prompts
- KET is processed if certain segments are found. If not, full document. Specific prompt is used
- Sub-technology is processed if KETs are found (with certain segments or full document)
- Results of processing are logged and saved to Excel summary

Use of cheaper model for segments and more expensive for KET / Subtech

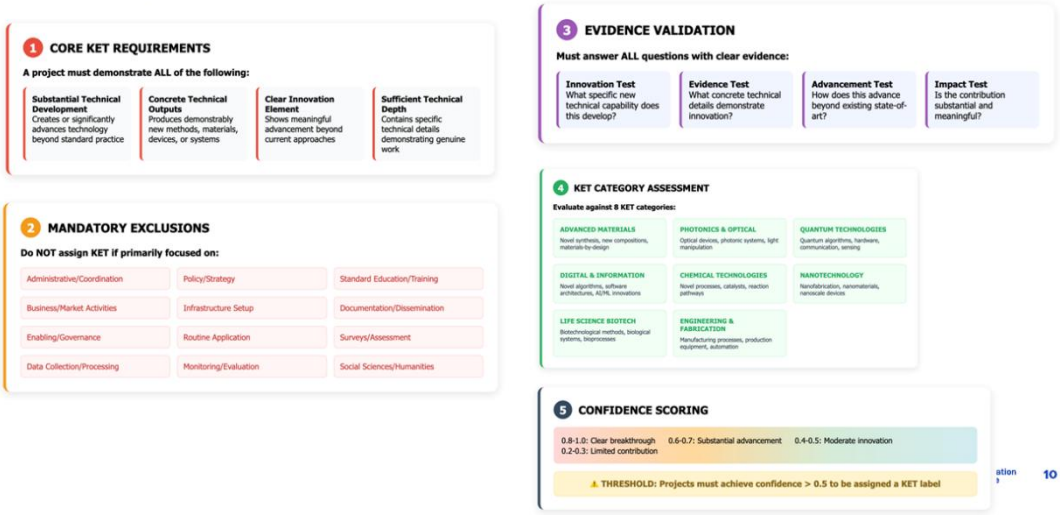
TNO innovation for life 8

Experiments



TNO innovation for life 9

Prompt setup



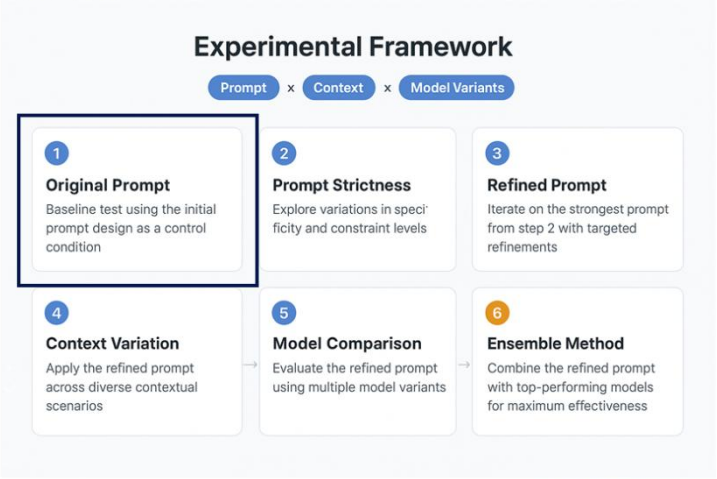
10

Initial GPT-4o-mini and GPT-5-mini

Feature	GPT-4o Mini	GPT-5 Mini
Release	May 2024 (introduced with GPT-4o as a smaller, faster variant)	August 2025 (released alongside GPT-5 flagship & Nano)
Core Purpose	Lightweight, cheaper, faster version of GPT-4o for general tasks	Mid-tier model balancing cost, speed, and reasoning between GPT-5 Nano and GPT-5 flagship
Capabilities	- Multimodal (text + vision, limited audio)- Optimized for **speed and efficiency**- Stronger than GPT-3.5 in reasoning- Suited for conversational AI and lighter workloads	- Better reasoning than GPT-4o Mini- Supports multimodality (text + images, possibly audio)- Handles more complex problem solving- Sits between GPT-5 Nano (speed) and GPT-5 flagship (depth)
Strengths	- Very low latency- Cost-efficient- Good for lightweight apps (chat, summarization, simple Q&A)- Solid reasoning for its size	- Stronger reasoning than 4o Mini- Better at coding, math, health, and complex queries- Larger context window than 4o Mini- Good balance of speed and capability
Weaknesses	- Weaker reasoning vs GPT-4o / GPT-5 models- Limited for highly technical or deep reasoning tasks- Smaller context window- Knowledge cutoff older than GPT-5	- Less powerful than GPT-5 flagship- May lag on very specialized or high-stakes domains- More costly/slower than Nano variant
Knowledge Cutoff	**May 2023**	**May 2024**
Best Use Cases	- Real-time chatbots- Summarization- Customer support- Everyday Q&A	- More advanced assistants- Coding helpers- Analytical tasks- Mid-complexity research / technical support

