pleias

KU LEUVEN

NVIDIA

Nemotron-Personas France and Belgium

Synthetic personas grounded in census data

Dr. ing. Pieter Delobelle
Lead AI scientist @ Pleias
Guest professor @ KU Leuven

France and Belgium join a growing collection

France was dataset #6 and Belgium was dataset #10

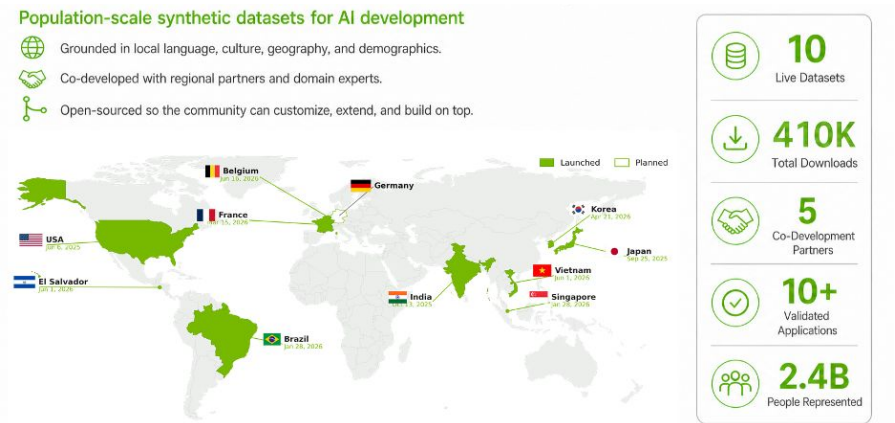


Figure: NVIDIA / SOOFI, 12 August 2026

Sample a person, then write the persona

Two steps produce each persona

A **Sample**

A Bayesian network, learned from the census, draws one person. OCEAN adds personality traits.

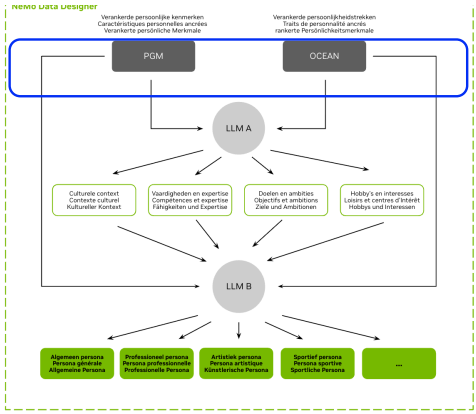


Figure: Nemotron-Personas-Belgium dataset card

pleias

Sample a person, then write the persona

Two steps produce each persona

A **Sample**

A Bayesian network, learned from the census, draws one person. OCEAN adds personality traits.

B **Generate**

LLM A writes the attributes of this person. LLM B uses them to write the personas.

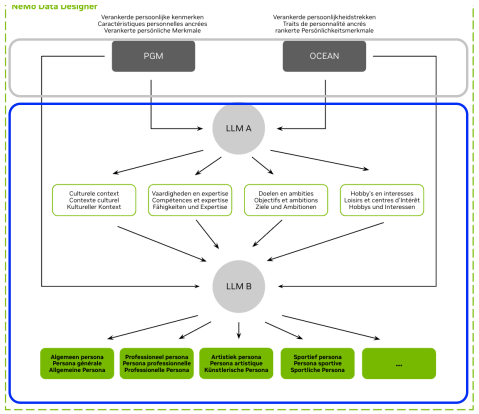


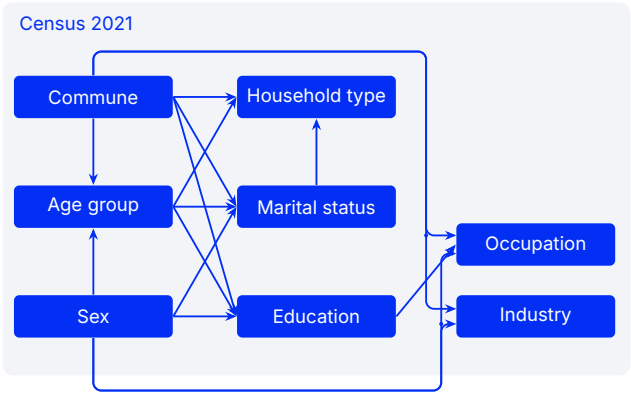
Figure: Nemotron-Personas-Belgium dataset card

pleias

STEP A · SAMPLE

The network is learned from the census

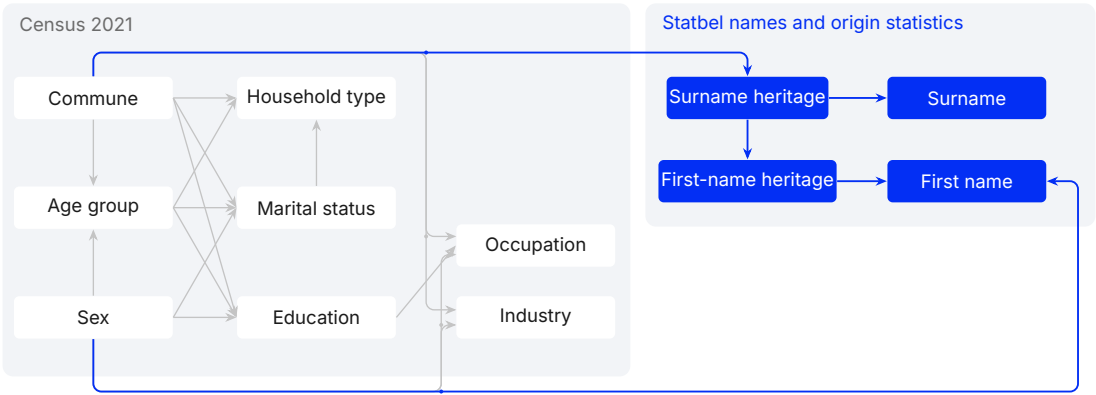
Belgium: Statbel tables per commune, age and sex



STEP A · SAMPLE

The network is learned from the census

Belgium: Statbel tables per commune, age and sex



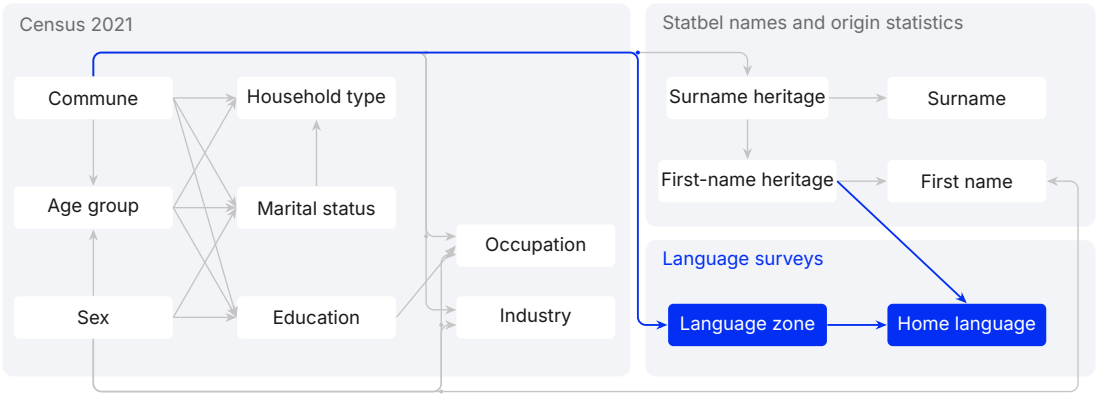
Marital status by age: population structure 2025. First-name heritage: Belgian Journal

pleias

STEP A · SAMPLE

The network is learned from the census

Belgium: Statbel tables per commune, age and sex



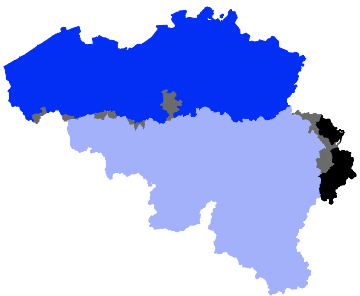
Marital status by age: population structure 2025. First-name heritage: Belgian Journal