TNO innovation for life 11

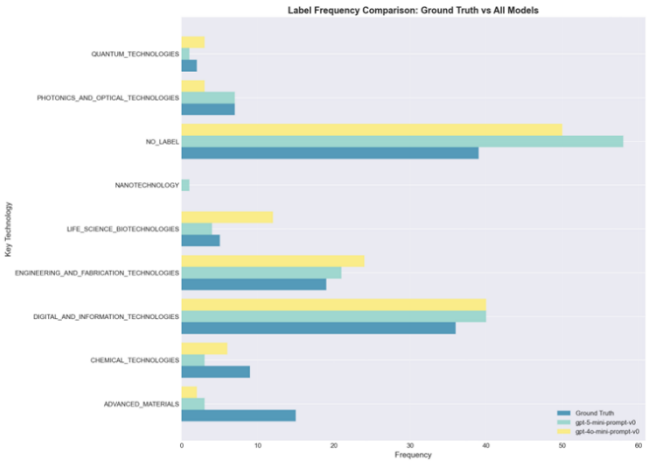
Experiments



TNO innovation for life 12

Experiment 1 Prompt and 2 model variants

Over prediction of no labels.



1. Over-prediction of NO_LABEL
Both models significantly over-predict NO_LABEL:
GPT-4o-mini: ~50 (vs 38 ground truth)
GPT-5-mini: ~57 (vs 38 ground truth)
GPT-5-mini shows more severe over-prediction bias

2. Minority Class Under-representation
Categories like QUANTUM_TECHNOLOGIES, NANOTECHNOLOGY, and ADVANCED_MATERIALS have very low ground truth frequencies. Evaluating those is difficult

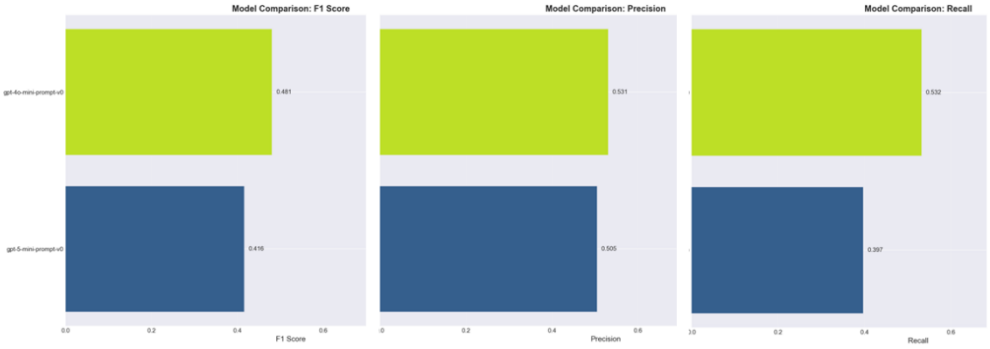
3. Moderate Success on M ior Cla
DIGITAL_AND_I RMAT TE S shows reasonable alignm ound truth
ENGINEERING_A IES shows mixed

Quality?

TNO innovation for life 13

Experiment 1 Prompt and 2 model variants

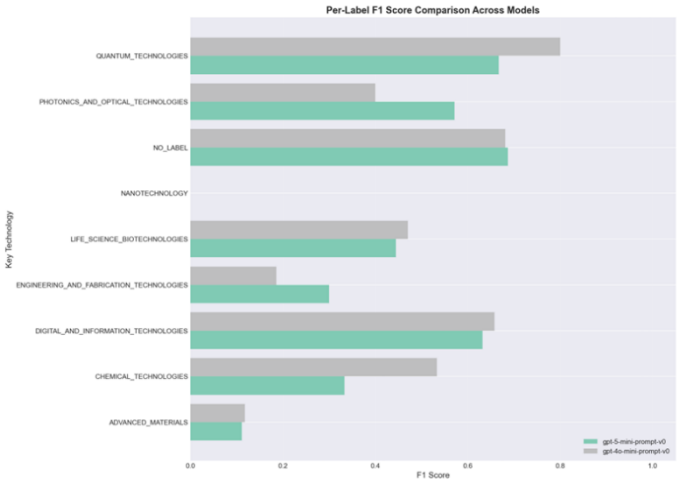
GPT 4o-mini tends to be slightly better, but overall scores both very poor