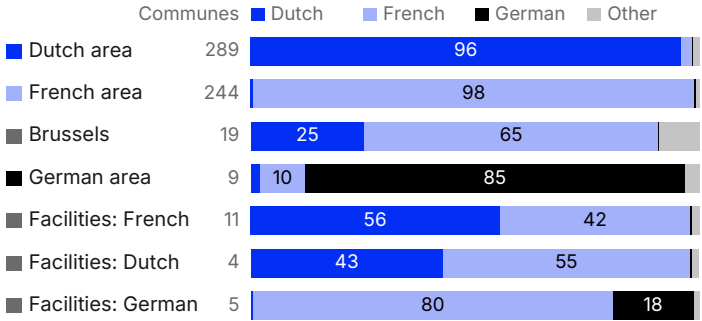
pleias

STEP A · LANGUAGE

Home language follows zone and first name

No census language question since 1961, so the zone priors come from surveys





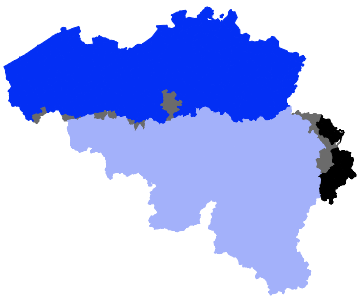
$P(\text{language} \mid \text{zone},\ \text{first-name heritage})$

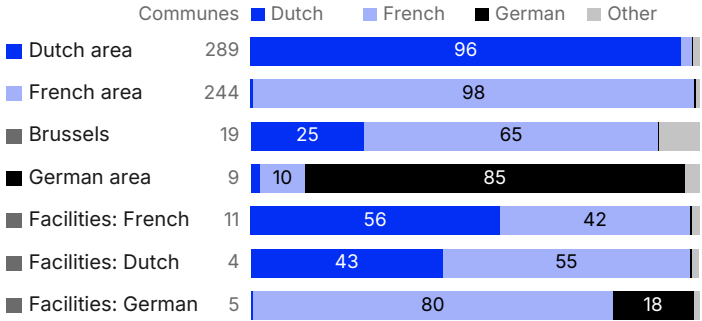
Shares in %, from 1M samples of the final model

STEP A · LANGUAGE

Home language follows zone and first name

No census language question since 1961, so the zone priors come from surveys





$P(\text{language} \mid \text{zone},\ \text{first-name heritage})$

In Brussels, 36% with a Flemish first name speak Dutch at home. With a Walloon first name: 22%.

Shares in %, from 1M samples of the final model

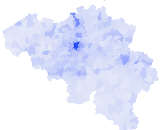
STEP A · NAMES

Heritage links the commune to the names

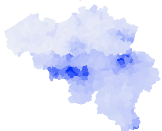
17 heritage groups, set per commune from Statbel origin statistics

Surname heritage per commune

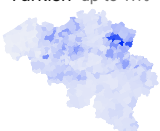
Maghreb up to 35%



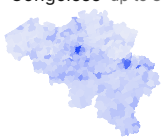
Italian up to 26%



Turkish up to 11%



Congolese up to 8%



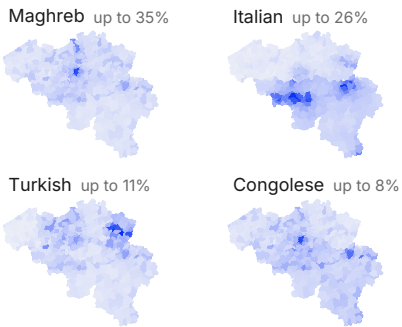
Name labels: DeepSeek-V3.2 and manual review. Name pairs: Flair NER on the Belgian Journal

STEP A · NAMES

Heritage links the commune to the names

17 heritage groups, set per commune from Statbel origin statistics

Surname heritage per commune



First-name heritage, given surname heritage (%)

Surname / first name	Flemish	Walloon	Maghreb	E. Eur.	Italian	Belgian	Dutch	Turkish	Other
Flemish	27	36				18			
Walloon		58				22			
Maghreb	17		54						
E. European	73								
Italian		33			24	15			
Belgian	11	50				23			
Dutch	67	18							
Turkish								67	

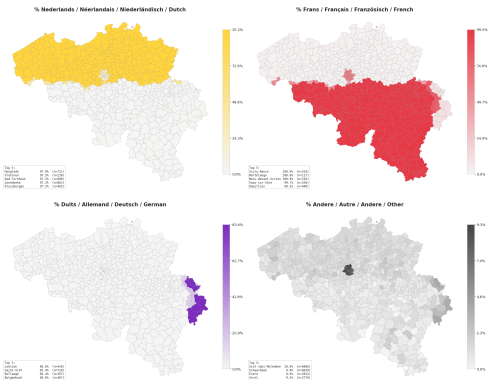
Name labels: DeepSeek-V3.2 and manual review. Name pairs: Flair NER on the Belgian Journal

pleias

STEP A · RESULT

The samples follow the language borders

Share of sampled personas per language and commune



Highest share per language

97.9%	Dutch, Hooglede
100%	French, Sivry-Rance
86.0%	German, Lontzen
10.0%	Other, Molenbeek