TNO innovation for life 14

Experiment 1 Prompt and 2 model variants

Performance varies between model across KETs



1. Model Performance Varies Significantly by Class
GPT-4o-mini shows more consistent performance across categories

GPT-5-mini has more variable performance - some high peaks, some low valleys

2. Class-Specific Strengths:
GPT-4o-mini excels at:
QUANTUM_TECHNOLOGIES (0.8 vs 0.67)
CHEMICAL_TECHNOLOGIES (0.52 vs 0.38)
GPT-5-mini excels at:
PHOTONICS_AND_OPTICAL_TECHNOLOGIES (0.55 vs 0.40)
ADVANCED_MATERIALS: Both perform very poorly (F1 < 0.15)

3. Moderate Performance Classes:
DIGITAL_AND_INFORMATION_TECHNOLOGIES: Both around 0.62-0.67
LIFE_SCIENCE_BIOTECHNOLOGIES: Both around 0.45-0.48
NO_LABEL: Reasonable performance (~0.67-0.70) despite over-prediction bias

The F1 scores reveal that **neither model is universally superior**

TNO innovation for life 15

What you could bump into: GPT-5 processing issues

"We've limited access to this content for safety reasons" for biology questions

Other

Working in chemical biology space and have now found that GPT5 is limiting responses on most questions that would be useful. Used to use it o3 a lot to find literature about my field and use it for troubleshooting etc but now it seems to be essentially useless as most answers get filtered. Even some really simple questions that would not have any malicious application seem to be getting overzealously filtered out.

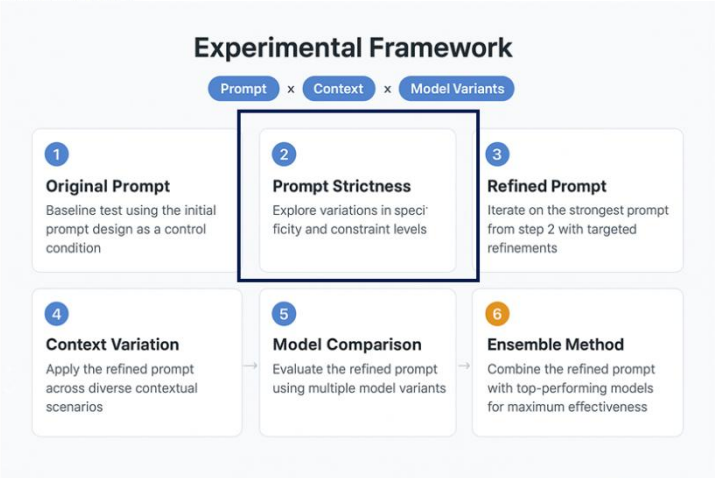
https://openai.com/index/preparing-for-future-ai-capabilities-in-biology/

"We've limited access to this content for safety reasons.This type of information may be use… benef… to harm people. We are continuously refining our work in this area, and you can read more about o… our blog post and Model Spec."

Small sidenote

TNO innovation for life 16

Experiments



TNO innovation for life 17

Experiment 2: Different formats of the prompt in "strictness of KET assignment"

Parameter changes across 5 dimensions of the prompt*

Feature	Strict (Original)	Balanced (Version 1)	Inclusive (Version 2)
Primary Goal	Identify high-impact, novel R&D.	Capture meaningful, valuable technical contributions.	Cast a wide net for any technical component.
Core Requirements	**Substantial** Technical Innovation. Requires a high-depth, novel advancement.	**Significant** Technical Contribution. Accepts meaningful improvements.	**A** Technical Contribution. Any technical work is sufficient.
Evidence Validation	Requires answers to 4 specific questions, including "Advancement beyond state-of-the-art" and "Substantial Impact."	Requires answers to 4 questions, but rephrases "Advancement" to "improves existing approaches" and "Impact" to "meaningful."	Requires answers to only 2 questions, focused on identifying a technical component and evidence. No "Advancement" or "Impact" test.
Confidence Threshold	**> 0.5** to classify as a KET.	**> 0.4** to classify as a KET.	**No threshold.** The sco for descriptive purposes only.
Final Rule	"When in doubt, classify as **NO_KET**."	"If evidence is present but not overwhelming, err on the side of giving a **KET**."	"Give a **KET** for any p that is not explicitly excluded."
Exclusions	Highly detailed list. For mixed projects, technical component must be **central and substantial**.	Highly detailed list. For mixed projects, technical component must be **central and significant**.	Simplified list. A project is **NOT** excluded if it has **ANY** technical component.

Gemini adjusted prompts were used

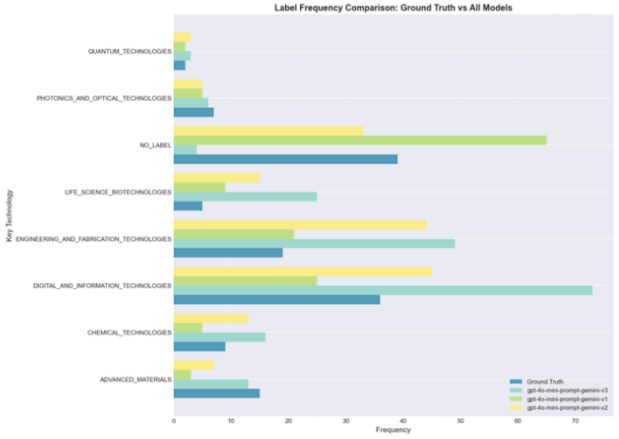
We show only GPT-4o-mini results since they tended to better than GPT-5-mini

* LLMs were used to produce different prompts given our initial prompt also which has a different structural setup

TNO innovation for life 18

Experiment 2 Prompt variants

Inclusive prompt: a lot of labels predicted, Strict least. Balanced in between



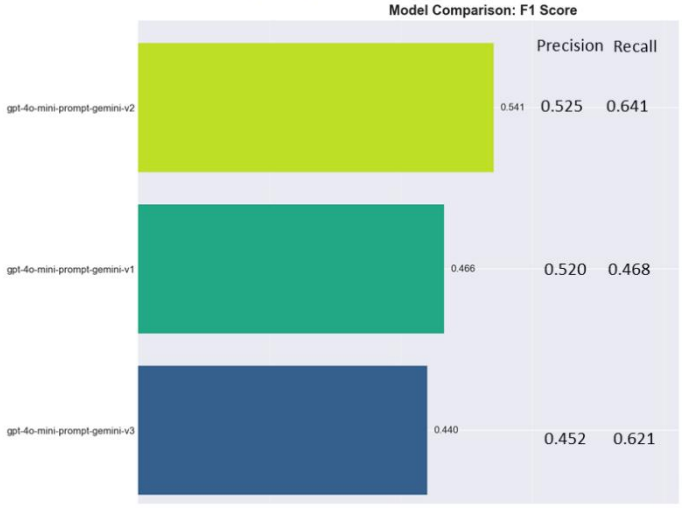
V3 Shows Severe Over-Prediction Bias:
•`DIGITAL_AND_INFORMATION_TECHNOLOGIES`:
• **Massive over-prediction** (~68 vs 35 GT)
• `NO_LABEL`: Significant over-prediction (~58 vs 38 GT)
•`LIFE_SCIENCE_BIOTECHNOLOGIES`: Triple over-prediction (~25 vs 8 GT)
• **Pattern**: V3 is extremely liberal/aggressive in classification

V1 & V2 Show Conservative Tendencies:
•`ADVANCED_MATERIALS`: **Severe under-prediction** (V1: ~3, V2: ~2 vs 15 GT)
•`PHOTONICS_AND_OPTICAL_TECHNOLOGIES`: Under-prediction across all models
•V1 slightly under-predicts `NO_LABEL` and `DIGITAL_AND_INFORMATION_TECHNOLOGIES`
V2 Shows Best Balance:
•Closest alignment to ground truth for `NO_LABEL` (~35 vs 38 GT)
•Reasonable performance on most categories
•Still struggles with minority classes

TNO innovation for life 19

Experiment 2 Prompt variants

Balanced (v2) seems overall sweet spot, but still not



Overall Performance Ranking:

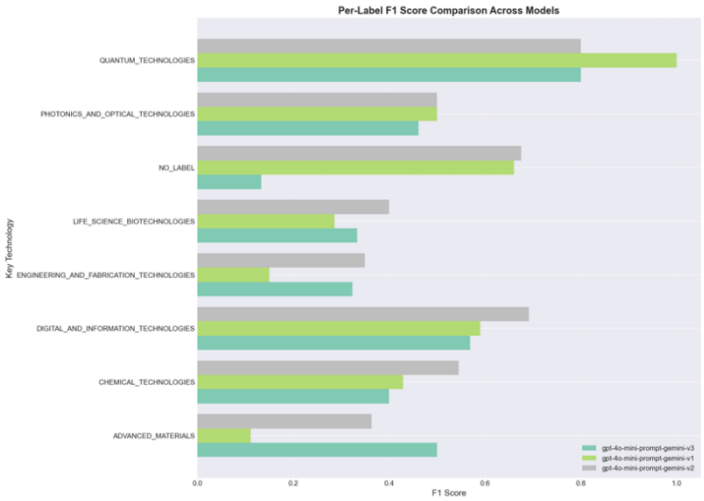
1.V2 (Best): F1=0.541, Recall=0.641, Precision=0.525
2.V1 (Middle): F1=0.466, Recall=0.468, Precision=0.520
3.V3 (Worst): F1=0.440, Recall=0.621, Precision=0.452