Belgium: 58% NL, 39% FR, 1% DE, 2% other

Figure: Nemotron-Personas-Belgium dataset card

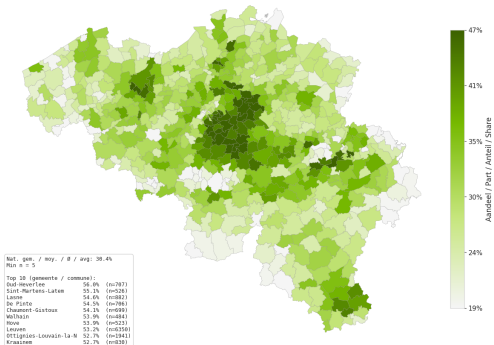
pleias

STEP A · RESULT

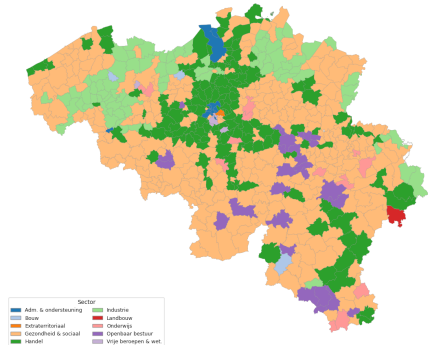
The samples keep the local patterns

Belgium, per commune

Higher education, age 25+



Largest industry



Figures: Nemotron-Personas-Belgium dataset card

pleias

STEP B · GENERATE

335k generated, 316k pass in all languages

Belgium: Gemma 4 31B on Jean Zay



France: generated with Nemotron 3 Super

pleias

STEP B · GENERATE

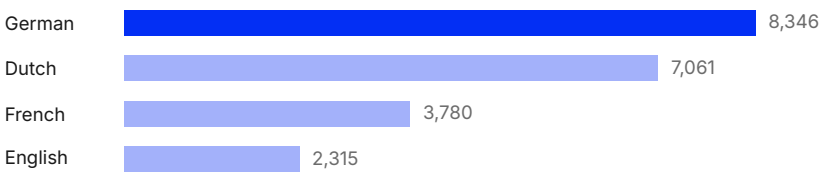
335k generated, 316k pass in all languages

Belgium: Gemma 4 31B on Jean Zay

335,000 personas generated → −19,144 (5.7%) → 315,856 pass in NL, FR, DE and EN → −15,856 surplus → 300,000 published per language

Dropped per language

a field is missing or uses another script



France: generated with Nemotron 3 Super

pleias

RESULT

Two open datasets on Hugging Face

Both released under CC BY 4.0

	Records	Personas	Languages	Communes	Tokens	LLM
France	1M	6M	1	31.6k	1.5B	Nemotron 3 Super
Belgium	300k	1.8M	4	581	1.9B	Gemma 4 31B

nvidia/Nemotron-Personas-France (March 2026) and nvidia/Nemotron-Personas-Belgium (June 2026)

RESULT

Two open datasets on Hugging Face

Both released under CC BY 4.0

	Records	Personas	Languages	Communes	Tokens	LLM
France	1M	6M	1	31.6k	1.5B	Nemotron 3 Super
Belgium	300k	1.8M	4	581	1.9B	Gemma 4 31B

One Belgian record: man, 37, Antwerp, technician in transport and storage

NL Herwig Andries is een nuchtere Antwerpse technicus met vette handen die zijn trots put uit de mechanica van de haven en de gezelligheid van het volkscafé.

DE Herwig Andries ist ein bodenständiger Antwerpener Techniker mit ölverschmierten Händen, der seinen Stolz aus der Hafenmechanik und der Geselligkeit der Eckkneipe zieht.

FR Herwig Andries est un technicien anversois pragmatique, les mains marquées par la graisse, qui tire sa fierté de la mécanique portuaire et de la convivialité du café du coin.

EN Herwig Andries is a down-to-earth Antwerp technician with grease-stained hands who finds his pride in port mechanics and the social buzz of the local pub.

nvidia/Nemotron-Personas-France (March 2026) and nvidia/Nemotron-Personas-Belgium (June 2026)

pleias

KU LEUVEN

NVIDIA

Both datasets are open

France and Belgium on Hugging Face

France



1M records, French

Belgium



300k records, 4 languages

Dr. ing. Pieter Delobelle
Lead AI scientist @ Pleias
Guest professor @ KU Leuven