Performance Pattern Analysis:
V2 - Optimal Balance:
•Highest F1 score indicates best overall performance
•Highest recall AND precision shows well-balanced prompt design
•Successfully optimizes both sensitivity and specificity
V1 - Conservative Approach:
•Lowest recall (0.468) suggests under-prediction/conservative classification
•Decent precision (0.520) indicates low false positive rate
•Pattern suggests prompt may be too restrictive
V3 - Liberal Approach:
•High recall (0.621) but lowest precision (0.452)
•Classic over-prediction pattern - catches most true positives but generates many false positives
•Prompt likely too permissive/broad
Conclusions:
V2 represents the most effective prompt engineering approach because:
•Best F1 score indicates optimal precision-recall tradeoff
•Achieves highest values in both individual metrics
•Demonstrates superior prompt design for this classification task

TNO innovation for life 20

Experiment 2 Prompt variants

V2 as stated best balance



1. V3's Poor Overall F1:
- Despite some strong per-class F1 scores, V3's aggressive over-prediction creates many false positives
- Classic high recall, low precision pattern

2. Minority Class Challenge:
- All models fail to adequately predict `ADVANCED_MATERIALS` and `PHOTONICS_AND_OPTICAL_TECHNOLOGIES`
- Suggests these classes need better examples or more distinctive prompt features

3. V1's Conservative Approach:
- Under-predicts most classes, leading to missed opportunities (lower recall)
- Explains why it has decent precision but lower overall F1

4. V2 Shows Best Balance:
Closest alignment to ground truth for NO_LABEL (~35 vs 38 GT)
Reasonable performance on most categories
Still struggles with minority classes

TNO innovation for life 21

Experiments

Experimental Framework

Prompt x Context x Model Variants

1 Original Prompt
Baseline test using the initial prompt design as a control condition

2 Prompt Strictness
Explore variations in speci· ficity and constraint levels

3 Refined Prompt
Iterate on the strongest prompt from step 2 with targeted refinements

4 Context Variation
Apply the refined prompt across diverse contextual scenarios

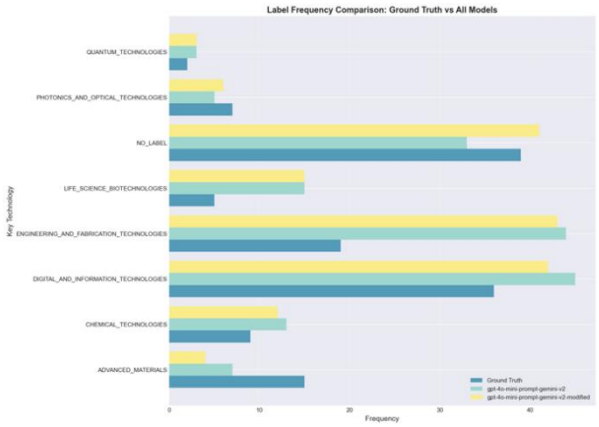
5 Model Comparison
Evaluate the refined prompt using multiple model variants

6 Ensemble Method
Combine the refined prompt with top-performing models for maximum effectiveness

TNO innovation for life 22

Experiment 3 Prompt variant : more specific information per KET

Fixing one problem, creating another



Key Changes from Original V2 to Modified V2:

Significant Improvements:

- `LIFE_SCIENCE_BIOTECHNOLOGIES`: Much closer to GT (~15 vs GT=8, compared to original ~25)
- `CHEMICAL_TECHNOLOGIES`: Better alignment (~12 vs GT=10, compared to original ~15)

New Problems Created:

- `ENGINEERING_AND_FABRICATION_TECHNOLOGIES`: **Severe over-prediction** (~45 vs GT=25, compared to original's perfect ~25)
- `NO_LABEL`: Increased over-prediction (~45 vs GT=38, compared to original's good ~32)

Maintained Issues:

- `ADVANCED_MATERIALS`: Still severe under-prediction (~3 vs GT=18), now even worse than original (~8)
- `PHOTONICS_AND_OPTICAL_TECHNOLOGIES`: Consistent under-prediction across both versions

Analysis Insights:

1. Trade-off Pattern:

- The modification improved performance on some chemical/biological categories
- But created aggressive over-prediction in engineering and general categories
- **Classic example of "fixing one problem, creating another"**

2. Modification Impact:

- Appears to have made the model more sensitive to engineering/fabrication terminology
- May have broadened the criteria for these categories too much
- Lost some of the balanced approach that made original V2 effective

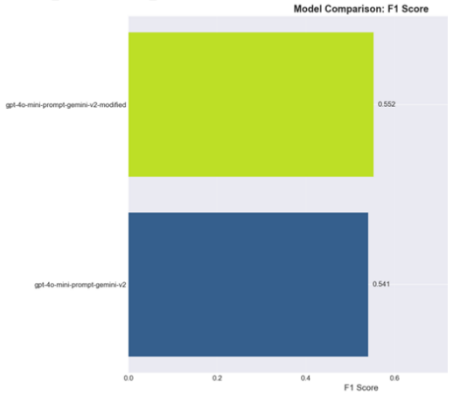
3. Overall Assessment:

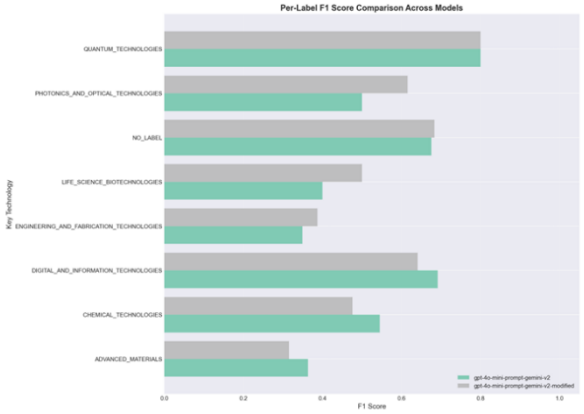
- **Original V2 likely still superior** due to better balance
- Modified version shows potential insights for targeted improvements
- Demonstrates how sensitive LLM classification is to prompt changes

TNO innovation for life 23

Experiment 3 Prompt variant : more specific information per KET

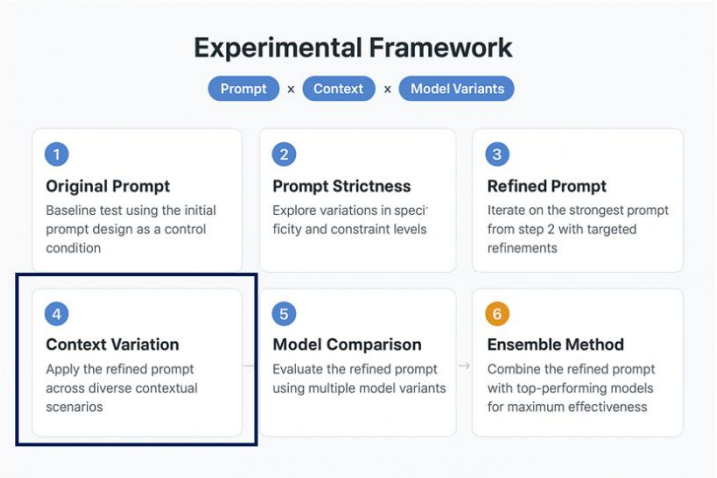
Best prompt using sections v2 versus modified prompt





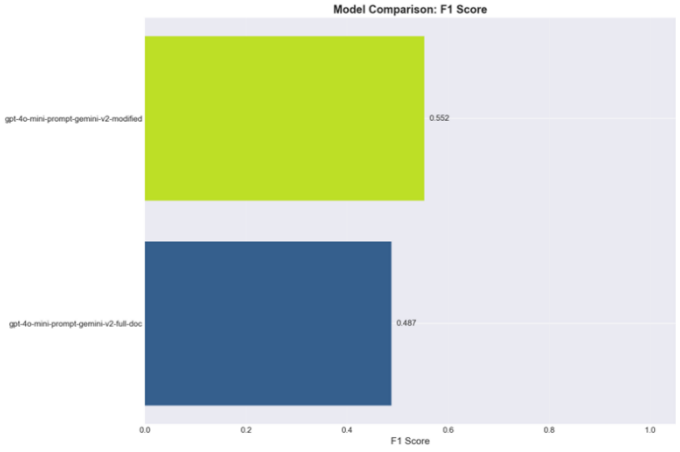
TNO innovation for life 24

Experiments



Experiment 4: Full document vs section processing: Context

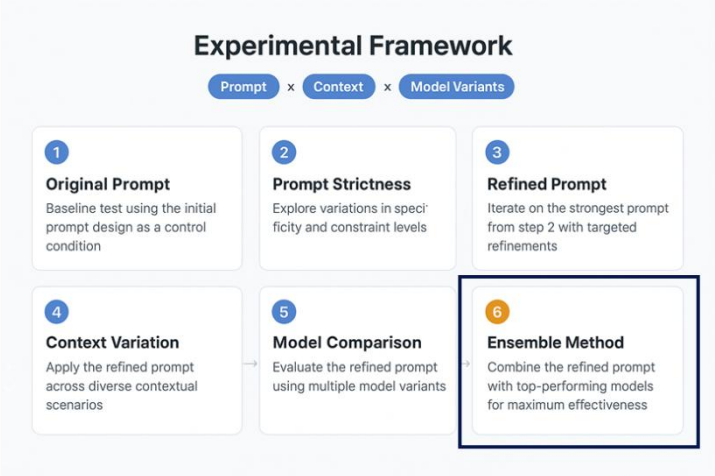
Best prompt using sections v2 vs full document v2 4o



* Giving too much context leads to degrading performance

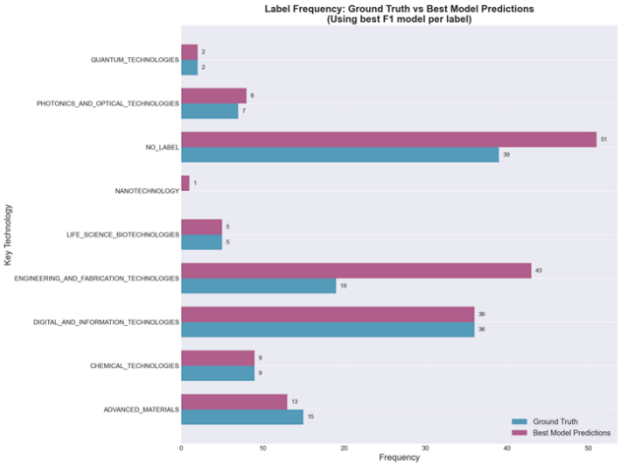
* Performance was less for all labels

Experiments



Ensemble based on F1 score

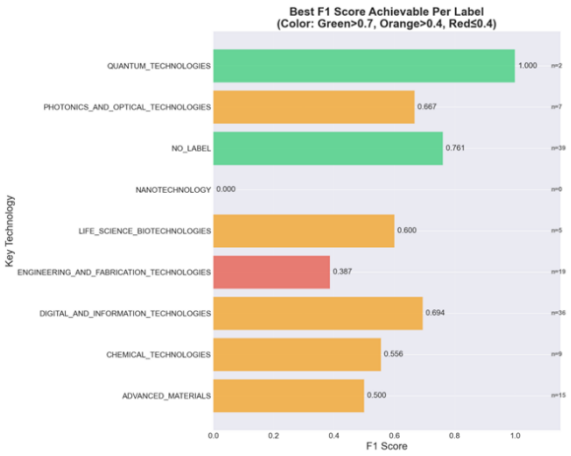
Put everything together with an ensemble (F1 score)



- Still over prediction for NO_LABEL (suggesting model default to this when uncertain) and ENGINEERING_AND_FABRICATION_TECHNOLOGIES
- Don't get too focused on low number of ground truth of KETs

Ensemble based on F1 score

In general poor to fair performance*



Metric Type	F1 Score	Precision	Recall
Weighted	**0.6397**	0.6037	0.7015
Macro	**0.5739**	0.5532	0.6129

Good (F1 > 0.7): QUANTUM_TECHNOLOGIES (1.0), NO_LABEL (0.761)

Fair (F1 0.6-0.7): DIGITAL_AND_INFORMATION_TECHNOLOGIES (0.694), PHOTONICS_AND_OPTICAL_TECHNOLOGIES (0.667)

Poor (F1 0.5-0.6): LIFE_SCIENCE_BIOTECHNOLOGIES (0.600), CHEMICAL_TECHNOLOGIES (0.556), ADVANCED_MATERIALS (0.500)

Very Poor (F1 < 0.5): ENGINEERING_AND_FABRICATION_TECHNOLOGIES (0.387)

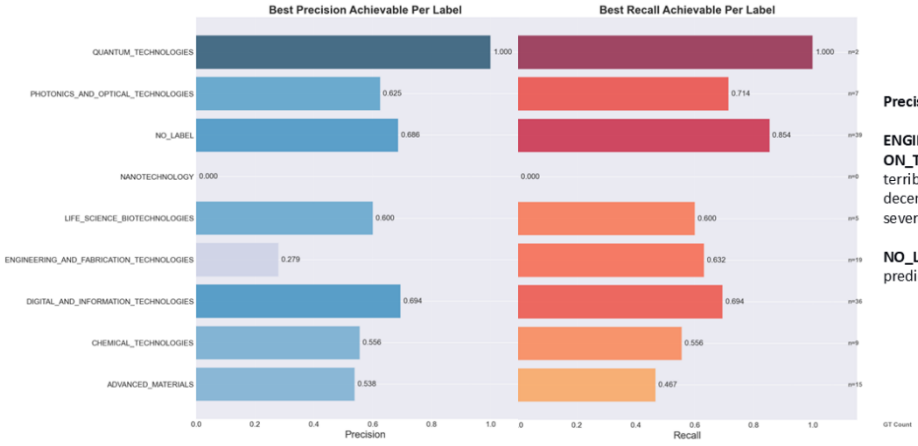
Failed: NANOTECHNOLOGY (0.000)

TNO innovation for life 29

*We have quite some imbalance, but still would need more data for variety of reasons

Ensemble based on F1 score

Precision and recall scores



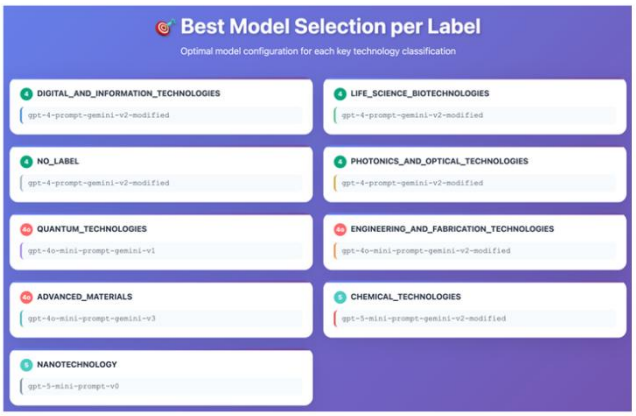
Precision-Recall Trade-offs

ENGINEERING_AND_FABRICATION_TECHNOLOGIES shows terrible precision (0.279) but decent recall (0.632), indicating severe over-prediction

NO_LABEL is significantly over-predicted,

TNO innovation for life 30

Mixture of models



Model	Labels Won
gpt-4-prompt-gemini-v2-modified	4
gpt-4o-mini variants	3
gpt-5-mini variants	2

No free lunch nor no free hunch

TNO innovation for life 31

Learnings

A lot of suboptimal decisions on the data / label side will fire back

Formalizing contribution to key technologies is not trivial let alone a translation towards an LLM when to output (or not) a key technology

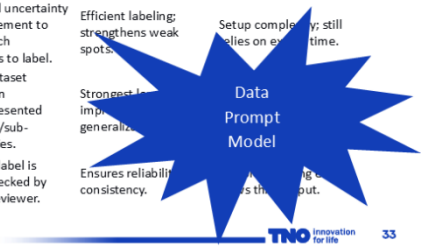
Prompt optimization can rapidly explode: need for more automated / programmatic process in such cases (e.g. using frameworks like DSPy and GEPA for automatic prompt optimization)

At end of the day, it begins with good quality (and enough) data

TNO innovation for life 32

Route forward

Approach	Effort Level	Expected Impact	Needed	Description	Pros	Cons
Expert & peer review of prompt and labelled data	Low	Low–Medium	Time from experts/labelers	Experts review the current prompt and labels for clarity and consistency.	Quick to implement; improves quality.	Doesn't fix gaps in underrepresented categories.
Prompt alternatives (manual design)	Medium	Medium	Time from experts to draft/test prompts	Create and test multiple prompt formulations systematically.	Low-cost; can yield noticeable improvements.	Trial-and-error; limited by data sparsity.
Prompt optimization (automatic methods)	Medium	Medium	Tools/infrastructure for automated optimization	Use automated tools to explore and optimize prompts at scale.	More systematic than manual design.	Needs infra; may overfit to limited dataset.
Active learning for targeted labeling	Medium–High	High	Workflow setup + expert labeling time	Use model uncertainty or disagreement to select which documents to label.	Efficient labeling; strengthens weak spots.	Setup comple y; still elies on e time.
Label more data in sparse categories	High	High	Significant expert labeling time; careful doc selection	Expand dataset focusing on underrepresented categories/sub-technologies.	Strongest i imp generali	
Assign peer reviewer for each new label	High	Medium–High	Double expert/labeler time for each doc	Each new label is double-checked by another reviewer.	Ensures reliabili consistency.	ut.



TNO innovation for life 33

Some remarks regarding the category's "overlap"

Digital and information technology is a very broad category and touches a lot of KIP documents: *"Digital and Information technologies is the collective term for all data- and information-driven technologies".*

Engineering and fabrication technology often touches Digital aspect as well: *"They include a wide range of technologies focused on developing, qualifying, and validating advanced manufacturing processes, machines, and equipment, as well as (digital) monitoring and control through sensors, digital technologies, and other equipment. "*

It is not strange they occur quite often in parallel: Evaluate in peer review

TNO innovation for life 34